Integration of theory, simulation, artificial intelligence and virtual reality: A four-pillar approach for reconciling accuracy and interpretability in computational spectroscopy

Integration of theory, simulation, artificial intelligence and virtual reality: a four-pillar approach for reconciling accuracy and interpretability in computational spectroscopy†

Vicenzo Barone*[a], Cristina Puzzarini[b] and Giordano Mancini,[a]

The established pillars of computational spectroscopy are theory and computer based simulations. Recently, artificial intelligence and virtual reality are becoming the third and fourth pillars of an integrated strategy for the investigation of complex phenomena. The main goal of the present contribution is the description of some new perspectives for computational spectroscopy, in the framework of a strategy in which computational methodologies at the state of the art, high-performance computing, artificial intelligence and virtual reality tools are integrated with the aim of improving research throughput and achieving goals otherwise not possible. Some of the key tools (e.g., continuous molecular perception model and virtual multifrequency spectrometer) and theoretical developments (e.g., non-periodic boundaries, joint variational-perturbative models) are shortly sketched and their application illustrated by means of representative case studies taken from recent work by the authors. Some of the results presented are already well beyond the state of the art in the field of computational spectroscopy, thereby also providing a proof of concept for other research fields.

1 Introduction

Experiment or theory? Which one should be trusted more? While this dispute has characterized all scientific fields for decades, the end of the last century witnessed a reconciliation, thus leading to the widespread employment of integrated experimental-theoretical approaches. Following on from this, a challenging issue arises: how far can the quantitative prediction of experimental and technological problems be pushed? This question has puzzled theoretical and computational chemists since the birth of quantum mechanics (QM). While the Schrödinger equation, in principle, contains the solution to all problems related to the nanoscopic world, its resolution and application have posed serious issues from the very beginning, thus originating the never-ending fight between feasibility, accuracy and interpretability. The continuous struggle against the limitations faced by scientists in their reconciliation has led to the development of multi-scale and multi-resolution approaches[1–4] that have been supported and driven by considerable improvements, especially during the last decade, in hardware and software. Computer-based simulations are nowadays considered routine facilities in science and represent the second pillar of an integrated approach. The spontaneous question arising is: what is next? The instinctive answer is data science, which is related to the availability of an unprecedented amount of high quality data and represents today the third pillar of an effective integrated strategy for the investigation of complex phenomena[5]. At the same time, we are witnessing a parallel and fast growth of a fourth pillar, represented by *data visualization*: with Big Data increasing at an unprecedented pace in all sectors of research and industry, data visualization becomes crucial not only to obtain a real insight into this data deluge,[6] but also for dissemination purposes, which are nowadays an integral part of any major project.

Molecular spectroscopy is a discipline underlying all areas of chemistry because of its ability to investigate different physical-chemical aspects of molecular systems in a non-invasive way. In the field of molecular spectroscopy, experimental outcomes have been traditionally considered as the unquestionable and definitive answers. However, the improvements of conventional spectroscopic techniques and the blooming of new ones have provided new information paralleled by the challenge of interpreting raw spectral data. In this scenario, experiment cannot do without theory, with the latter providing key inputs for guiding, supporting and complementing experimental determinations and analyses.

Traditionally, the interpretation of spectroscopic data with QM

[a] *Scuola Normale Superiore, Piazza dei Cavalieri 7, I-56126 Pisa,Italy*
[b] *Dipartimento di Chimica "Giacomo Ciamician", Università di Bologna, Via F. Selmi 2, I-40126 Bologna, Italy*

is based on obtaining molecular geometries at the best possible level attainable[7,8]. The procedure usually starts from a few candidate structures and employs the local search techniques (i.e. geometry optimizations) available in electronic structure codes. In these calculations, the starting point is often the available knowledge together with chemical intuition. The growth in size of the systems to be investigated, however, implies much more complex potential energy surfaces (PESs), with a large number of minima (some of comparable stability) that makes relying on this approach prone to bias. As a consequence, wide span exploration methods that combine (step wise) preliminary cheap simulations and/or machine learning (ML) methods are becoming increasingly important in all spectroscopic studies[9–11]. However, even if all theoretical issues could be solved, computational spectroscopy would not automatically join its experimental counterpart as a routine procedure in the analysis of challenging systems as long as vis-á-vis comparisons of the raw data, i.e. the molecular spectra, cannot be performed in a systematic, robust and user-friendly way[12]. The ultimate "ingredient" to improve the vis-á-vis comparison is the unprecedented possibility of 3D representations of scientific data as well as immersive scenarios to explore macro-/microscopic environments[13] employing Immersive Virtual Reality (IVR) and, to a lower extent, Augmented Reality (or AR/VR for short) to solve the problems related to data visualization in molecular modeling.

In summary, theory, simulation, machine learning and data visualization represent the four pillars upon which a complete integration of theory and experiment in the field of molecular spectroscopy can be built. In this manuscript, we provide examples of the recent developments concerning each of the four pillars addressing specific challenges in different aggregation states

- Gas phase: rotational spectroscopy is the technique of choice. The recent introduction of laser-ablation and broadband techniques[14–17] has opened to the investigation of biomolecules and large molecular complexes in the gas phase and with high resolution. This, however, has led scientists to face the huge problem of correctly describing and characterizing large flexible systems.

- Diluted solutions: optical and chiroptical spectra are of utmost interest in this context and their interpretation requires the average over a large number of low-energy structures issuing from solvent fluctuations. In this context, the four-partner strategy alluded above is fundamental for obtaining the correct atomistic interpretation of the observation, otherwise not accessible. In particular, AR/VR can play an invaluable role in perceiving complex spectroscopic outcomes, like, e.g., magnetic fields for chiroptical spectra.

- Crystals: vibrational (infrared, IR, and Raman) spectra are the sources of the most valuable information and they are strongly tuned by environmental effects. Together with the challenges addressed above, one has to face the problem of generating and including in the computational model the correct periodic environment both from computational and graphical points of view.

To detail this "four-pillar strategy" in the next section we analyze its most distinctive features with special reference to the different steps of the pipeline and to the benefits of employing AR/VR tools. Next, different spectroscopic techniques and aggregation states are considered with reference to some recent studies performed in our laboratories. We start from semi-rigid molecules in the gas phase and then we proceed to analyze the role of environmental effects in the solid state or in innocent solvents (which can be properly described by polarizable continuum models). Subsequently, we tackle the problem of flexibility, thereby showing that the effectiveness and reliability of systematic or stochastic approaches can be overcome by methods based on the artificial intelligence paradigm. Finally, we consider the problem of environmental fluctuations, which become a central issue for non-innocent solvents (especially water) and biological environments.

2 General workflow: the four pillar strategy

In general terms, the overall strategy starts with a pre-processing stage (preferably performed in an AR/VR context) in which perception tools are used to analyze the studied system in terms of chemically-based concepts, like resonance, inductive effects, etc., to dissect intra- and inter-molecular interactions, to disentangle hard and soft degrees of freedom and, possibly, to generate –by means of unsupervised tools– a local environment closely resembling the experimental one. Next, proper computations are performed to obtain all the needed structural and spectroscopic parameters, thereby employing QM, molecular mechanics (MM), stochastic and ML approaches. Finally, in the post-processing stage, vis-á-vis comparisons between simulated and experimental spectra can be performed and the physical-chemical properties of the studied system can be explored in an AR/VR scenario.

In the Introduction, it has been mentioned that IVR (or AR/VR) is the most recent pillar added to the computational spectroscopy tools. However, we will discuss this aspect at the very beginning since the other three pillars can be envisioned as tools which may be used within an IVR context. The importance of the data-visualization pillar in the overall strategy cannot be underestimated; indeed, without graphical data representations (charts, diagrams, chemical formulas and drawings) the results of complex experimental or simulation studies would provide very limited scientific insights.[18,19] This is not surprising because one-third of neurons in the brain cortex are activated when dealing with visual information. In other words, when visual information can be immediately understood (for a chemist, most of the ball-and-stick representations can be deemed as "direct") or properly codified (via unambiguous icons, gliphs and symbols), it can be decoded faster. However, one can still argue on the importance of IVR with respect to traditional 2D visualization given the increased development efforts that AR/VR requires. In this respect, its unique features are:

- the possibility and ability of displaying data on several layers and on much larger surface with respect to a 2D screen;

- multisensorial analysis with data sonification (auditive input can perform better than visual one in certain cases, most

notably in perceiving anomalies);[20,21]

- *proprioception*: the capability to perceive own position and movement in space not from sight but from the kinesthetic inputs from the skeletal muscle apparatus.[22]

Focusing on the last point, it is noted that proprioceptive feedbacks are fundamental in approaching both real and virtual objects, inducing natural reactions that help in the interaction with the objects. This is important in general, but becomes crucial for applications such as engineering and chemistry, where the objects of study are three-dimensional and span different scales. In general terms, IVR or AR/VR can fill the gap between the human perception and the molecular world, thereby speeding up the process of understanding. As a matter of fact, VR allows one to transfer molecular systems and phenomena from their typical Angstrom/nanosecond scale to the human meter/second scale. Such a natural exploration of the nano-world permits to tap into human intuition, thus enabling the indirect abstraction processes required for relating microscopic events to their macroscopic consequences.

The benefits of immersive visualization have been recognized since the first pioneering steps of the technology[23,24]. However, the adoption of IVR or AR/VR for applications beyond the proof-of-concept stage has become commonplace only in the last few years,[25–27] owing to the introduction of cheap, off the shelf devices[28]. The situation has developed to the point that well-known remote teaching providers offer courses devoted to the use of AR/VR for scientific visualization. Furthermore, several companies (e.g., https://nanome.ai/) now offer services specifically related to AR/VR and molecular modeling.

In the last decade, we have developed, alongside traditional 2D tools for visualization and interaction such as the so-called virtual multifrequency spectrometer, VMS[29], and Ulysses[30] for the parameterization of intramolecular force fields, a fully immersive molecular graphics system denoted as Caffeine,[31–33] which was based on expensive Cave 3D systems. Although Caffeine is deficient in many typical features of the more mature 2D molecular viewers, it allowed to show the potentialities of AR/VR applied to molecular modeling.[34,35] * In the last years, the release of 3D helmets together with new AR devices and widespread adoption of game engines (mostly Unity, https://unity.com/, and Unreal, https://www.unrealengine.com) for non game purposes made the technologies of Caffeine obsolete. Using Unity, a tool to explore analytical PES and simple chemical reactions has then been realized[36]. Now, we are conveying all the experience developed over the years in a new package named *Tardis*.[37]. Tardis is an AR/VR tool for research and education, which allows to cross in a seamless way different time and space scales, thus allowing to perceive in an intuitive way the molecular world and to interact with the results of molecular simulations and QM computations. Quantitative information can also be added by exploiting the possibilities offered by AR. A flavor of these features is provided by Figure 1.

