Infrared Spectroscopy of Chemically Diverse Carbon Clusters: A Data-Driven Approach

Florent Calvo, Aude Simon, Pascal Parneix, Cyril Falvo, Clement Dubosq

Infrared spectroscopy of chemically diverse carbon clusters: a data driven approach

Florent Calvo,[*,†] Aude Simon,[‡] Pascal Parneix,[¶] Cyril Falvo[¶,†] and Clément Dubosq[§]

†*University Grenoble Alpes, CNRS, LiPhy, 38000 Grenoble, France*

‡*Laboratoire de Chimie et Physique Quantiques LCPQ/FeRMI, UMR5626, Université de Toulouse (UPS) and CNRS, Toulouse, France*

¶*Université Paris-Saclay, CNRS, Institut des Sciences Moléculaires d'Orsay, 91405, Orsay, France*

§*Laboratoire de Chimie et Physique Quantiques LCPQ/IRSAMC, UMR5626, Université de Toulouse (UPS) and CNRS, Toulouse, France*

E-mail: florent.calvo@univ-grenoble-alpes.fr

Abstract

Carbon clusters exhibit a broad diversity of topologies and shapes, encompassing fullerene-like cages, graphene-like flakes, and more disordered pretzel-like and branched structures. Here we examine computationally their infrared spectra in relation with these structures, from a statistical perspective. Individual spectra for broad samples of isomers were determined by means of the self-consistent charge density functional based tight-binding method, and an interpolation scheme is designed to reproduce the spectral features by regression on a much smaller subset of the sample. This interpolation proceeds by encoding the structures using appropriate descriptors and selecting them through a principal component analysis, Gaussian regression or inverse distance weighting providing the nonlinear weighting functions. Metric learning is employed to reduce the global error on a preselected testing set. The interpolated spectra satisfactorily reproduce the specific spectral features and their dependence on size and shape, enabling quantitative prediction away from the testing set. Finally, the classification of structures within the four proposed families is critically discussed through a statistical analysis of the sample based on iterative label spreading.

Keywords: Infrared spectrum, carbon clusters, statistical analysis, machine learning, density-functional tight-binding

1 Introduction

The interest in the formation of fullerenes in the gas phase dates back from their experimental discovery.[1,2] Recently, the astronomical detection of neutral[3] and cationic[4] fullerenes in the interstellar medium has revived this topic and notably led to several proposed formation mechanisms involving either smaller or larger building blocks.[5,6] Such astrophysical observations could be achieved by spectroscopy, for which the convenient signature of the highly symmetric buckminsterfullerene was the key to its identification.

Spectroscopy is undoubtedly the method of choice on which astrochemists rely to better understand the fullerene road under such harsh astrophysical conditions.

Until now, the traditional strategy for interpreting astrophysical spectra of complex molecules, in both the infrared (IR) or optical ranges, has mostly rested on the comparison with spectra usually obtained from electronic structure calculations offering a sufficient trade off between accuracy and efficiency, density-functional theory (DFT) and its time-dependent version representing the most popular methods of choice.[7,8] In some cases such as polycyclic aromatic hydrocarbons (PAHs), further refinements have been attempted in order to account for anharmonicity or finite temperature effects in vibrational spectra,[9–11] or to model the IR emission cascade spectrum originating from a UV excitation.[12]

Unraveling the molecular compounds responsible for the spectroscopic features observed in astrophysical objects like planetary nebulae can also be tackled from a more statistical perspective, without assuming specific molecular structures but focusing instead on broad relevant samples generated by intuition or, in a much less biased way, by means of high-throughput atomistic simulations.[13] In particular, upon categorizing conformers into different structural families, their generic spectral features can be inferred from the individual contributions of each member. Such an approach was found fruitful in assisting the interpretation of astronomical spectra of planetary and protoplanetary nebulae,[14] highlighting in particular the potential role of compounds with a significant aromatic content such as fullerenes. While such a statistical averaging washes away the detailed spectroscopic information pertaining to individual conformers, it is consistent with the relatively poor knowledge of such astrophysical regions and the likely presence of many chemical species. However, extending such a bottom-up strategy to broader classes of compounds having a larger size or containing heteroatoms is not straightforward, because of the very fast increase in the number of stable conformers and the difficulty of sampling their potential energy landscape as exhaustively as possible. Moreover,

for each conformer a suitable method is needed to determine their spectrum, which for chemical accuracy requires either advanced (polarizable) force fields or schemes offering an explicit description of electronic structure, even simplified.

While the two tasks of sampling the energy landscape and determining the individual spectra can be addressed successively along the lines of a multiscale description, the computational effort associated with the spectral determination can be particularly heavy for large samples, even with efficient methods such as density-functional based tight-binding (DFTB). However, extracting the spectral features from a statistical sample seems also naturally suited to be tackled by machine learning (ML) techniques, a broad range of computational approaches that have become spectacularly popular in chemical physics and physical chemistry in the recent years.[15,16] Within the general context of relating properties to structure, several groups have recently shown the benefits of employing ML for vibrational spectroscopy[17–19] through a variety of approaches aiming to represent potential energy and electric dipole moment surfaces within perturbative frameworks,[20] to condense molecular information into topological descriptors,[21] or to numerically solve the quantum nuclear dynamics problem by better partitioning the various degrees of freedom.[22]

In the present contribution, we explore several ML ideas to reconstruct the infrared spectrum of carbon clusters in a statistical sense, from a limited sample and using interpolation techniques in a multidimensional feature space, supervision being introduced through metric learning. We notably show how such an interpolation scheme can predict the spectral trends for clusters that are not members of the initially chosen sample. We also use a clustering algorithm to further address the classification of conformers into structural families.

In the next section, we briefly describe how the samples of conformers were obtained and the infrared spectra reconstructed from these samples. In Sec. 3 we present several statistical analyses of the spectra in terms of their convergence within the sample, their

reconstruction by interpolation assisted by metric learning, and we discuss the typical error of the method on individual and collective spectra. Sec. 5 discusses further the classification of the structures into families based on iterative label spreading[23] rather than ad hoc order parameters alone. Finally, some concluding remarks are given in Sec. 6.

2 Sample generation

Our systems of interest are carbon clusters C_n of selected sizes n, which exhibit a great diversity of structures that have been the specific subject of earlier contributions.[13,24] Here we summarize the main steps of the sample generation, and how the infrared spectra were obtained from it.

2.1 Force field exploration

The samples of isomers for C_{24}, C_{42}, and C_{60} were originally produced by replica-exchange molecular dynamics simulations based on the REBO potential,[25] with details given in Ref. 13. For these simulations, the clusters were enclosed in spherical containers to prevent dissociation at high temperature, and various simulations using different radii were performed to enhance sampling, fixing the densities at $\rho = 0.025$, 0.15, 0.4, and 1.7 g.cm^{-3} for the three cluster sizes. These densities were chosen on a trial-and-error basis in order to generate clusters with different structural trends. In particular, lower densities are needed to favor the very extended branched conformers, while higher densities tend to produce cages more efficiently.