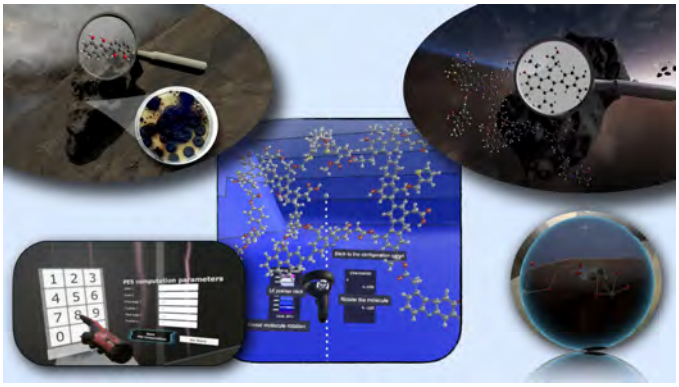

Fig. 1 Tardis, the latest AR/VR component of our ecosystem, is used for the preview in the macroscopic scenario (the molecules shown around a rock, on a meteorite and in a spherical drop) and for pre- and post-processing at the molecular level (e.g. exploration of potential energy surfaces).

As mentioned above, both the pre- and post-processing of computational spectroscopy tasks can be performed in an AR/VR context. Concerning the former (pre-processing), complex molecules are built and analyzed in an intuitive way, and suitable fragments can be selected for deeper analysis. Regarding post-processing, 2D PESs can be explored exactly in the same way as 3D maps in the macroscopic world (through a link to the Avatar code[36]) and simulated spectra can be compared to their experimental counterparts (through an interface to VMS[12]). Some features of the pre-processing stage have already been mentioned, whereas different aspects of post-processing are illustrated in later sections with reference to specific case studies.

Once AR/VR tools have led the researcher into the microscopic world, artificial intelligence comes into play for molecular perception, which is the set of methods and techniques designed to handle, in a clever and automatic way, chemical data for both cheminformatics and computational needs. The goal is to employ the least possible amount of input data (in principle just the raw chemical formula or the so-called SMILES[38]), to derive automatically all the required chemical information[39]. This task is performed by the Proxima tool[40], which determines the molecular topology. Contrary to the standard practice of employing discrete indexes (e.g. single, double and triple bonds), Proxima is able to perform a continuous perception, as recently proposed with a more limited scope in ref. 41. Furthermore, hydrogen bonds are detected between lone-pair electrons and acidic hydrogens. Once the initial topology is set, the next task is to generate all possible isomers, tautomers and, whenever chiral centers are detected, enantiomers or diastereoisomers. In particular, the detection of a tautomeric form is performed using the lone pair information: if these lone pairs are sufficiently close to hydrogens that are capable of being removed from their locations, then these are identified as tautomeric centers. In the gas phase, only neutral tautomers are generated, whereas tautomers also involving charge separation are formed if the presence of a solvent is indicated. Next, full lists of enantiomers (for chiral carbon atoms)

* The code was released under the GPL3 license: https://bitbucket.org/sns-smartlab/caffeine/src/master/

and isomers (at present for imines, alkenes and allenes) can be generated. Chiral centers are not only detected, but can also be inverted, and all possible diastereoisomers resulting from this process can be generated. Then, inter-molecular interactions can be taken into account by building either solvation shells (liquid state)[42] or periodic replicas of a given unitary cell (solid state). In the former case, a spherical drop (see Figure 1) is generated automatically for a panel of available solvents, whereas in the latter case, the cell parameters can be provided using different formats.

The outcome of the pre-processing stage is passed to the selected electronic structure code (ESC) for computing the structural, energetic and spectroscopic parameters of semi-rigid systems. In the case of flexible molecules, the exploration of soft degrees of freedom is managed by a new software including both stochastic and genetic algorithms and linked to some ESCs[10]. Finally, atomistic environments (e.g. non innocent solvents or large non-periodic biological systems) require a more articulated pipeline, which will be illustrated in a dedicated section.

Some applications of the general strategy (outlined above) will be illustrated by means of specific examples. In all cases, after the sought spectrum is obtained, the VMS tool[12] takes care of the graphical representation and vis-á-vis comparison with experiment, possibly also in an immersive AR/VR framework.

3 Semi-rigid molecules

In the present context, we will use the term semi-rigid to indicate molecular systems showing a single well-defined energy minimum (sometimes referred to as rigid) and/or characterized by potential energy surfaces containing only a few low-lying energy minima separated by high energy barriers. As a matter of fact, the study of these types of systems does not involve the problem of exploring a huge number of soft degrees of freedom, which requires instead going beyond deterministic (local optimization) methodologies.

3.1 Unraveling molecules in the gas phase: a virtual-experimental study.

Among different molecular spectroscopic techniques, rotational spectroscopy, owing to its intrinsic high resolution and high sensitivity, is one of the most powerful tools for investigating the structure and dynamics of molecules and molecular complexes in the gas phase[7,43–46]. Indeed, rotational spectra contain a wealth of accurate information on different molecular parameters, which are hardly or even not accessible from other experimental methodologies. In particular, the analysis and assignment of rotational spectra lead to the very accurate derivation of rotational constants, which are the leading terms of this spectroscopic technique and are strongly related to the structure of the molecular systems under consideration[8,43,47,48]. However, extracting such information from the rotational constants is a challenging task, often unfeasible without the support of quantum-chemical computations[7,8,43,48]. Actually, already the analysis and the assignment of rotational spectra can be a major challenge.

Most of molecules/molecular systems are asymmetric rotors,

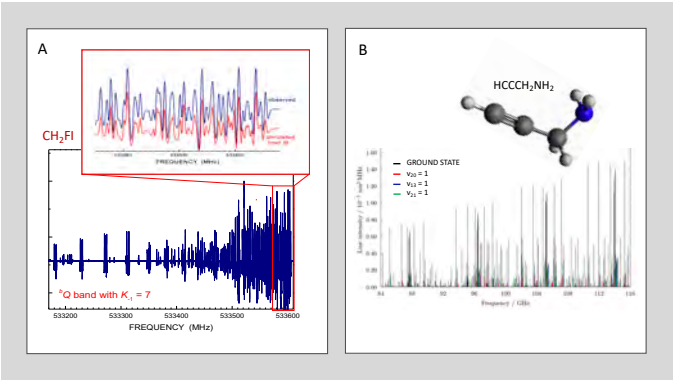

Fig. 2 A) Example of the rotational spectrum of an asymmetric rotor: the case of fluoroiodomethane[49]. B) Example of rotational spectra in the ground and excited vibrational states: the case of propargylamine[50].

thus showing complicated and dense rotational spectra, as evident in Figure 2A. The situation can be further complicated by the concomitant presence of different conformers and/or isomers, or low-lying vibrational states whose rotational spectra are well visible in a spectral recording (see Figure 2B). In such cases, the spectral analysis and assignment are strongly hampered. From a quick look at Figure 2A, the great difficulty of dealing with the rotational spectrum resulting from more species concomitantly present can be easily imagined, a task that cannot be accomplished without computational guidance whenever no previous information is available. While the practice of computing spectroscopic parameters is well established, an approach based on vis-á-vis simulation and artificial intelligence is here envisaged for taking a step forward.

3.1.1 From the structure to the rotational spectrum and back

The close correlation between the rotational spectrum and molecular structure is revealed by rotational constants, the leading terms of rotational spectroscopy. Indeed, within the rigid-rotor approximation, the rotational Hamiltonian can be written as

$$H_{rot} = \sum_i B_i J_i^2 \ , \tag{1}$$

where the sum runs over the inertial axis $i = a,b,c$ (so that, $B_a = A$), and B_i is defined as

$$B_i = \frac{\hbar^2}{2\pi I_i} \ . \tag{2}$$

where $\hbar$ is the reduced Planck constant. According to Eq. (2), rotational constants (B_i) are inversely proportional to the corresponding components of the inertia tensor $\mathbf{I}$ (I_i), which in turn is related to the molecular structure:

$$\mathbf{I} = \sum_K M_K (R_K^2 \mathbf{1} - \mathbf{R}_K \mathbf{R}_K^T) \ , \tag{3}$$

where the sum runs over all nuclei K in the molecule. Note that the masses M_K refer to atomic masses. The nuclear coordinates $\mathbf{R}_K$ needed to calculate the inertia tensor $\mathbf{I}$ are those of a rigid molecular configuration (the equilibrium structure), which is ob-

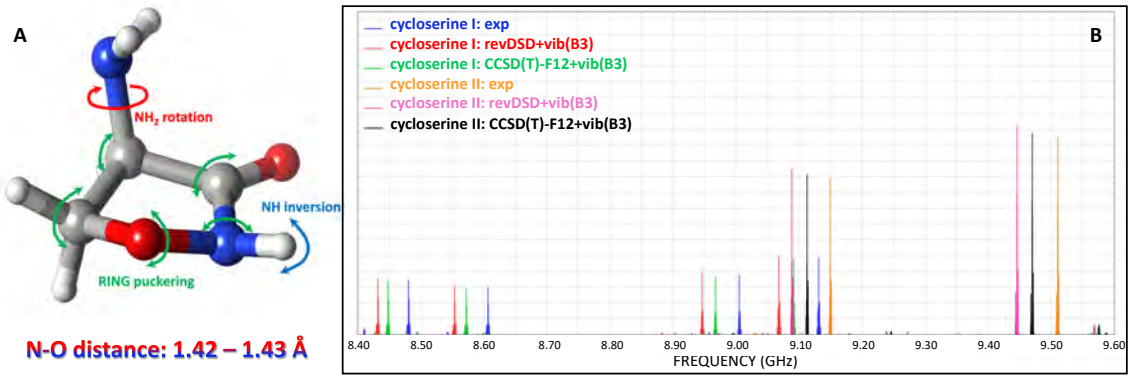

Fig. 3 A) Molecular structure of cycloserine (the form II is shown). B) Comparison between experimental and calculated rotational spectra [51].

tained in quantum-chemical calculations by means of geometry optimization.

While for a semi-rigid molecule the search of the equilibrium structure is a well-defined task, for flexible molecules the exploration of the PES can become a huge challenge, as will be addressed in Section 4. However, for a reduced number of soft degrees of freedom (e.g., hindered rotation, inversion or ring puckering), a systematic search can be still performed with conventional techniques. To describe how to go from structural determinations to the simulation of the rotational spectrum, and then how to derive geometrical information from the latter, we consider the case study offered by cycloserine [51].

Cycloserine has three soft degrees of freedom, namely the rotation of the NH_2 group, the NH inversion and the ring puckering (see Figure 3A). The latter is characterized by only two nonequivalent twisted energy minima. Each minimum can be split in two forms due to the NH inversion, thus leading to four possible structures. Then, for each of them, the NH_2 rotation can lead to three staggered conformers (two gauche and one anti structure). Overall, a total of 12 possible energy minima is thus expected for cycloserine. However, focusing on the isolated species in the gas phase, only two forms actually "survive" because all the others relax to them, as shown in ref. 51, due to low energy barriers. These two forms have been denoted as 'cycloserine I' and 'cycloserine II', the former being more stable than the latter by $\sim$204 cm^{-1}.

While a preliminary study of all conformers has been carried out using a double-hybrid density functional (see ref. 51 for details), to accurately evaluate the rotational constants to be used in the prediction of the rotational spectrum of cycloserine, the equilibrium structures of the two stable species have been optimized using the explicitly correlated CCSD(T)-F12 method [53] in conjunction with the cc-pVDZ-F12 basis set [54]. Going beyond the rigid-rotor approximation, while equilibrium rotational constants (B_i^e) are straightforwardly obtained from equilibrium geometries, to predict the experimental features, we need to move from the bottom of the well to the vibrational ground state. Within second-order vibrational perturbation theory (VPT2) [55,56], the rotational constants of the latter are given by:

$$B_i^0 = B_i^e + \Delta B_i^0 = B_i^e - \frac{1}{2}\sum_{r=1}^{N} \alpha_r^i \,, \qquad (4)$$

where the superscript "0" denotes the vibrational ground state and the sum is taken over all fundamental vibrational modes r. The α_r^i's are the so-called vibration-rotation interaction constants. From a computational point of view, anharmonic force field calculations are required to compute the vibrational corrections to rotational constants (ΔB_i^0) [43,57–60]. For cycloserine, vibrational corrections computed at the B3LYP-D3BJ/jul-cc-pVDZ level [61–63] and equilibrium rotational constants computed at the rev-DSD-PBEP86/jun-cc-pVTZ [63–65]) or CCSD(T)-F12/cc-pVDZ-F12 level, lead to the simulated spectra shown Figure 3B, thereby exploiting the VMS-ROT software [52]. In Figure 3, the comparison with the experimental counterparts is also reported. As evident, the agreement is impressive also in the case of DFT calculations. Indeed, the computational simulations of Figure 3B guided the assignment of the experimental spectrum. Furthermore, thanks to the direct access to spectroscopic databases such as the Cologne database [66], the VMS-ROT software allows for cleaning the recorded spectrum from all features due to photofragmentation products. The latter are present because cycloserine was brought in the gas phase using the laser ablation technique.

The saturated five-membered ring of cycloserine is also found in isoxazolidines, which are important scaffolds in drug-design chemistry. The most remarkable feature of this five-membered ring is the presence of adjacent nitrogen and oxygen atoms. The rotational spectroscopy study carried out in ref. 51 allowed the first accurate determination of this N-O distance. In addition to accurate computational evaluations, the so-called semi-experimental approach [57,59,67] has been used. Within this approach, the equilibrium structure is obtained by a least-squares fit of the molecular structural parameters to the equilibrium moments of inertia (I_e^i) straightforwardly derived from the semi-experimental equilibrium rotational constants (B_i^{SE}):

$$B_i^{SE} = B_i^{0,exp} - \Delta B_i^{0,calc} \,, \qquad (5)$$

with the $B_i^{0,exp}$'s being the experimental ground-state rotational constants and the $\Delta B_i^{0,calc}$'s computed according to Eq. (4). Since the rotational constants of only one isotopic species were available, the semi-experimental approach has been applied to the evaluation of the N-O bond length in cycloserine I and II (see Figure 3A), while keeping the other structural parameters fixed

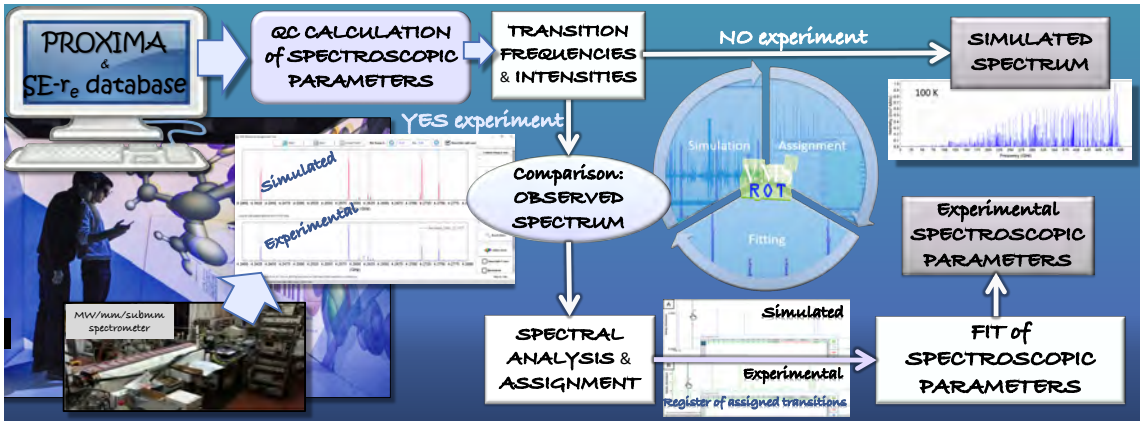

Fig. 4 The VMS-ROT [52] software integrated in a VR environment.

at the CCSD(T)-F12/cc-pVDZ-F12 values. Notably, the accuracy of the latter was verified by analogous evaluations for hydroxilamine (NH_2OH), for which a full semi-experimental equilibrium structure could be determined.

3.1.2 What is that molecule? The virtual spectrometer

As demonstrated in the previous section, rotational spectroscopy is a powerful tool for structural characterization of known molecules. However, is this spectroscopic technique also able to identify an unknown compound in a gas mixture? A possible route toward this goal is illustrated in Figure 4, with the starting point being Proxima [40] linked to VMS-ROT. This tool incorporates quantum-chemical predictions to be used as starting points for predicting and/or analyzing experiments. In this step, the user is supported by a "build-in tool" to automatically generate the input files and then run the quantum-chemical calculations. Once the spectroscopic parameters are available, the SPFIT/SPCAT program [68] (fully integrated in VMS-ROT) is used to predict or analyze the spectra, with an intuitive and user-friendly GUI (graphical user interface) being available for creating the input files and running the code. The output files are then automatically displayed (synthetic spectrum) and, if the experimental spectrum is available, the assignment and fitting (spectroscopic parameters) procedure are performed relying on the experimental-calculated difference.

As introduced in Section 2, Proxima is able to build a molecular frame starting from the atomic content knowledge. Being linked to the database of the semi-experimental equilibrium structures* (SE-re) [59,60] and that of the linear regression approach (LRA) [60], Proxima is able to provide VMS-ROT with an accurate prediction of the molecular structure at the equilibrium, and thus an accurate prediction of the equilibrium rotational constants (see Figure 5A). The computational module (VMS-Comp) of VMS-ROT then allows to obtain the missing data by running quantum-chemical calculations, with the Gaussian code [69] being fully integrated [52].

Let us consider the case where, in the gas-phase mixture, a molecular species containing a given number of H, C and O atoms is formed, e.g. $H_2C_2O_2$ (see Figure 5A). Proxima is employed to generate all possible structures for the given raw chemical formula. In the subsequent step, for each structure, Proxima makes use of the LRA information to provide a reliable prediction of the geometrical parameters. In fact, the LRA database collects the typical values of the most common bond lengths and angles as well as the associated confidence intervals [60]. While these first steps are part of a black-box procedure, the molecular structures can be further improved by relying on the SE-re database (see Figure 5B). In the molecular system under consideration, the VMS-ROT user has to identify fragments that correspond to molecules present in the latter database. The SE-re structures of these fragments are then employed to improve the geometrical parameters. A graphical representation of this procedure for structural refinement is shown in Figure 5B, where the specific case of propargyl alcohol is presented. In this molecule, two fragments can be envisaged, HCC- and -CH_2OH, whose accurate structural parameters can be obtained from the SE-re database (acetylene and methanol, respectively). For the C-C linkage, the LRA value is instead kept.

Regardless of how the molecule is built, VMS-ROT proceeds with (straightforwardly) deriving the rotational constants from the structural information. Before being able to predict the corresponding rotational spectra, quantum-chemical calculations are automatically submitted by the program itself. These aim at collecting the missing information: not only vibrational corrections to rotational constants as well as centrifugal distortion parameters, but also energetics and electric dipole moments. While the first data are required to predict the rotational spectra (i.e. the frequencies at which the rotational transitions lie), the latter two pieces of information are needed for predicting the intensities of the rotational transitions. The relative energy of the various isomers, tautomers and/or conformers allows the derivation of the Boltzmann distribution, which is -however- not self-sufficient because the intensities intrinsically depend on the dipole moment values. Once all this information is available, VMS-ROT predicts the rotational spectrum of each structure of relevance (i.e. those molecular systems whose spectrum is sufficiently intense). All spectra are then combined and displayed together (thereby ex-

* The SE equilibrium structures are available for download from `smart.sns.it`.

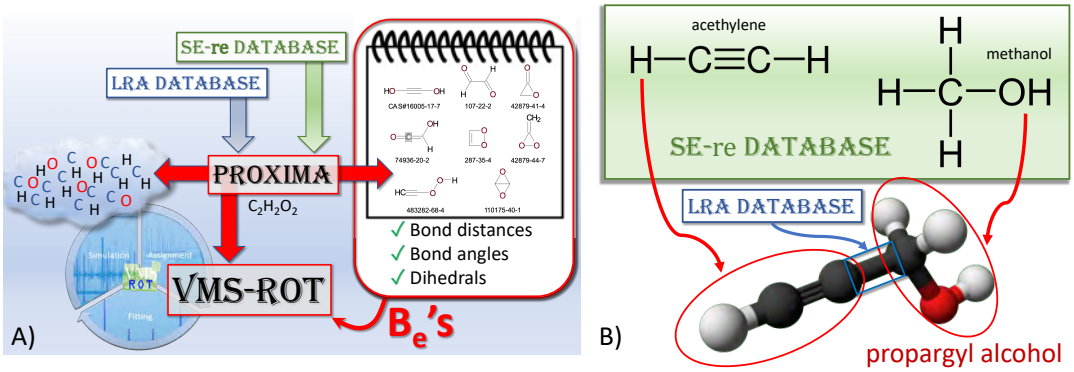

Fig. 5 Proxima: from raw chemical formula to rotational constants.

ploiting the VMS-Draw module[12]), ready to be compared with the experimental counterpart. It is from this comparison that the user can understand what are the molecular species present in the gas-phase mixture.