New simulations were also performed for C_{33} and C_{52} to assess the performance of the interpolation scheme introduced below, away from the testing set that only includes data for C_{24}, C_{42} and C_{60}. For these additional clusters, we used a more efficient criterion based on bond connectivity in a parallel tempering Monte Carlo framework: a configuration was rejected if two subclusters are separated from each other by more than 3 Å. Except for

this difference, the parameters were otherwise similar, with 28 temperatures distributed in a geometrical fashion in the range 500–5500 K, and 4 additional temperatures around the melting point.

The total numbers of distinct conformers obtained with the REBO potential are listed in Table 1 for the five clusters.

Table 1: Numbers of distinct isomers in the samples produced by the simulations with the REBO potential, and after local reoptimization at the DFTB level.

Cluster size	REBO sample	DFTB sample
24	51 901	44 341
33	62 974	52 682
42	240 305	196 519
52	260 746	150 475
60	656 438	309 167

2.2 Reference infrared spectra

Because it lacks any information about the dipole moment surface, the REBO potential energy surface used to generate the structures is not suited to infrared spectroscopy. Instead we resorted to the self-consistent-charge density-functional-based tight-binding method[26] for this purpose, uusing the dispersion correction parameters of Ref. 14. The individual absorption spectrum associated with each structure was determined in the harmonic approximation after local reoptimization. Throughout this article (and in the supplementary material), this method will be denoted simply as DFTB. Structures saved periodically and locally reoptimized using the REBO potential were further refined using the DFTB method to calculate their IR absorption spectrum. The numbers of distinct structures obtained after this reoptimization step, listed in Table 1, are always lower than the initial sample size, because occasionally several starting points endded up into the same minimum (a reduction in information and pool size would also naturally occur by conducting the optimizations in the opposite way). Here we should also note that the

loss of a significant fraction in the initial sample size could be partly alleviated using ML techniques that screen the structure before deciding whether they are worth considering for further optimization.[27]

A standard double harmonic approximation was employed to calculate the IR absorption spectrum with the DFTB method, without any scaling factor for the frequencies. This choice is motivated by the methodological nature of the present study, rather than on any intention of a direct comparison with existing experiments or observations. It should be noted that, for the present systems, different scaling factors should also be applied in different spectral ranges for such a comparison,[14] which also leads to narrow spectral ranges being uncovered and appearing as spurious holes.

The spectra of entire structural families were determined as in Ref. 14 by simple addition over all contributions from individual members and no particular weighting. Individual spectra were made smoother by Gaussian broadening with a 5 cm^{-1} width over the 0–3000 cm^{-1} range with a 3 cm^{-1} bin size. Such parameters were necessary in order to satisfactorily describe both the broad plateaus in the 500–1500 cm^{-1} range, as well as the few narrow peaks occuring at various frequencies depending on the structural family.

Finally, and following our earlier work,[14] the high-frequency part of all IR absorption spectra was further attenuated using the blackbody radiation law at 300 K in order to make comparison with IR *emission* astrononomical spectra more realistic. All DFTB calculations were performed with the deMonNano software package.[28]

2.3 Structural classification

At the sizes n covered here of $n = 24$, 42, and 60, carbon clusters exhibit a broad variety of structures that can be categorized using order parameters. Following our earlier analysis[13,14] we use the fraction of sp^2 atoms and a dimensionless asphericity shape parameter β to sort these structures into four main families: (i) the cages family, which includes the archetypal fullerenes, consists of a high sp^2 content and a low asphericity; (ii) the flakes

family, notably including graphene cuts, also has a high fraction of sp^2 atoms but a significant asphericity; (iii) pretzel-like structures,[24] containing several chains of sp^1 atoms but not necessarily associated with marked deformations from the sphere; (iv) branched structures also dominated by long chains of sp^1 atoms but significantly aspherical.

The shape parameter β is one of the three quantities that can be obtained by assigning the three principal momenta of inertia I_k of the conformer following the Hill-Wheeler representation as

$$I_k = \frac{2}{3} r_c^2 \left[1 + \beta \sin \left(\gamma + \frac{(4k-3)\pi}{6} \right) \right], k = 1, 2, 3, \tag{1}$$

where r_c is the gyration radius and the angle γ measures the cluster triaxiality. With such a definition, β lies in the range 0–1, small values indicating nearly spherical shapes while high values are typical of prolate ellipsoids. The fraction of sp^2 atoms is straightforwardly obtained in the DFTB scheme from the orbital populations.[14] In the REBO model, it is simply estimated from the relative number of carbon atoms with exactly three nearest neighbors.

With such dimensionless order parameters at hand, all conformers were categorized into either of the four families, with sizes that are given in Table 2 for each of the five clusters.

Table 2: Sizes of the configurational samples for all carbon clusters considered in this work, as described by the DFTB method.

System	cages	flakes	pretzels	branched
C_{24}	11	740	6312	37282
C_{42}	27350	40740	36013	92416
C_{60}	82111	78182	38141	110733
C_{33}	13	4764	202	47703
C_{52}	20331	56067	741	73336

This partitioning of isomers between structural classes was initially introduced without reference to their spectroscopic response, and in the following we challenge this clas-

sification using additional tools.

3 Statistical analyses

3.1 Convergence of global spectra

For each cluster size, the IR spectra associated with each of the four structural families result from the combination of all individual spectra from each member of the family, and we first discuss how the resulting spectra are sensitive to the diversity among the corresponding sample. Individual IR spectra are generally highly resolved, because the clusters are not very large and display at most 172 fundamental modes, of which only a subset are IR active. However, the combination of many such spectra into an average spectrum for the corresponding family will likely yield much smoother features all the more than the family is large and diverse.

Here we introduce an error measure to quantify the discrepancy between a reference spectrum $\mathcal{I}_{\rm ref}(\omega)$ and an estimated spectrum $\mathcal{I}(\omega)$, from the average sum of the integrated difference between them on the relevant interval of interest:

$$\begin{aligned} \mathcal{E} &= \frac{1}{\omega_{\rm max}-\omega_{\rm min}} \int_{\omega_{\rm min}}^{\omega_{\rm max}} |\mathcal{G}_{\rm ref}(\omega) - \mathcal{G}(\omega)|\, d\omega, \\ \mathcal{G}_{\rm ref}(\omega) &= \int_{\omega_{\rm min}}^{\omega} \mathcal{I}_{\rm ref}(x) dx \\ \mathcal{G}(\omega) &= \int_{\omega_{\rm min}}^{\omega} \mathcal{I}(x) dx \end{aligned} \tag{2}$$

with $\omega_{\rm min} = 0$, $\omega_{\rm max} = 2500$ cm^{-1}. This error measure is not as sensitive to spectral shifts as the more standard least square error in which the first integral operates directly on $|\mathcal{I}_{\rm ref} - \mathcal{I}|$ or its square, and this was notably recognized recently by Kovacs and coworkers in the ML approach to the IR spectroscopy of polyaromatic hydrocarbons.[21] Here the reference spectrum is simply defined for each family as the spectrum obtained from the full corresponding sample.