3.2 Crystals of semi-rigid molecules: vibrational spectroscopies

In this section, we give some hints about the extension of the strategy illustrated above for molecules in the gas phase to molecular crystals, provided that proper representations and computational tools are employed for treating periodic environments. This is illustrated with reference to part of a recent study[70] devoted to a journey across different aggregation states of creatinine (see Figure 6), the end product of nitrogen metabolism in vertebrates, following the evolution of its tautomeric equilibrium[71] by means of the distinctive features of rotational (gas phase) and vibrational (condensed phases) spectra.

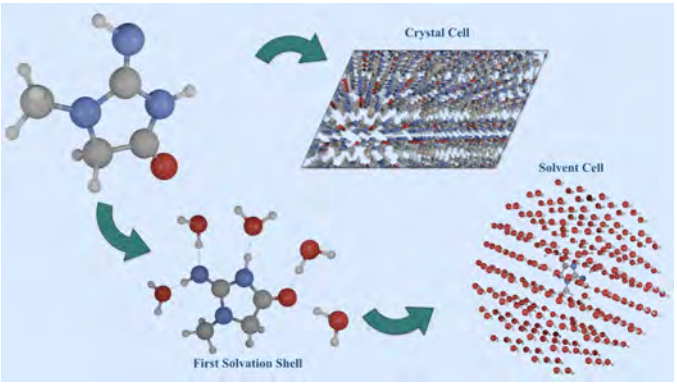

Fig. 6 Models for the imine tautomer of creatinine (I) in the gas phase, aqueous solution and crystal: starting from the isolated molecule, water is added. The first solvation-shell is explicitly incorporated, while the bulk effects are added to describe the solution. If intermolecular bonds are established among creatinine molecules, the crystal is obtained.

The first step is to build a 3D model of creatinine and to generate all the different possible tautomers and isomers either in an augmented reality framework or by more conventional 3D graphics on commodity hardware. Next, the structures of all the above species are passed to an electronic structure code (in the present context, the Gaussian software[69]) and optimized at a low-cost level of theory (here B3LYP in conjunction with a double-zeta basis set). The structures and harmonic force fields of those energy minima falling within pre-selected thresholds are recomputed at higher level (here B2PLYP-D3/jun-cc-pVTZ[63,72–75]). Accurate energies and properties are finally evaluated by means of composite schemes (here, the so-called "cheap scheme", ChS[76,77]). If required (e.g. for benchmarking purposes), the geometries of selected species can be reoptimized using more advanced methodologies (here CCSD(T)-F12/cc-pVDZ-F12[53,54]). Transition states ruling the interconversion between different isomers need also to be located if only species separated by sufficiently high barriers can be detected experimentally. However, this step has not been yet integrated in the general platform and must be performed separately.

In the case of creatinine, all computations and experiments agree in indicating that ketonic forms are considerably more stable than enolic tautomers, but the situation is more involved concerning the relative stability of the imine (I) and amine (A) ketonic forms.[71,78–81] The imine tautomer is explicitly shown in Figure 6, whereas the amine form is obtained obtained by transferring an hydrogen atom from the cyclic to the exocyclic imine group. According to state-of-the-art electronic computations, the global energy minimum of the isolated system is the imine tautomer (which exists in two isomers, E and Z, nearly degenerate in energy), with the A species being 7.7 kJ mol^{-1} less stable. The interconversion between I and A forms involves a high energy barrier (more than 200 kJ mol^{-1}), and thus both forms could be observable in rotational spectra. Indeed, this prediction was experimentally confirmed, thus demonstrating unequivocally that different tautomeric forms are present in non-negligible amounts in the gas phase.

Focusing on the solid state, the starting point is the proper and accurate description of the periodicity of the crystal. Using the parameters of the unitary cell and the symmetry group, our software can build either a cluster of predefined dimensions or the input stream for periodic computations (using, e.g., the Crystal code[82]). In the latter case, the output can be represented and managed either by three-dimensional graphics or in a AR/VR framework. The creatinine crystal belongs to the P21/c space group and it contains four molecules per unit cell. In the original X-ray study[80], which dates back to 1955, it was not possi-

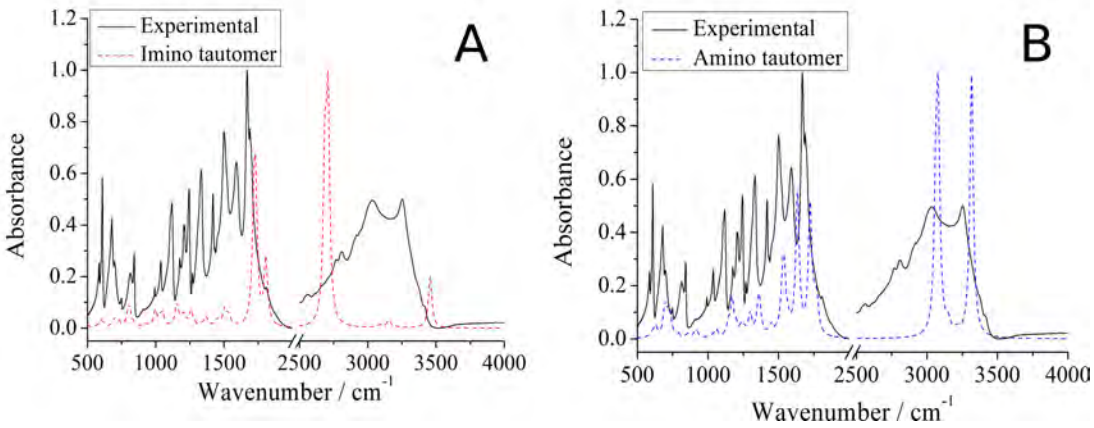

Fig. 7 The experimental IR spectrum of creatinine is compared with those computed for the imine (A) and amine (B) tautomers.

ble to identify the position of the hydrogen atoms, thus leaving open the question about the presence of the imine and/or amine tautomer. A more recent investigation[71] suggested that only the amine form is present in the crystal and this has been confirmed by periodic DFT (B3LYP) computations[70]. The latter indicate that the crystal of creatinine in the amine form is more stable than the imine counterpart by more than 100 kJ mol^{-1}. Since there are four molecules in the unit cell, the stabilization brought by a single amino molecule over the imine one can be estimated to be around 30 kJ mol^{-1}.

The results discussed above are in contrast with previous interpretations of the vibrational spectra of the solid compound in terms of both the amine and imine tautomers[81]. In order to settle the question, we have computed the IR spectra of both the imine and amine crystals and compared the results with a new experimental spectrum (see Figure 7). The spectrum computed for the amine form fits the experimental data much better than that for the imine tautomer, especially for what concerns the fingerprint region between 1500 and 1650 cm^{-1}, which shows the bands arising from the NH_2 bending (see Figure 7A). Such a mode is obviously not present in the computed spectrum of the imine species. In conclusion, all the experimental and computational evidences concur in suggesting that only the amine tautomer is present in the solid state.

4 Flexible molecules

A quantitative description of physical-chemical properties (structural, thermochemical and spectroscopic) of flexible molecules requires at least a full characterization of all low-energy minima and the transition states ruling their interconversion. Both the exhaustiveness of the sampling and the accuracy of the underlying quantum-chemical model concur to shape the final outcome, which is –however– also affected by the ill-defined role of possible error compensations. Flexibility is also accompanied by local fluctuations around the located structures, which lead to broadening effects in conventional spectroscopic experiments, whereas time-resolved experiments (not analyzed explicitly in the following) require additional considerations[83]. The crucial role played by conformational sampling in shaping the physical and spectroscopic behavior has stimulated the development of several computer codes devoted to effective PES explorations[84–90], which –in some cases– combine different methods in a single pipeline.

Exploration methods (in the present context we use the terms *PES exploration* and *conformer search* as synonyms) can be classified in terms of the underlying model used to compute the energy and its derivatives (QM, MM, QM/MM, simple geometric criteria)[91,92] as well as of the search strategy. The first aspect will be dealt when analyzing specific case studies, whereas here we discuss in some detail the available search strategies (see Figure 8). Systematic searches are the methods of choice for molecules characterized by a small number of soft degrees of freedom[93,94], as illustrated in the case of cycloserine in section 3.1. However, their cost increases steeply with the number of degrees of freedom, unless chemical insights are incorporated in the process[95–97]. In the present context, we focus our attention on the so-called *metaheuristics*[98], namely iterative generation procedures that guide subordinate heuristics by combining different concepts for *exploring* and *exploiting* a search space. Several methods of this kind have been proposed[99–109], whose common background is to mimic the efficiency of natural phenomena (e.g., collective intelligence or natural selection) to produce additional solutions with respect to those obtained by deterministic local optimizations[110].

Here, we assume that a PES exploration proceeds along the following steps:

- a (set of) solution(s) is initially built from a molecular topology;
- the energy of the candidate solution(s) is evaluated (by employing an electronic structure code);
- the solution(s) undergo some kind of coordinate transformation;
- new energies are obtained;
- from the current and previous solution(s), a new (set of) candidate solution(s) is obtained following the chosen algorithm.

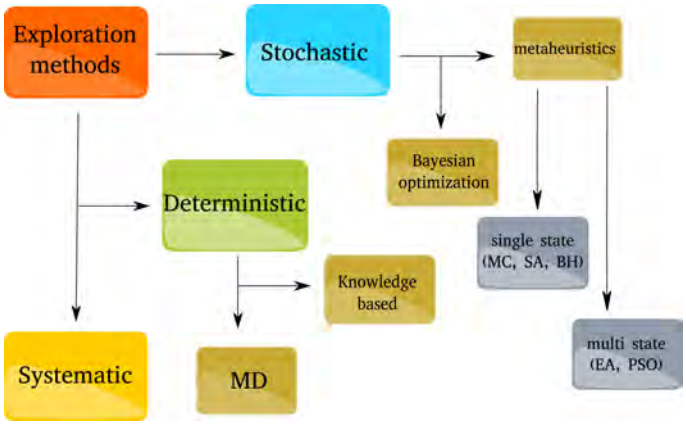

Fig. 8 A simplified taxonomy of PES exploration methods. Note that in actual applications different techniques can be combined.

The last three steps are repeated until some convergence criterion is satisfied or a predefined number of steps is exceeded. All these methods[111] try to address the problems of *exploration* (or diversification), i.e. they are able to escape local minima and avoid premature convergence, thus reaching far away regions of the search space, and *exploitation* (or intensification), i.e. they are able to improve a given (set of) solution(s) by an extensive search of its(their) neighbourhood. Noted is that for a successful conformation search, exploration is the critical factor because exploitation can be straightforwardly carried out by geometry optimization[102]. Furthermore, at each step of the optimization procedure, the more traditional single-state methods (Monte Carlo, MC; simulated annealing, SA; or basin hopping, BH) evaluate a single candidate solution at each step, whereas multi-state methods (e.g., evolutionary algorithms, EA) act on sets including several solutions. An additional option is offered by methods based on Bayesian optimization, which have recently been applied to the exploration of the conformational PES of amino acids in the gas phase[112,113]. Finally, we note that different flavors of metadynamics have been exploited and automatic procedures for the automatic selection of the underlying collective variables have been proposed[114,115].

In the following, the effectiveness of metaheuristics in dealing with flexible molecules is illustrated with reference to a milestone system, namely the neutral form of glycine (H_2NCH_2COOH) in the gas phase. Its conformational landscape is ruled by three dihedral angles and is tuned by the formation of intra-molecular hydrogen bonds. Recent systematic studies (based on state-of-the-art QM methods) agree in forecasting 8 low-energy minima[116–118], 6 of which (see Figure 9A) have been characterized by different spectroscopic techniques. The number of calculations needed to recover all eight minima (when possible) by different methods can be compared to the upper limit of $12^3 = 1728$ points needed for a systematic exploration of the three dihedral angles with a resolution of 30°. The results of a recent benchmark study[9] show that (i) stochastic (MC) explorations combined with geometry optimizations work well and (ii) cheap semi-empirical methods (such as DFTBA[119], PM7[120,121] and GFN2-xTB[122,123]) can be profitably used for the geometry optimization step. The results of the MC search are compared in Figure 10A with those issuing from SA and $(\lambda+\mu)$ EA[124] methods. The $(\lambda+\mu)$ model is a Genetic Algorithm in which each generation is actually formed by a fraction of old (μ) and new (λ) specimens; its choice is motivated in ref. 10. All QM calculations were performed carrying out a geometry optimization at the PM7 level followed by a single point energy evaluation at the B3LYP/6-31+G(d)-D3 level, which was used to calculate the MC/SA acceptance ratios or the EA fitness. It is quite apparent that the single state methods (MC and SA) are systematically outperformed by the multi state counterpart (EA) when comparing runs with a similar number of electronic structure calculations (by far the most expensive part of this type of procedure). Furthermore, a huge improvement is obtained with respect to systematic approaches.

The spectroscopic consequences of the conformational equilibrium of glycine are well illustrated with reference to IR spectra. From a general point of view, we recall that quantitative results can be obtained only taking into the proper account anharmonic contributions in the simulated spectra by means of VPT2 for small-amplitude motions and one-dimensional quasi-variational treatments for large-amplitude motions[8]. Next, inspection of Figure 9B shows that some IR features can be used as diagnostics for the presence of specific conformers (see colored regions). Finally, the overall IR spectrum is obtained by averaging the IR spectra of the various conformers weighted for their Boltzmann population. Figure 9B demonstrates that, at room temperature, the experimental IR spectrum is recovered by accounting for the three most stable conformers.

Let us now consider a larger system, neutral threonine in the gas phase, for which a full systematic search would be unfeasible. Recently, Szidarovszky and co-workers systematically explored the conformational space of threonine by performing 7776 energy evaluations at the B3LYP/6-311++G(d,p) level[125]. Seven of these conformers were subsequently refined at the MP2/6-311++G(d,p) level and experimentally identified with rotational spectroscopy[126]. In ref. 10, we used threonine as a benchmark to test and tune another EA variant, denoted as $(\lambda+\mu)$ Island Model (IM). In addition to mixing old and new structures to form the current solution at any given step, in the IM approach, new solutions can be obtained only by subsets of existing ones (called islands); limited exchange of structures between islands is allowed during the search. The searches on the threonine PES were carried out using the same two-level method employed for computing the energy of glycine, with geometry optimizations performed with the PM7 method[10]. The outcomes of different exploration strategies, shown in Figure 10B, point out the comparatively low efficiency of MC, which actually never recovers more than 39 structures out of 56, present in the reference data set. Finally, Figure 10C shows that the IM approach is the best performer, indeed being able to locate nearly all the structures with less than 1500 electronic structure computations.

5 Molecules in solution

As mentioned in the previous sections, the prediction of the spectra of flexible molecules in the gas phase is a non trivial task, indeed requiring a careful balance between feasibility and ac-

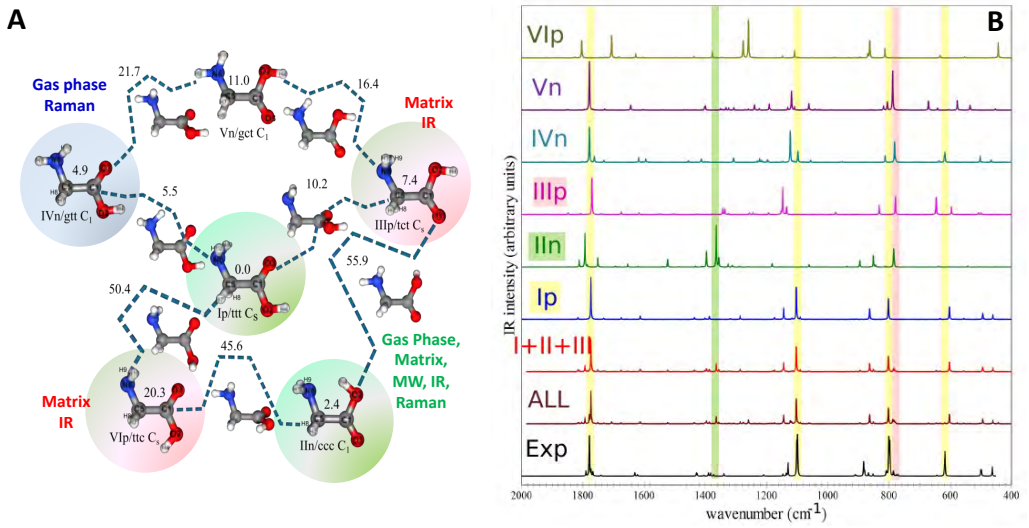

Fig. 9 A) The first six low-lying conformers of glycine. B) Comparison of the experimental IR spectrum of neutral glycine with the calculated IR spectra of the six conformers of panel A together with that resulting from a Boltzmann average of the first three species.

curacy[8]. When moving to solutions that are poorly described by continuum models, the difficulty is further exacerbated by the need of complementing the exhaustive sampling of soft internal degrees of freedom with the fluctuations of the environment[127–130]. The most widely used approaches (e.g., Quantum Mechanics-Molecular Mechanics, QM/MM) follow a multi-scale strategy in which a relatively small part of the system (e.g., the chromophore) is treated at the highest possible QM level, whereas the remaining (large) part (including remote regions of the solute and the solvent, possibly beyond the cybotactic region) is treated at a lower QM or MM level[1–3,131,132].

Broadly speaking, a general QM/MM pipeline includes the following main ingredients:

1. creation of the initial model;

2. sampling of the complete system, carried out at a QM or classical level including an appropriate representation of the environment either by a discrete (atomistic) or continuum (mean field) model or by a combination of both;

3. selection of a representative number of system configurations to perform the subsequent high-level calculations;

4. QM/MM calculations of energies and spectra for the chosen structures.

All these steps play a comparable role in determining the final accuracy of the results and, here, we sketch how such a stack may be built starting from the initial creation of the chromophore of interest and proceeding step by step to the final estimation of the property under study. The first step, the creation of an initial (set of) topology(-ies) for the system of interest (for short, solute hereafter) and a solvent box, has been discussed in section 2. Hence, in the following section, we focus our attention on points 2 to 4.

5.1 Molecular dynamics under Non Periodic Boundary Conditions

As is well known, molecular mechanics describes the PESs of molecular systems in terms of specific (and simple) functional forms for different kinds of interactions, which depend on a limited number of parameters and are trained for specific systems and conditions. This is commonly called force field (FF) and the accuracy of the classical sampling strictly depends on the quality of the underlying FF parameters[133–137]. This has stimulated a large amount of work devoted to increase the reliability of FFs, which -in recent years- has seen artificial intelligence come into play through the replacement of traditional FFs by trained artificial neural networks (ANN)[138–140], possibly guided by some underlying physical model[141]. ML comes into play also when a traditional FF needs to be expanded: the most widespread FFs in soft-matter simulations are trained for specific domain[142–144]. Hence, the inclusion of a new molecule (perhaps small with respect to the size of system) implies a new parameterization. This can be guided by ANNs using existing parameters, thus achieving the same or better accuracy at a fraction of the computational cost[145]. An additional alternative is offered by Direct Chemical Perception which soughs to assign FF parameters without the encoding forced by atom types[146]. We will not consider explicitly these possibilities in the following, making, instead, reference to specific FFs, whose parameters are optimized against quantum-chemical results for a single system[134], possibly employing genetic algorithms[136].