To analyse the sensitivity of the spectra towards the structural diversity among the set, Monte Carlo simulations have been performed in which the spectra were reconstructed from a random limited subset. For each structural family, 1000 random subsets of conformers were chosen and the spectra averaged from their specific contributions.

The average errors obtained for the four families of C_{60} isomers are shown in Fig. 1 as a function of sample size. The converged spectra obtained from the full samples are

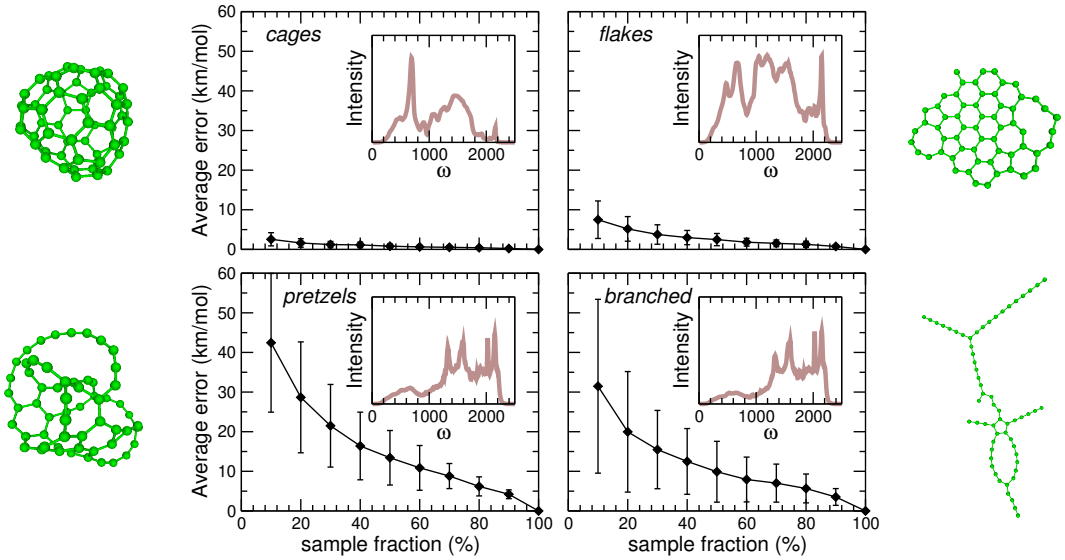

Figure 1: Average error in describing the IR spectra of the four structural families of C_{60} clusters, by selecting only a subset of the entire corresponding sample, with error bars that indicate the standard fluctuations around the average. The insets show the reference IR spectra obtained for the full samples for each family. Next to each panel corresponding to a specific family, a typical structure is also depicted.

themselves also shown as insets for subsequent reference, together with representative conformers of each family for visual identification.

Clearly the IR spectra obtained for the cages and flakes do not depend as much on the underlying sample as the more disordered pretzels and branched conformers, for which the average error is about one order of magnitude larger. The spectral features of cages and flakes are reasonably converged already with a few thousands of the corresponding structures, while this number is closer to tens of thousands for pretzels, and exceeds it for branched structures.

Similar results are obtained for C_{42}, but for C_{24} the numbers of cages and flakes are much smaller and the resulting spectra show a much greater dependence on the subset selection. The corresponding spectra and their statistical convergence are given as Supplementary information.

3.2 An interpolation scheme for IR spectra

The previous statistical analysis revealed that the spectral features associated with structural families are differently robust depending on the diversity within each family. We now attempt to use this information to reduce the amount of conformers needed to reproduce the global spectra in a more systematic and automated fashion that could also make it transferable to other systems.

More precisely, we use interpolation to relate the spectrum $\mathcal{I}_{\mathbf{R}}(\omega)$ of an arbitrary conformer $\mathbf{R}$ to the known spectra $\mathcal{I}_k(\omega)$ from reference conformers $\mathbf{R}_k$ of a fixed sample, much smaller than the complete set of conformers generated by simulation. In a first natural approach, linear interpolation is employed as

$$\widetilde{\mathcal{I}}_{\mathbf{R}}(\omega) = \sum_k g_{\mathbf{R},k} \mathcal{I}_k(\omega). \tag{3}$$

where the weight $g_{\mathbf{R},k}$ measures the degree of similarity between structures $\mathbf{R}$ and $\mathbf{R}_k$, and is a simple but nonlinear function of a distance $d_{\mathbf{R},k}$. Here we considered two kernels, namely Gaussian regression (GR) $g(d) = g_0 \exp(-\gamma d^2)$ with g_0 a normalization factor, as well as inverse distance weighting (IDW) $g(d) = 1/d^2$ if all $d > 0$, or, if $d = 0$ for some member of the set, then $g(d) = 1$ for this member and $g(d) = 0$ for all other members.

The distance $d_{\mathbf{R},k}$ is defined by a metric based on N descriptors that we collectively denote as $\mathbf{q}(\mathbf{R}) = \{q_j, j = 1, \ldots, N\}$. These descriptors were chosen to cover various molecular features of relevance in vibrational spectroscopy, and they are detailed below.

Next we select a reduced sample of conformers on which to interpolate the spectra for

arbitrary structures. In a first approach, we mesh the order parameter space in β and sp^2 fractions and only include those conformers that fall near the regular positions on this 2D grid, within the grid resolution for each system, namely 10×10 for C_{24} and C_{42}, and 8×8 for C_{60}.

Structures selected for C_{24}, C_{42} and C_{60} are all depicted onto the corresponding maps in Fig. 2. This sample of the totally available set of conformers amounts to about 2.4%

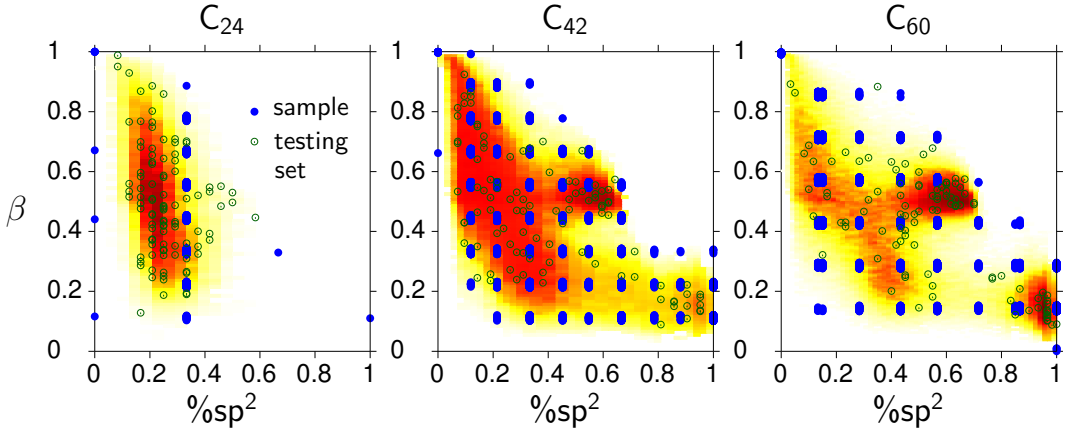

Figure 2: Distributions of locally optimized structures of carbon clusters C_n projected onto the plane of sp^2 fraction and asphericity shape parameter β, in logarithmic scale. The full blue and empty green circles mark the location of the interpolation sample and the testing sets members, respectively.

members only, or 13314 distinct structures.