Given a reliable FF, classical Molecular Dynamics (MD) is perhaps the most widely used tool for sampling the phase space of molecular systems in condensed phase[147,148]. Although periodic boundary conditions (PBC) are usually employed for minimizing the error related to the finite dimensions of the simulation box, they are also prone to several artifacts[149–152]. For this reason, in the last few years, we have linked to the Proxima tool[40] (which generates the initial regular drop of predefined dimen-

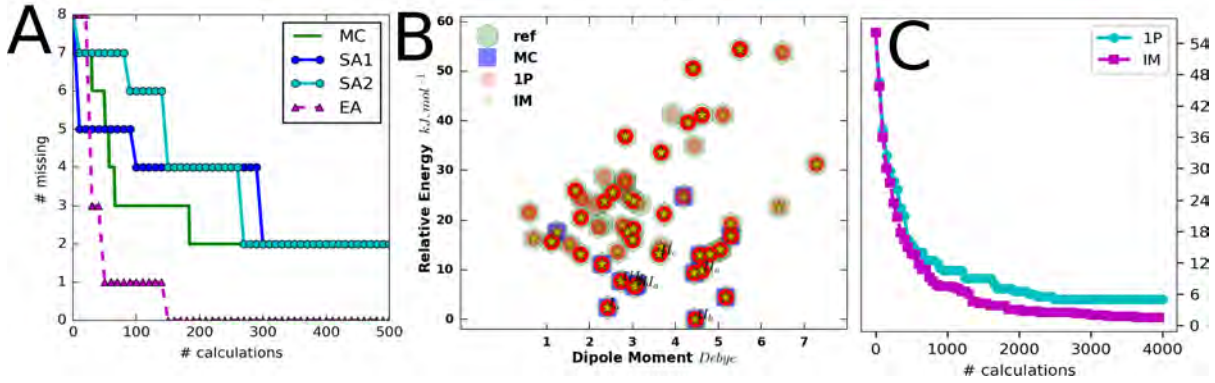

Fig. 10 A) Exploration of the glycine PES: the number of calculations performed with MC (green full line[9]), SA (blue and cyan circles) and the EA (magenta triangles) is shown. B) Exploration of the threonine PES: the 56 minima found after refinement by Szidarovszky et al.[125] (green hexagons) and those retrieved by different stochastic exploration runs are shown in a feature space constituted by dipole moment magnitude and relative stability with respect to the global energy minimum; in particular, the combination of the sets of structures found by MC (blue square), single population (1P, red small circle) and Island Model (IM, yellow star) are shown. C) Exploration of the threonine PES: average number of missing structures (within 15 kJ mol^{-1} above the global energy minimum) as a function of the number of quantum-chemical calculations for single population (cyan circles) and Island Model (magenta squares) runs.

sions containing the solute and a suitable number of explicit solvent molecules) a new MD engine[153,154] based on non periodic boundary conditions (NPBC)[155,156]. The NPBC approaches usually require a lower number of explicit solvent molecules with respect to PBC ones and provide a more natural choice for QM methods based on localized basis functions. In particular, in our GLOB (General Liquid Optimized Boundary) model[157–159], the regular drop mentioned above (usually a sphere) has elastic walls and is embedded in a polarizable continuum. The interaction potential between the explicit and continuum (mean field, MF) parts of the system is partitioned into an electrostatic (el) and a steric (st) part ($U_{MF} = U_{el} + U_{st}$). The former contribution is described by means of a polarizable continuum (e.g. CPCM[160]), whereas the latter is empirically modeled by an optimization procedure targeting the experimental bulk density. For each solvent, the final numerical representation of U_{st} is then fitted to a simple form (usually a polynomial in r or r^{-1}) and added to the library of potentials for common solvents, already available from previous studies[161,162].

Recently[163], we have added the possibility of simulating rigid-body fragments using the rotational velocity Verlet (RVV) integrator[164], which is based on a quaternion formalism[165–167]. When applicable, this kind of rigid-body simulations allows to use long time-steps (up to 4 fs), provides remarkable energy conservation[164,168], and does not suffer from constrain limitations in simulating specific moieties. The stability of the implemented integrator can be appraised by looking at Figure 11 (panels A and B), where it can be observed that, irrespective of the used time step, it outperforms the standard velocity Verlet counterpart. Figure 11C shows, instead, the effect of the inclusion of the U_{MF} term on the average density (and fluctuations) in concentric shells of equal volume in the simulation of a spherical chloroform box. As can be observed, without the U_{MF} term the system presents a density gradient that is clearly an artifact due to limited size, but the anisotropy vanishes adding U_{el} and a properly trained U_{st} term.

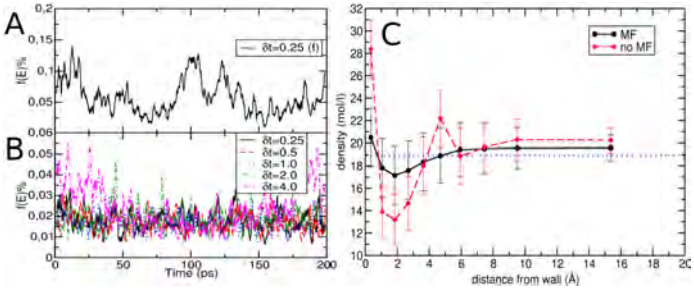

Fig. 11 NPBC simulations. A) Total energy fluctuation (percentage) for a 200 ps pure acetonitrile (simulated with flexible bonds) trajectory starting from a box equilibrated at 298K. The fluctuation is plotted as $f(E) = \sqrt{\langle (E - \langle E \rangle_t)^2 \rangle_t} / \langle E \rangle_t$ where $\langle E \rangle_t$ is the average energy. B) Total energy fluctuation for acetonitrile trajectories carried out with the RVV integrator and increasing integration time step. C) Deviations from bulk density with and without the U_{MF} term in an acetonitrile box[163].

5.2 Partition of the PES in basins and sub basins

Having generated the initial structure(s) (see section 2) and performed MD simulations to explore the phase space, the sampled configurations can be used to model spectroscopic properties. This usually means to perform some kind of convolution of the generated data. The simplest possible approach is to take frames with a fixed interval and possibly increase the density of this selection until convergence of the estimated properties. This approach, however, suffers from two major drawbacks: (i) it may require a too large number of calculations and (ii) it does not provide any insight into structure-property relationships. Therefore, general and robust procedures are needed to select frames and maximise the signal-to-noise ratio[169], with the most effective alternatives being represented by unsupervised learning (UL) techniques, such as cluster analysis[170], self-organizing maps[171] or combinatorial optimization[172]. Cluster analysis[128,173,174] (or just clustering for short) is carried out on a set of specific structural parameters and/or properties (the *feature space*) in which the similarity/dissimilarity of groups of objects is measured and,

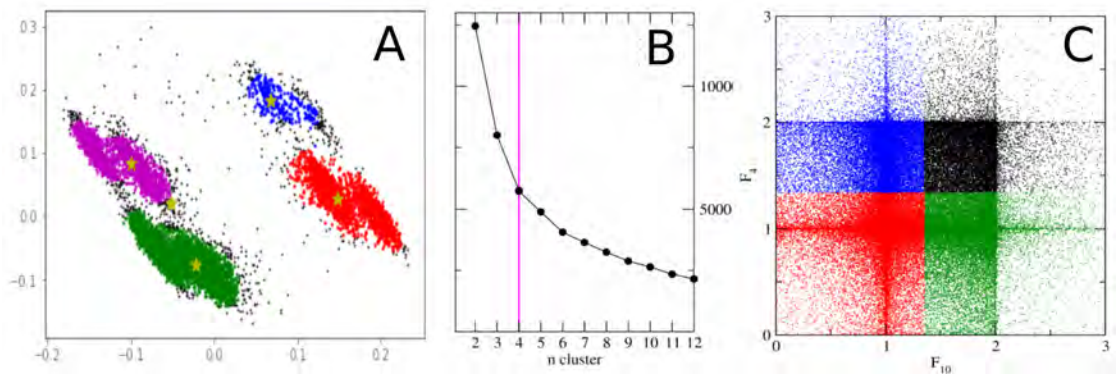

Fig. 12 A) clustering of molecular frames for tyrosine zwitterion in aqueous solution using the Principal Components of dihedral angle trajectory as feature space. Clusters are shown as dots of different colors and the corresponding centroids as stars. B) Clustering of molecular frames for uracil in aqueous solution: the sum of squared distances of cluster samples from their closest center as a function of the number of clusters is shown. C) Clustering of molecular frames for uracil in aqueous solution: the final selection of four clusters (evidenced by the red line) is employed to assign different colors to the sampled points.

hence, nearby objects (simulation frames in the present case) are assigned to the same group; the feature space is usually defined by a set of (transformed) coordinates, but it can actually include any set of properties. Since the *right number of clusters* cannot be estimated a priori *a priori*, validation criteria can (and should) be applied to select sensible parameter values for the chosen algorithm [170,175,176].

Figure 12 shows some examples of clustering simulation frames. In particular, Figure 12A shows the results of a cluster analysis carried out with a density-based method (HDBSCAN [177]) on a two-dimensional feature space obtained from a principal component analysis [178] of the dihedral angles of tyrosine zwitterion in aqueous solution [179]. Figure 12B and 12C show, instead, the use of the so-called F function (a continuous function quantifying the strength of hydrogen bonds) to partition the snapshots of a MD simulation of uracil in aqueous solution into four different clusters by means of the Partition Around Medoids method [180]. As already mentioned, the feature space is not restricted to geometrical parameters, but it can result from the combination of properties belonging to different domains and even vector fields. A specific example is given in Figure 13, which shows the clustering of the magnetically induced current density (MICD) around the metal atom of a ruthenium complex [169] and the analysis of the results in an IVR environment using the Caffeine viewer [32] (Figure 13B).

5.3 Simulation of spectra

The results of the procedures described in the previous sections provide the necessary background for the computation of different kinds of spectra in condensed phases. The general automatic procedure we have devised to accomplish this task is summarized in Figure 14 for the specific case of optical and chiroptical spectra.

The starting points are the centroids of the basins determined according to the procedures described in the previous section. Next, for each of them, the following two steps are followed:

1. The required structural, energetic and spectroscopic parameters are evaluated at the highest possible quantum-chemical level for the solute. If possibly, a reduced number of strongly bound solvent molecules is incorporated in the calculation.

2. Vibrational spectra of one or more electronic states are computed using the generalized VPT2 (GVPT2) approach to take mechanic and electric/magnetic anharmonicity into account. [56,181,182] If needed, vibrational modulation (vibronic) effects on optical or chiroptical spectra are computed using models based on the Franck-Condon principle and a continuum description (PCM) of bulk solvent effects, as described in detail in refs. 183–185.

Both time dependent (TD) and time independent (TI) formulations of this general approach are available, but in the following we will make reference only to the TI route, in which the band-shape is obtained as the sum of the individual transitions between the vibrational states of the initial and final electronic states. Either vertical or adiabatic models can be employed for semi-rigid molecules, but the presence of large amplitude internal motions favors the use of the vertical approach, with the so-called Vertical Gradient (VG) model [186] only requiring the vibrational frequencies of the initial state and the forces of the final state at the equilibrium geometry of the initial one. [183–185]

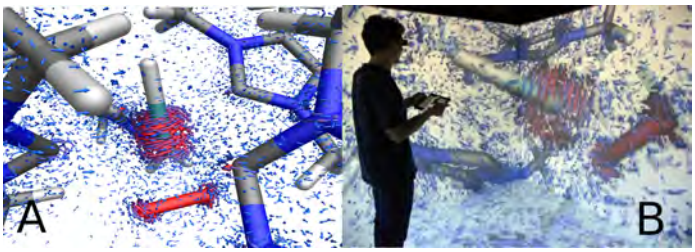

Fig. 13 A) Hierarchical clustering of the MICD in a ruthenium complex induced by a magnetic field parallel to the Ru–H bond. The colour and size of arrows are proportional to the density of vectors in a local neighbourhood. B) analysis of the results in an IVR environment.

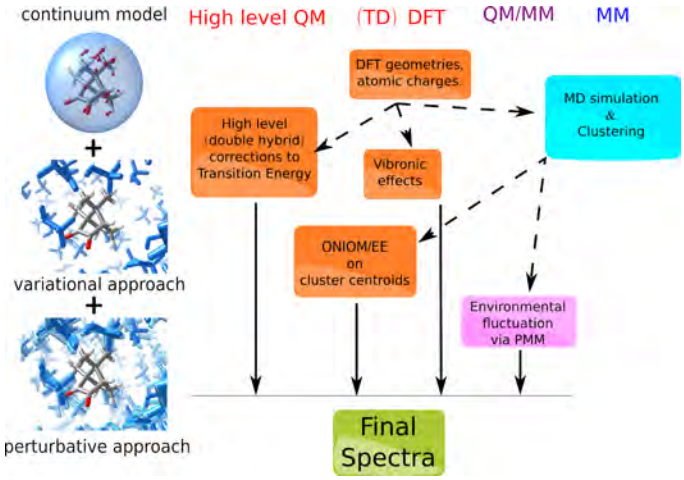

Fig. 14 Complete workflow for the simulation of spectra of flexible molecules in condensed phase.

Indeed, remarkable results can be often obtained already at this level, as shown by several recent studies.[187,188] For purposes of illustration, the VCD spectra of two bimetallic complexes in acetonitrile solution recently analyzed in our laboratory are considered (see Figure 15). The computational setup treats explicitly either the bare complex or the supermolecule formed with the two acetonitrile molecules in axial positions completing the coordination sphere of the metal, whereas bulk solvent effects are taken into account by means of the polarizable continuum model, which does not add further degrees of freedom to the analyzed system. The computation of reliable VCD spectra for these complexes is particularly demanding due to the dimension and flexibility of the system coupled to the presence of a heavy metal atom, and to the extreme sensitivity of some spectroscopic signatures to both stereo-electronic and environmental effects. In particular, a proper description of the different spectral features requires both the inclusion of explicit acetonitrile molecules in axial positions and Boltzmann averaging of low-energy structures. For instance, although the -,+,- sequence of intensities in the region between 1470 and 1600 cm^{-1} is present in most of the simulated spectra issuing from the different structures, significant differences are apparent: only exploiting the Boltzmann averaging of the individual spectra allows to approach the experimental shape (see Figure 15).

More generally, once the spectra for the selected reference structures are computed, they need to be combined with the effect of solvent motions to obtain the final spectra. To this end, we have recently devised a new model[179,189] that merges variational and perturbative approaches to get an effective yet accurate evaluation of solvatochromic shifts for a large panel of spectroscopic techniques. For a reference frame of each basin/cluster, the sought spectrum is evaluated by a variational procedure in which the QM model system is embedded in a set of (possibly polarizable[153,190]) point charges representing the environment (ONIOM/EE).[191] Then, for all the other frames of each cluster, the fluctuations of environmental effects with respect to the reference configuration are computed *a posteriori* by a perturbative approach (the perturbed matrix method, PMM[189]). In details, for each frame of each sub-trajectory, the diagonal matrix of the eigenvalues of the reference configuration is perturbed by another matrix representing the difference of the electrostatic potential between the considered frame and the reference value. Diagonalization of the overall matrix provides a set of eigenvalues (electronic states) representing the instantaneous effects of the embedding environment issuing from MD trajectories. The operator representing the variation between the perturbing effects exerted by the environment in each frame and the reference is obtained by expanding the perturbing electrostatic potential within the atomic region around each atomic center.[192]

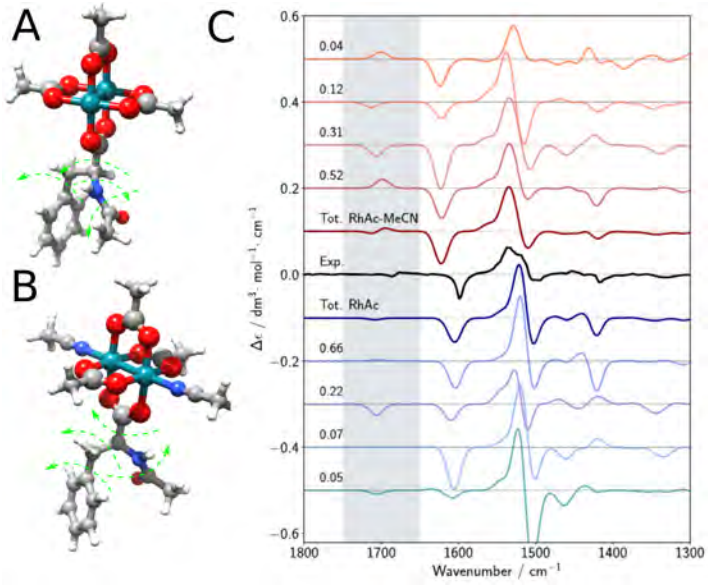

Fig. 15 The $[Rh_2(O\text{-}Phe\text{-}Ac)(O\text{-}Ac)_3]$ (Rh_2Ac) complex (A) and the (Rh_2Ac-MeCN) complex (b). C) Experimental (black lines) and theoretical harmonic VCD vibrational spectra of individual structures and Boltzmann average (bold lines). The spectra with or without explicit acetonitrile molecules are shown in shades of red or blue, respectively. Theoretical line-shapes have been convoluted by means of gaussian distribution functions with half-width at half-maximum of 10 cm^{-1}.

As an example of the reliability of this approach, Figure 16 compares the computed one photon emission (OPE, panel A) and circularly polarized luminescence (CPL, panel B) spectra of camphor in methanol solution with their experimental counterparts. In this case, the separation of the initial MD trajectory in basins was performed using a three-dimensional feature space based on the number of hydrogen bonds between the solvent and the carbonyl group of the solute. It is apparent that the simulated spectra are in remarkable agreement with experiment and reproduce, in particular, the sign alternation of the chiroptical spectrum, a feature that can be only recovered by going well beyond the standard computational models, thereby including vibronic contributions, intra-molecular flexibility and environment fluctuations. For full details, the reader is referred to ref. 194.

A second example is offered by the UV absorption spectrum of the neutral form of tyrosine in acetonitrile (Figure 16C). In this case, the underlying MD trajectory was partitioned in disjoint basins with reference to a feature space spanning the six princi-

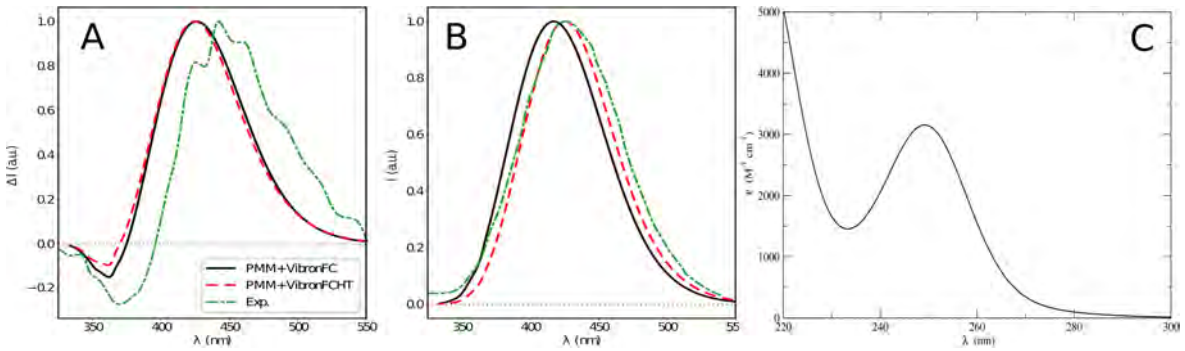

Fig. 16 OPE (A) and CPL (B) spectra of camphor in methanol, as obtained by combining the ONIOM/EE-PMM procedure outcome with VG|FC (black full line) and VG|FCHT (dashed red line) models to simulate the vibronic coupling; the corresponding experimental spectra (from ref. [193]) are also reported (green dot-dashed line). C) Tyrosine in acetonitrile electronic absorption spectrum as obtained by applying the ONIOM/EE-PMM procedure.

pal components of the dihedral angle distrbution. As a matter of fact, neither the results based on a single reference structure nor the conventional PMM approach provided reasonable spectra, whereas the ONIOM/PMM model reproduces the full ONIOM results (requiring about 200 variational computations) with just 6 ONIOM computations followed by inexpensive PMM perturbations.

6 Conclusions

A journey through the meanders of a general "four-pillar strategy" has been carried out. The different parts of the general platform under active development for implementing this strategy have been illustrated by means of some case studies involving molecules of different sizes and flexibility in different aggregation states. The common thread of our journey is molecular spectroscopy because of its capability in disclosing the underlying physical-chemical properties. However, this can be only accomplished by means of vis-á-vis comparisons of experimental and simulated spectra. While the expression 'simulated spectrum' might recall a simple diagram based on computed spectroscopic parameters, we have demonstrated that this is instead the result of a long process that starts from the raw chemical formula and proceed thanks to machine learning, quantum chemistry and virtual reality.

We are fully aware that the proposed strategy still requires other scientific and technological developments, but we hope to have provided convincing evidences that we already dispose of flexible, robust and reliable tools. Further research along the route of the "four-pillar strategy" is thus very promising and could open new exciting perspectives for the study of the complex systems and processes of current scientific and technological interest.

Conflicts of interest

There are no conflicts to declare.

Acknowledgements

The authors thank the Avogadro Staff at SNS for managing the HPC systems and Dr. Federico Lazzari, Marco Fusé and Balasubramanian Chandramouli for useful discussions.