Once descriptors are known, a metric can be chosen for d and the interpolation method can be applied straightforwardly to predict the IR spectra from the reference spectra on the reduced sample. However, the quality of the prediction can be improved using ideas from machine learning methods, and here metric learning will be employed.

3.3 Descriptors encoding molecular structures

Twenty features or descriptors were chosen to characterize the various isomers generated by molecular simulation. In the context of vibrational spectroscopy we have primarily used descriptors involved in the local bond properties involving between 2 and 4 con-

nected atoms, as well as more global properties related to the overall shape of the entire configuration or its bond topology seen as a graph. The descriptors are detailed and numbered as follows. As they should, they are all invariant over global translations, rotations, and permutations among atoms.

For a N-atom cluster, and from the gyration tensor $\mathbf{S} = \{S_{\alpha\beta}\}$,

$$S_{\alpha\beta} = \frac{1}{N}\sum_i r_i^\alpha r_i^\beta,$$

where r_i^α is the component $\alpha = x, y, z$ of atom i, we can extract its trace, giving the square gyration radius R_g^2, while the higher moments give insight into the so-called asphericity and prolateness parameters A and P.[29] Introducing the traceless tensor $\mathbf{D}$ we thus define the three first descriptors as

$$\begin{aligned} R_g^2 &= \mathrm{Tr}\,\mathbf{S} && \text{(D1)} \\ A &= 3\frac{\mathrm{Tr}\,(\mathbf{D}^2)}{2R_g^4} && \text{(D2)} \\ P &= 3\frac{\mathrm{Tr}\,(\mathbf{D}^3)}{2R_g^6} && \text{(D3)} \\ \mathbf{D} &= \mathbf{S} - \frac{R_g^2}{3}\mathbf{I}, \end{aligned}$$

where $\mathbf{I}$ is the 3×3 identity matrix. Note that only R_g^2 is size sensitive, A and P being both dimensionless quantities. Also note that the asphericity parameter A is not strictly identical to the parameter β used to classify the structures into the four families, for which Eq. (1) was used.

From the structure we then define bond connectivities using a cut-off radius of 1.85 Å, the average bond length and the fluctuations around this average providing our next descriptors

$$\bar{r} = \langle r_{ij}\rangle_{nn}, \qquad \text{(D4)}$$

$$\delta r = \left[\langle r_{ij}^2\rangle_{nn} - \langle r_{ij}\rangle_{nn}^2\right]^{1/2}, \tag{D5}$$

where the subscript nn indicates that average is taken over nearest neighbors. Likewise, for any triplet i, j, k of connected atoms we define the average angle $\bar{\theta}$ and the fluctuation around this average as

$$\bar{\theta} = \langle\theta_{ijk}\rangle_{nn} \tag{D6}$$

$$\delta\theta = \left[\langle\theta_{ijk}^2\rangle_{nn} - \langle\theta_{ijk}\rangle_{nn}^2\right]^{1/2}, \tag{D7}$$

and similarly torsion angles $\phi_{ijk\ell}$ are identified and included in the set for all quadruplets i, j, k, ℓ of connected atoms:

$$\bar{\phi} = \langle\phi_{ijk\ell}\rangle_{nn} \tag{D8}$$

$$\delta\phi = \left[\langle\phi_{ijk\ell}^2\rangle_{nn} - \langle\phi_{ijk\ell}\rangle_{nn}^2\right]^{1/2}. \tag{D9}$$

From the global connectivity we also calculate the adjacency matrix A_{ij}, which we diagonalize into the set of eigenvalues $\{\alpha_k\}$ from which the next descriptors are extracted as

$$\alpha_{\min} = \min_k \alpha_k \tag{D10}$$

$$\alpha_{\max} = \max_k \alpha_k \tag{D11}$$

$$\bar{\alpha}_+ = \langle\alpha_k^+\rangle \tag{D12}$$

In the above equations, $\langle\alpha_k^+\rangle$ denotes the average over all strictly positive eigenvalues (the trace and average over all eigenvalues being strictly zero for any adjacency matrix).

The four next descriptors are also related to bond connectivity but in more direct connection with the chemical nature of the carbon bonds. More specifically, the numbers of atoms that are singly (N_1), doubly (N_2), triply (N_3), and quadruply (N_4) connected to

other carbon atoms are identified and their fractions define the descriptors as

$$f_1 = N_1/N \tag{D13}$$

$$f_2 = N_2/N \tag{D14}$$

$$f_3 = N_3/N \tag{D15}$$

$$f_4 = N_4/N \tag{D16}$$

In a first approximation, these fractions measure the relative amounts of carbon atoms that terminate a chain or are hybridized as sp^1, sp^2, and sp^3, respectively. In particular, f_3 is used together with β to assign the REBO structures into the four families.

The next descriptor is also topological and defined as the normalized meshedness χ of the graph made by carbon atoms.[30] It is defined from the numbers of edges N and bonds N_b as

$$\chi = \frac{N_b - N + 1}{2N - 5} \tag{D17}$$

Finally, in relation with the aromatic nature of many carbon nanostructures, we have considered the numbers $\mathcal{N}_k$ of cycles of length $k = 5$–7 and normalized them to yield three more descriptors

$$\bar{\mathcal{N}}_5 = \mathcal{N}_5/N \tag{D18}$$

$$\bar{\mathcal{N}}_6 = \mathcal{N}_6/N \tag{D19}$$

$$\bar{\mathcal{N}}_7 = \mathcal{N}_7/N. \tag{D20}$$

3.4 Principal components analysis

Each structure in the samples generated for C_{24}, C_{42} and C_{60} was assigned a point in 20-dimensional space using the above listed descriptors. Because the various dimensions possibly cover quite different numerical ranges, we further standardize the dataset by

shifting each value with respect to the average, and dividing it by the mean square fluctuation. To assess the quality of the dataset and remove possible redundancies among descriptors, we next analyze the data in terms of their principal components (PCs).

After diagonalization of the PC matrix, the eigenvalues are ordered by decreasing value, and the eigenvectors corresponding to the highest eigenvalues are employed to determine the number of dimensions needed to capture the greatest amount of the data and their dispersion. Such an analysis is performed independently for the three datasets corresponding to C_{24}, C_{42} and C_{60} samples. The percentage of explained variance obtained from the eigenvectors of the PC matrix is given as an inset in Fig. 3 for the lowest 10 dimensions that correspond to the highest eigenvalues. Most of the variance can thus

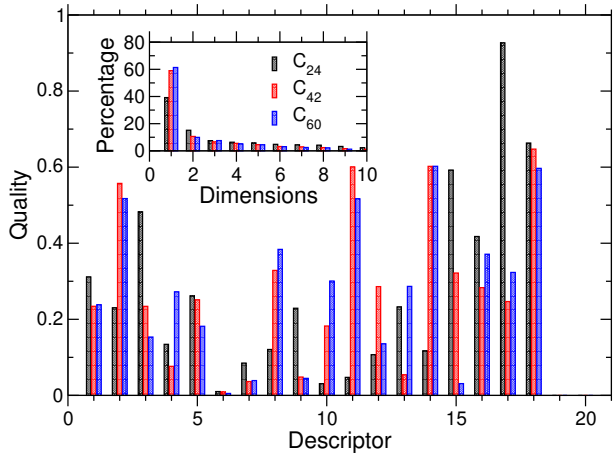

Figure 3: Quality of the description provided by using only the principal components corresponding to the 5 highest eigenvalues of the PC matrix, for the three samples of C_{24}, C_{42}, and C_{60}. The inset shows the percentage of variance explained by the successive principal components ordered by decreasing eigenvalue, up to dimension 10.

be explained with only 5 principal components, especially for the samples corresponding to the two larger clusters, the data being somewhat more widespread for C_{24}. We interpret this difference as reflecting the poorer ability of this small cluster to form regular cages, which contribute significantly to the other samples but are only very few (eleven) in this case.

The quality of the representation of the data by the 5 principal components associ-

ated with the highest eigenvalues of the PC matrix was determined from the sum of the corresponding square eigenvector coefficients. It is shown in the main part of Fig. 3 for the three samples against descriptor number, and reveals the different contributions of these descriptors to the set of 5 principal components. While the results for C_{24} are again slightly different from the two other systems, we find some common patterns that were not anticipated when the descriptors were originally designed: the angular terms for triplets of connected atoms essentially do not contribute, as is the case for the normalized numbers of 6- and 7-atom rings. In contrast, we find the shape parameters of asphericity and prolateness, and the fractions of atoms with 2 and 3 neighbors to be quite well represented in those principal components. These results are consistent with our initial but empirical choice of order parameters to sort the structures into the four families of cages, flakes, pretzels, and branched isomers.[13]

3.5 Metric learning

The previous PC analysis indicates that 10 descriptors are sufficient to capture the structural diversity amont the available set, and we thus discard those contributing the least to the 5 highest components, namely the average bond length and its fluctuation, the average bond angle and its fluctuation, the fluctuation in torsion angles, the maximum and average of positive eigenvalues of the adjacency matrix, and the fractions of 5-, 6-, and 7-rings.

Distances are thus evaluated in the remaining 10-dimensional set, and for each kernel $g(d)$ we introduce a specific metric $d(\mathbf{R})$ as a set of strictly positive numbers a_j such that

$$d^2_{\mathbf{R},k} = \sum_{j=1}^{N} a_j [q_j - q_j^{(k)}]^2,$$

where q_j and $q_j^{(k)}$ denote the corresponding descriptors for conformers $\mathbf{R}$ and $\mathbf{R}_k$, respectively.

To improve the quality of the interpolation scheme, we then introduce some degree of supervision to the algorithm by training the metric so the spectra predicted for a specific set of structures (a testing set) mimics as best as possible their true IR absorption spectra determined by the DFTB method. The testing set was chosen by randomly selecting 300 conformers among the totally available sample, 100 for each of the C_{24}, C_{42} and C_{60} systems, only ensuring none of them also belongs to the interpolation sample as this would make error evaluation non differentiable with the IDW kernel. The testing sample is also depicted on the density plots in Fig. 2.

For a given kernel, the metric is then optimized by minimizing a global error $\mathcal{E}$ over the whole testing set using again a Monte Carlo procedure, following here a zero temperature Metropolis acceptance rule in which random moves that do not decrease the error are rejected, alternating with steps of gradient-based local minimizations.

From a practical perspective, all statistical computations were performed with homemade codes.

4 Performance of the interpolation scheme

4.1 Individual IR spectra

The general behavior of the interpolation scheme can be first discussed on the example of individual structures. Figure 4 shows the reference and interpolated IR spectra obtained for members of the testing set that contribute the most and the least to the final error obtained after metric learning, using the GR and IDW kernels. The corresponding conformers are also shown on this figure. As already mentioned, the IR spectra of specific conformers are highly resolved for such relatively small clusters. By combining the contribution of various individual spectra, the interpolated spectrum is expected a priori to be much smoother and convey the statistically dominant features. For the two kernels, the best prediction is obtained for cage isomers of C_{60}, whose spectra exhibit rather

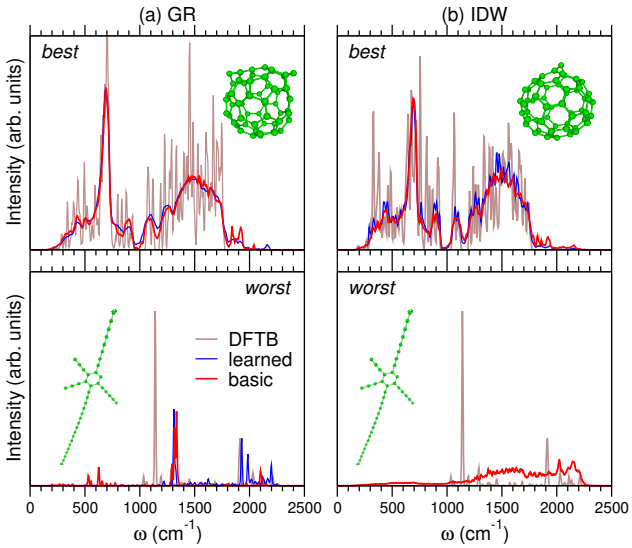

Figure 4: Most similar ('best') and dissimilar ('worst') individual IR absorption spectra, as defined from their lowest and highest contribution to the global error among the testing set, obtained for the GR (left panels) and IDW (right panels) kernels after metric learning (blue lines), or without (red lines). The DFTB reference spectra are shown as thick brown lines, and the corresponding conformers are depicted next to the spectra.

large absorption bands near 700 cm^{-1} and a main peak, as well as in the 1000–1700 cm^{-1} range with a broader bump. Here the interpolated spectrum manages to reproduce these features and especially the low intensity parts.

In contrast, the most dissimilar spectra are obtained for much more disordered, branched-type structures exhibitting fewer but highly intense IR active modes. The same conformer is responsible for this highest contribution to the error for both kernels. Interestingly, the Gaussian weight manages to preserve this highly resolved character, but the most intense peak near 1150 cm^{-1} is not reproduced. This suggests that none of the reference spectra in the interpolation sample matches this specific spectrum well enough to be able to produce anything similar in output. In turn, it suggests excessive structural dissimilarities between this specific conformer and all members from the sample. Inverse distance weighting produces an excessively smooth spectrum lacking intense features, as the result of multiple conformers from the sample producing similar weights.

These results indicate that the interpolation scheme can produce very different types

of spectra, typically smoother but not necessarily always smooth either since the weighting scheme is highly nonlinear. For comparison, we have also shown in Fig. 4 its prediction without carrying the metric learning stage, i.e. including all 20 descriptors in the metric and not optimizing its coefficients a_j. The most similar spectra are weakly altered, except at high frequency where spurious peaks are found near 2000 cm^{-1}. With inverse distance weighting, the interpolated spectrum is also even smoother and the error to the reference spectrum is increased by about 25%. Concerning the most dissimilar spectra, the absence of supervision produces an even worst prediction for Gaussian regression, with spurious peaks near 550 cm^{-1} and 2000 cm^{-1}, without improving on the most intense peak at 1150 cm^{-1}. However, the much broader averaging achieved by inverse distance weighting is preserved without metric learning, the predictions being very similar.

4.2 Global IR spectra

The global spectra predicted for the four structural families of C_{60} after averaging on all their members are shown in Fig. 5, in comparison with the reference DFTB spectra and for the two interpolation schemes employing the GR or IDW kernels. The corresponding spectra predicted for the families of C_{24} and C_{42} are given as supplementary information.

Overall, the spectral features are fairly well reproduced by both interpolation schemes, in the entire relevant spectral range. In agreement with the purely statistical analysis of Fig. 1, the greater discrepancies are found for the pretzels and branched families, for which convergence of the global spectrum is the slowest with set size. The most significant deviation occurs for the flakes, in the 1000–1200 cm^{-1} range, where interpolation with inverse distance weighting notably underestimates the IR intensity by a few percents, while the narrow peak at 2150 cm^{-1} is overestimated.

Without metric learning, the spectrum obtained for cages is barely affected, but more significant differences are found for the other families, especially with inverse distance weighting where the broad region 1200–2200 cm^{-1} is notably underestimated.

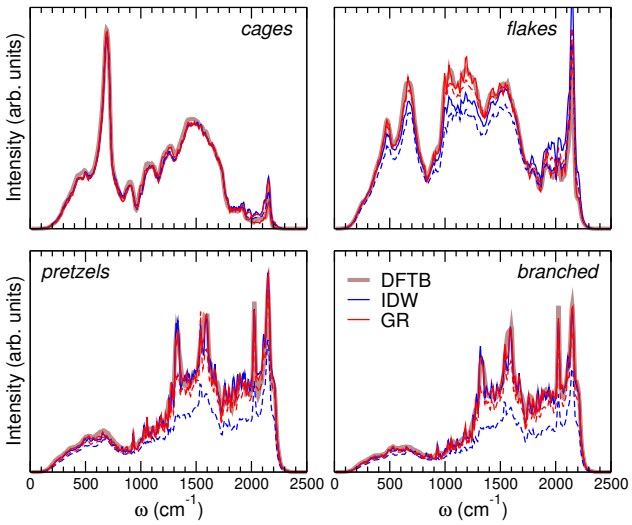

Figure 5: IR spectra of the four structural families of C_{60} clusters, obtained from DFTB electronic structure calculations (thick brown lines) and from sample interpolation employing the GR and IDW kernels (red and blue lines, respectively). The dashed lines show the corresponding predictions of the interpolation method without metric learning.

It is also quite striking that the spectral features of the pretzels and branched families are very similar, except for minor intensities differences in the broad band at 600 cm^{-1}. This spectroscopic similarity obviously results from structural similarities that can be ascribed to the predominance of sp^1 carbon chains in both cases, as discussed further below. The performance of the interpolation scheme was also tested on a more challenging case, namely a carbon cluster different from any member of the sample and testing sets. The energy landscape of C_{52} explored using a combined REBO-DFTB methodology provided broad samples of cages, flakes, and branched structures, although pretzels are fewer, their numbers being given in Table 2. Their IR spectra were determined for each conformer individually using DFTB, and accumulated to yield a global spectrum for each of the four families, serving again as reference.

The interpolation schemes were also applied to predict the IR spectra from the sole information of the conformers provided by their descriptors. These spectra are depicted in Fig. 6. For this cluster the main features of the IR spectra are reproduced very satisfactorily for the four families, the main deviations being found again in the 1000–1200 cm^{-1}

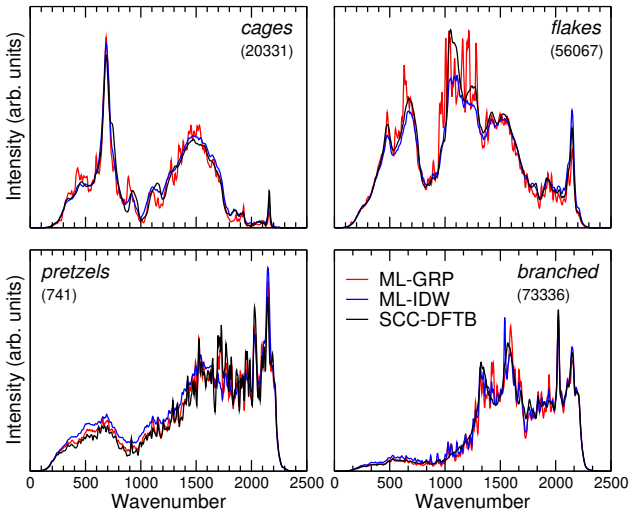

Figure 6: IR spectra predicted for structural families of C_{52} by sample interpolation using the GR and IDW kernels (red and blue lines, respectively), compared to the DFTB reference spectra (thick brown lines).

range for the flake conformers, Gaussian regression performing somewhat better than inverse distance weighting. The spectrum for the pretzels is more resolved owing to the much smaller size of this family, but the trends are again correctly described by both interpolation schemes. For the cages, the reference spectrum is mostly in error near 1000 cm^{-1}, while the spectrum for the branched family is remarkably described.

The interpolation schemes were also applied to the structures directly produced at the initial REBO level of modeling, without refining them using DFTB. The results, given as supplementary information, remain satisfactory for the more ordered cages and flakes families but for the disordered structures several spurious peaks are produced upon interpolation, especially when employing Gaussian regression, indicating that some conformers are not correctly recognized from their descriptors. This issue could probably be addressed by training the metric on a set made from REBO structures, rather than the DFTB conformers presently used. For the smaller cluster C_{33}, the number of cages and pretzels is rather reduced (see Table 2) and the resulting spectra are not particularly smooth. Yet the trends predicted by the interpolation schemes are again rather good, the main discrepancy appearing once again for the main band of the flakes family spectrum

near 1000–1200 cm^{-1}, which further indicates that not enough representative members of these specific conformers are included in the sample. As was the case for C_{52}, the interpolation scheme for this smaller cluster performs not as well when the descriptors are those of the REBO structures (see supplementary information), although it is worth noting that the smoothest spectrum, obtained for the flakes family, is not affected to such an extent.

5 Clustering analysis by iterative label spreading

The above analysis has shown that interpolation on a much reduced sample of representative data can reproduce the structural trends of collective but also individual spectra, especially with metric learning. It was also found that the IR spectra obtained for two of the four families are very similar, namely the pretzels and the branched structures. They were arbitrarily distinguished from one another based on the asphericity parameter, pretzel conformers being more spherical, hence having a lesser tendency for exhibitting terminating carbon chains. Here were return to this issue of structural partitioning but ignoring now such intuitive criteria and turning instead to the purely statistical analysis of iterative label spreading (ILS).[23] ILS is a semisupervised clustering method chosen for its ability to handle dense sets of points. Briefly, ILS proceeds by reordering all members of the set, labelling them each after the other, picking the next labelled member as the closest from the already labelled set. Once all members are labelled, groups of points within the set forming families are expected to be separated by peaks in the minimum distance. ILS requires a metric for distance evaluation between two arbitrary members of the set, and we naturally use the metrics optimized for the interpolation scheme for either Gaussian regression or inverse distance weighting. As in Ref. 23 we chose the starting points as the set barycenters. Fig. 7 shows the resulting partitioning for the structural database of C_{60}, as emerging from ILS with the metric optimized for Gaussian regression. Similar results are obtained with the other metric optimized for inverse distance weighting. To relate

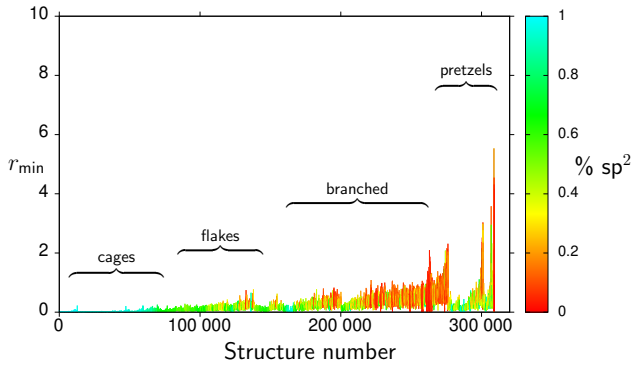

Figure 7: Clustering analysis for the isomers of C_{60} by iterative label spreading, from the metric optimized for the GR kernel. The various minima are numbered (abscissa) by increasing minimum distance (ordinate) to the set of already labelled members. Each line is colored according to its fraction of sp^2 atoms.

these partitionings with our initial sorting into the four families of cages, flakes, pretzels and branched structures, we further tag each labelled member with the value of its sp^2 fraction, which is the most discriminating of the two order parameters with the asphericity β. The corresponding graph tagged with the value of β is given as supplementary information. While there is no obvious jump in the minimum distance, the structures are clearly ordered within the four families identified empirically. The first ones lying closer to the set barycenter are the cages (highest sp^2, lowest β content) and are followed by the flakes (medium sp^2, higher β). The more disordered structures follow, first with the branched structures (lowest sp^2 values, highest β) and finally the pretzels (low to medium sp^2, low β).

These results, and especially the proximity between pretzels and branched structures, are fully consistent with their spectroscopic similarities already noted in Figs. 5 and 6 which themselves originate from the common chemical motifs of long sp^1 carbon chains. However, the ILS analysis also suggests the possibility of subgroups inside some of the four main families, particularly for the more disordered pretzels and especially branched cases. Further partitioning among these families would require additional order parameters for a more complete characterization.

6 Concluding remarks

Unravelling the relation between structure and vibrational or electronic spectra is one of the important issues in modern chemical physics in the gas phase. The interpolation scheme explored in this work attempts to address this issue from an automated statistical perspective. In this respect it is part of the broad current effort to use machine learning techniques to assist spectral determination. The interpolation scheme developed here is relevant for extracting spectroscopic trends over a statistical set of structures, and cannot claim chemical accuracy for individual conformers. Its main appeal resides in its systematic nature and in not being limited to chemically similar structures such as pure aromatics. While it can be used as a nonsupervised method, optimization of the metric was found to improve the performance especially in the cases where the structural diversity was the highest among the set. As with conventional machine learning approaches, the performance of the present method is naturally limited by the existence of sufficiently similar members in the sample on which interpolation is performed.

With respect to neural networks, the present approach involves far fewer parameters but requires a conformational sample. Its main computational interest was motivated in the need to bypass the heaviest numerical effort associated with the determination of numerous individual IR spectra, structural optimization itself being mostly achieved at a lower level of theory (here, the REBO force field).

Classification of the structures into the four families of cages, flakes, pretzels and branched conformers was also confirmed by performing independently a clustering analysis based on iterative label spreading. Here it could be interesting to compare the predictions of this method with deep learning approaches that are also becoming more widespread in physical chemistry as a classification tool.

In the future the interpolation scheme could be extended and improved along several directions. For pure computational efficiency, the k-nearest neighbor algorithm could be used to limit the number of members in the sample to be included for each newly pre-

dicted spectrum, even though the number of neighbors and the metric should be both optimized themselves self-consistently as hyperparameters. Another natural improvement could be in the selection of the interpolating sample, which was here taken from a systematic mesh of conformers based on dedicated order parameters. A Monte Carlo procedure could be introduced to further optimize the members of this sample in their capability to describe the spectra of the testing set. Alternatively, Bayesian inference could be envisaged as well for even smaller samples.

From the physical chemistry perspective, the interpolation method could also be applied to systems exhibiting even greater chemical diversity, such as hydrogen-, oxygen- or nitrogen-containing compounds, for which REBO or other reactive potentials[31,32] could be employed. The inverse problem of finding pools of structures that match observational data could also be tackled stochastically, under given constraints of available samples in terms of sizes, compositions, and individual conformations. Here also, ML techniques appear as naturally suited to such a purpose.[33]

Acknowledgments

The authors gratefully acknowledge financial support by the Agence Nationale de la Recherche (ANR) Grant No. ANR-16-CE29-0025, and the GDR EMIE 3533.

Supporting Information Available

Statistical convergence of the global IR spectra for C_{24} and C_{42} families. Interpolated spectra for the families of C_{33} with isomers described at the DFTB level. Interpolated spectra for the families of C_{33} and C_{52} with isomers described at the force field level. Iterative label spreading classification for C_{60} isomers, tagged by their asphericity β. Optimized metrics for the GR and IDW kernels.

References

(1) Kroto, H. W.; Heath, J. R.; O'Brien, S. C.; Curl, R. F.; Smalley, R. E. C_{60}: Buckminsterfullerene. *Nature* **1985**, *318*, 162-163.

(2) Curl, R. F. On the formation of the fullerenes. *Phil. Trans. R. Soc. Lond. A* **1993**, *343*, 19-32.

(3) Cami, J.; Bernard-Salas, J.; Peeters, E.; Malek, S. E. Detection of C_{60} and C_{70} in na Young Planetary Nebula. *Science*, **2010**, *329*, 1180-1182.

(4) Berné, O.; Mulas, G.; Joblin, C. Interstellar C_{60}^{+}. *Astron. Astrophys.* **2013**, *550*, L4.

(5) Dunk, P. W.; Adjizian, J. J.; Kaiser, N. K.; Quinn, J. P.; Blackney, G. T.; Ewels, C. P.; Marshall, A. G.; Kroto, H. W. Metallofullerene and fullerene formation from condensing carbon gas under conditions of stellar outflows and implication to stardust. *Proc. Natl. Acad. Sci. U.S.A.* **2013**, *110*, 18081-18086.

(6) Chuvilin, A.; Kaiser, U.; Bichoutskaia, E.; Besley, N. A.; Khlobystov, A. N. Direct transformation of graphene to fullerene. *Nature Chem.* **2010**, *2*, 450-453.

(7) Bauschlicher Jr, C. W.; Bakes, E. L. O. Infrared spectra of pilycyclic aromatic hydrocarbons (PAHs). *Chem. Phys.* **2000**, *262*, 285-291.

(8) Malloci, G.; Mulas, G.; Joblin, C. Electronic absorption spectra of PAHs up to vacuum UV. *Astron. Astrophys.* **2004**, *426*, 105–117.

(9) Calvo, F.; Basire, M.; Parneix, P. Temperature effects on the rovibrational spectra of pyrene-based PAHs. *J. Phys. Chem. A* **2011**, *115*, 8845-8854.

(10) Simon, A.; Rapacioli, M.; Lanza, M.; Joalland, B.; Spiegelman, F. Molecular dynamics simulations on $[FePAH]^{+}$ π-complexes of astrophysical interest: anharmonic infrared spectroscopy. *Phys. Chem. Chem. Phys.* **2011**, *13*, 3359-3374.

(11) Mackie, C. J.; Candian, A.; Huang, X.; Maltseva, E.; Petrignani, A.; Oomens, J.; Buma, W. J.; Lee, T. J.; Tielens, A. G. G. M. The anharmonic quartic force field infrared spectra of three polycyclic aromatic hydrocarbons: Naphthalene, anthracene, and tetracene. *J. Chem. Phys.* **2015**, *143*, 224314.

(12) Basire, M.; Parneix, P.; Pino, T.; Bréchignac, Ph.; Calvo, F. Modeling the anharmonicinfrared emission spectra of pahs: Application to the pyrene cation; in PAHs and the Universe, edited by Joblin, C. and Tielens, A. G. G. M. *EAS Publications Series* **2011**, *46*, 95.

(13) Bonnin, M. A.; Falvo, C.; Calvo, F.; Pino, T.; Parneix, P. Simulating the structural diversity of carbon clusters across the planar-to-fullerene transition. *Phys. Rev. A* **2019**, *99*, 042504.

(14) Dubosq, C.; Falvo, C.; Calvo, F.; Rapacioli, M.; Parneix, P.; Pino, T.; Simon, A. Mapping the structural diversity of C_{60} carbon clusters and their infrared spectra. *Astron. Astrophys.* **2019**, *625*, L11.

(15) Ceriotti, M.; Clementi, C.; von Lilienfeld, O. A. Machine learning meets chemical physics, *J. Chem. Phys.* **2021**, *154*, 160401.

(16) Prezhdo, O. V. Advancing Physical Chemistry with Machine Learning. *J. Phys. Chem. Lett.* **2020**, *11*, 9656-9658.

(17) Gastegger, M.; Behler, J.; Marquetand, P. Machine learning molecular dynamics for the simulation of infrared spectra. *Chem. Sci.* **2017**, *8*, 6924.

(18) Hu, W.; Ye, S.; Zhang, Y.; Li, T.; Zhang, G.; Mukamel, S.; Jiang, J. Machine learning protocol for surface-enhanced Raman spectroscopy. *J. Phys. Chem. Lett.* **2019**, *10*, 6026-6031.

(19) Härkönen, T.; Roininen, L.; Moores, M. T.; Vartiainen, E. M. Bayesian quantification for coherent Anti-Stokes Raman scattering spectroscopy. *J. Phys. Chem. B* **2020**, *124*, 7005-7012.

(20) Lam, J.; Abdul-Al, S.; Allouche, A.-R. Combining Quantum mechanics and machine-learning calculations for anharmonic corrections to vibrational frequencies. *J. Chem. Th. Comput.* **2020**, *16*, 1681-1689.

(21) Kovács, P.; Zhu, X.; Carrete, J.; Madsen, G. K. H.; Wang, Z. Machine-learning prediction of infrared spectra of Interstellar Polycyclic Aromatic Hydrocarbons, *Astrophys. J.* **2020**, *902*, 100.

(22) Gandolfi, M.; Rognoni, A.; Aleta, C.; Conte, R.; Ceotto, M.. Machine learning for vibrational spectroscopy via divide-and-conquer semiclassical intial value representation molecular dynamics with application to N-methylacetamide. *J. Chem. Phys.* **2020**, *153*, 204104.

(23) Parker, A. J.; Barnard, A. S. Selecting appropriate clustering methods for materials sciences applications of machine learning. *Adv. Theory Simul.* **2019**, *2*, 1900145.

(24) Kim, S. G.; Tománek, D. Melting the fullerenes: A molecular dynamics study. *Phys. Rev. Lett.* **1994**, *72*, 2418-2421.

(25) Brenner, D. W.; Shenderova, O. A.; Harrison, J. A.; Stuart, S. J.; Ni, B.; Sinnott, S. B. A second-generation reactive empirical bond order (REBO) potential energy expression for hydrocarbons. *J. Phys.: Cond. Matt.* **2002**, *14*, 783-802.

(26) Elstner, M.; Porezag, D.; Jungnickel, G.; Elsner, J.; Hangk, M.; Frauenheim, T.; Suhai, S.; Seifert, G. Self-consistent-charge density-functional tight-binding method for simulations of complex materials properties. *Phys. Rev. B* **1998**, *58*, 7260-7268.

(27) Jesus, W. S.; Prudente, F. V.; Marques, J. M. C.; Pereira, F. B. Modeling microsolvation clusters with electronic-structure calculations guided by analytical potentials and predictive machine learning techniques. *Phys. Chem. Chem. Phys.* **2021** *23*, 1738-1749.

(28) Heine, T.; Rapacioli, M.; Patchkovskii, S.; Frenzel, J.; Köster, A.; Calaminici, P.; Duarte, H. A.; Escalante, S.; Flores-Moreno, R.; Goursot, A. et al, deMonNano **2009** `http://demon-nano.ups-tlse.fr/`

(29) Wei, G.; Eighinger, B. E. On shape asymmetry of Gaussian molecules. *J. Chem. Phys.* **1990**, *93*, 1430-1435.

(30) Buhl, J.; Gautris, J.; Sole, R. V.; Kuntz, P.; Valverde, S.; Deneubourg, J. L.; Theraulaz, G. Efficiency and robustness in ant networks of galleries. *Eur. Phys. J. B* **2004**, *42*, 123-129.

(31) Ni, B.; Lee, K.-H.; Sinnott, S. B. A reactive empirical bond order (REBO) potential for hydrocarbon-oxygen interactions. *J. Phys.: Cond. Matt.* **2004**, *16*, 7261-7275.

(32) Senftle, T. P.; Hong, S.; Islam, M. M.; Kylasa, S. B.; Zheng, Y.; Shin, Y. K.; Junkermeier, C.; Engel-Herbert, R.; Janik, M. J.; Aktulga, H. Me. et al. The ReaxFF reactive force-field: development, applications, and future directions. *npj Comput. Mat.* **2016**, *2*, 15011.

(33) Zhou, C.; Ieritano, C.; Hopkins, W. S.. Augmenting Basin-Hopping with techniques from unsupervised machine learning: application to spectroscopy and ion mobility. *Front. Chem.* **2019**, *7*, 519.

Graphical TOC Entry