Notes and references

1 M. Levitt, *Angew. Chem. Int. Ed.*, 2014, **53**, 10006–10018.
2 M. Karplus, *Angew. Chem. Int. Ed.*, 2014, **53**, 9992–10005.
3 A. Warshel, *Angew. Chem. Int. Ed.*, 2014, **53**, 10020–10031.
4 L. W. Chung, W. M. C. Sameera, R. Ramozzi, A. J. Page, M. Hatanaka, G. P. Petrova, T. V. Harris, X. Li, Z. Ke, F. Liu, H.-B. Li, L. Ding and K. Morokuma, *Chem. Rev.*, 2015, **115**, 5678–5796.
5 *The Fourth Paradigm: Data-intensive Scientific Discovery*, ed. A. J. G. Hey, Microsoft Research, 2009.
6 T. Hey and A. E. Trefethen, *Science*, 2005, **308**, 817–821.
7 C. Puzzarini and V. Barone, *Acc. Chem. Res.*, 2018, **51**, 548–556.
8 C. Puzzarini, J. Bloino, N. Tasinato and V. Barone, *Chem. Rev.*, 2019, **119**, 8131–8191.
9 B. Chandramouli, S. Del Galdo, M. Fusé, V. Barone and G. Mancini, *Phys. Chem. Chem. Phys.*, 2019, **21**, 19921–19934.
10 G. Mancini, M. Fusè, F. Lazzari, B. Chandramouli and V. Barone, *J. Chem. Phys.*, 2020, **153**, 124110.
11 S. Grimme, F. Bohle, A. Hansen, P. Pracht, S. Spicher and M. Stahn, *J. Phys. Chem. A*, 2021, **77**, 10.1021/acs.jpca.1c00971.
12 V. Barone, *WIREs Comp. Mol. Sci.*, 2016, **6**, 86–110.
13 T. Weymuth and M. Reiher, *Chimia*, 2021, **75**, 45–49.
14 A. Lesarri, S. Mata, J. C. Lopez and J. L. Alonso, *Rev. Sci. Instrum.*, 2003, **74**, 4799–4804.
15 G. G. Brown, B. C. Dian, K. O. Douglass, S. M. Geyer, S. T. Shipman and B. H. Pate, *Review of Scientific Instruments*, 2008, **79**, 053103.
16 S. Mata, I. Pena, C. Cabezas, J. C. López and J. L. Alonso, 2012, **280**, 91–96.
17 G. B. Park and R. W. Field, *J. Chem. Phys.*, 2016, **144**, 200901.
18 E. R. Tufte, *The visual display of quantitative information*, Graphics Press, Cheshire, Conn, 2nd edn, 2001.
19 A. Telea, *Data visualization: principles and practice*, CRC

Press, Taylor & Francis Group, Boca Raton, Second edition edn, 2015.
20 A. P. Jallouk and P. T. Cummings, *J. chem. Theory Comput.*, 2014, **10**, 1387–1394.
21 N. Sawe, C. Chafe and J. Treviño, *Frontiers in Communication*, 2020, **5**, 46.
22 L. R. Squire, *Encyclopedia of neuroscience*, Academic Elsevier, [London], 2009.
23 S. Bryson, *Communications of the ACM*, 1996, **39**, 62–71.
24 A. van Dam, D. H. Laidlaw and R. M. Simpson, *Comp. Graph.*, 2002, **26**, 535–555.
25 A. Aspuru-Guzik, R. Lindh and M. Reiher, *ACS Central Science*, 2018, **4**, 144–152.
26 G. N. Simm, A. C. Vaucher and M. Reiher, *J. Phys. Chem. A*, 2019, **123**, 385–399.
27 S. Amabilino, L. A. Bratholm, S. J. Bennie, A. C. Vaucher, M. Reiher and D. R. Glowacki, *J. Phys. Chem. A*, 2019, **123**, 4486–4499.
28 P. Cipresso, I. A. C. Giglioli, M. A. Raya and G. Riva, *Frontiers in Psychology*, 2018, **9**, 2086.
29 D. Licari, A. Baiardi, M. Biczysko, F. Egidi, C. Latouche and V. Barone, *J. Comput. Chem.*, 2015, **36**, 321–334.
30 V. Barone, I. Cacelli, N. De Mitri, D. Licari, S. Monti and G. Prampolini, *Phys. Chem. Chem. Phys.*, 2013, **15**, 3736–3751.
31 A. Salvadori, A. Brogni, G. Mancini and V. Barone, *Augmented and Virtual Reality*, Springer International Publishing, Cham, 2014, vol. 8853, pp. 333–350.
32 A. Salvadori, G. Del Frate, M. Pagliai, G. Mancini and V. Barone, *Int. J. Quantum Chem.*, 2016, **116**, 1731–1746.
33 A. Salvadori, M. Fusè, G. Mancini, S. Rampino and V. Barone, *J. Comp. Chem.*, 2018, **39**, 2607–2617.
34 S. Marks, J. E. Estevez and A. M. Connor, *Towards the Holodeck: Fully Immersive Virtual Reality Visualisation of Scientific and Engineering Data*, ACM Press, 2014, pp. 42–47.
35 M. P. Haag and M. Reiher, *Faraday Discuss.*, 2014, **169**, 89–118.
36 M. Martino, F. Salvadori, F. Lazzari, L. Paoloni, S. Nandi, G. Mancini, V. Barone and S. Rampino, *J. Chem. Theory Comput.*, 2020, **41**, 1310–1323.
37 M. Martino, *Ph.D.Thesis*, 2021.
38 H. Van der Waterbeemd, R. E. Carter, G. Grassy, H. Kubiny, Y. C. Martin, M. S. Tute and P. Willett, *Pure Appl. Chem.*, 1997, **69**, 1137–1152.
39 J.-G. Sobez and M. Reiher, *J. Chem. Inf. Mod.*, 2020, **60**, 3884–3900.
40 F. Lazzari, A. Salvadori, G. Mancini and V. Barone, *J. Chem. Inf. Model.*, 2020, **60**, 2668–2672.
41 L. Jaillet, S. Artemova and S. Redon, *J. Mol. Graph. Mod.*, 2017, **77**, 350–362.
42 G. N. Simm, P. L. Turtscher and M. Reiher, *J. Comp. Chem.*, 2020, **41**, 1144–1155.
43 C. Puzzarini, J. F. Stanton and J. Gauss, *Int. Rev. Phys. Chem.*, 2010, **29**, 273–367.
44 J. L. Alonso and J. C. López, in *Microwave Spectroscopy of Biomolecular Building Blocks*, ed. A. M. Rijs and J. Oomens, Springer International Publishing, 2015, pp. 335–401.
45 J. Wang, L. Spada, J. Chen, S. Gao, S. Alessandrini, G. Feng, C. Puzzarini, Q. Gou, J.-U. Grabow and V. Barone, *Angew. Chem. Int. Ed.*, 2019, **58**, 13935–13941.
46 F. Xie, M. Fusè, A. S. Hazrah, W. Jaeger, V. Barone and Y. Xu, *Angew. Chem. Int. Ed.*, 2020, **59**, 22427–22430.
47 W. Gordy and R. L. Cook, *Microwave Molecular Spectra*, Wiley, 1984.
48 C. Puzzarini, *Phys. Chem. Chem. Phys.*, 2013, **15**, 6595–6607.
49 C. Puzzarini, G. Cazzoli, J. C. López, J. L. Alonso, A. Baldacci, A. Baldan, S. Stopkowicz, L. Cheng and J. Gauss, *J. Chem. Phys.*, 2011, **134**, 174312.
50 Degli Esposti, C., Dore, L., Puzzarini, C., Biczysko, M., Bloino, J., Bizzocchi, L., Lattanzi, V. and Grabow, J.-U., *Astronomy & Astrophysics*, 2018, **615**, A176.
51 E. R. Alonso, M. Fusé, I. León, C. Puzzarini, J. L. Alonso and V. Barone, *J. Phys. Chem. A*, 2021, **0**, null.
52 D. Licari, N. Tasinato, L. Spada, C. Puzzarini and V. Barone, *J. Chem. Theory Comput.*, 2017, **13**, 4382–4396.
53 G. Knizia, T. B. Adler and H.-J. Werner, *J. Chem. Phys.*, 2009, **130**, 054104.
54 K. A. Peterson, T. B. Adler and H.-J. Werner, *J. Chem. Phys.*, 2008, **128**, 084102.
55 I. M. Mills, *Vibration-Rotation Structure in Asymmetric-and Symmetric-Top Molecules in Molecular Spectroscopy: Modern Research*, 1972, **1**, 115.
56 V. Barone, *J. Chem. Phys.*, 2005, **122**, 014108.
57 F. Pawłowski, P. Jørgensen, J. Olsen, F. Hegelund, T. Helgaker, J. Gauss, K. L. Bak and J. F. Stanton, *J. Chem. Phys.*, 2002, **116**, 6482–6496.
58 C. Puzzarini, J. Heckert and J. Gauss, *J. Chem. Phys.*, 2008, **128**, 194108.
59 M. Piccardo, E. Penocchio, C. Puzzarini, M. Biczysko and V. Barone, *J. Phys. Chem. A*, 2015, **119**, 2058–2082.
60 E. Penocchio, M. Piccardo and V. Barone, *J. Chem. Theory Comput.*, 2015, **11**, 4689–4707.
61 A. D. Becke, *Phys. Rev. A*, 1988, **38**, 3098–3100.
62 S. Grimme, S. Ehrlich and L. Goerigk, *J. Comput. Chem.*, 2011, **32**, 1456–1465.
63 E. Papajak, J. Zheng, X. Xu, H. R. Leverentz and D. G. Truhlar, *J. Chem. Theory Comput.*, 2011, **7**, 3027–3034.
64 G. Santra, N. Sylvetsky and J. M. L. Martin, *J. Phys. Chem. A*, 2019, **123**, 5129–5143.
65 T. H. Dunning, *J. Chem. Phys.*, 1989, **90**, 1007–1023.
66 H. S. Müller, F. Schlöder, J. Stutzki and G. Winnewisser, *J. Mol. Struct.*, 2005, **742**, 215–227.
67 P. Pulay, W. Meyer and J. E. Boggs, *J. Chem. Phys.*, 1978, **68**, 5077–5085.
68 H. M. Pickett, *J. Mol. Spectrosc.*, 1991, **148**, 371–377.
69 M. J. Frisch, G. W. Trucks, H. B. Schlegel, G. E. Scuseria,

M. A. Robb, J. R. Cheeseman, G. Scalmani, V. Barone, G. A. Petersson, H. Nakatsuji, X. Li, M. Caricato, A. V. Marenich, J. Bloino, B. G. Janesko, R. Gomperts, B. Mennucci, H. P. Hratchian, J. V. Ortiz, A. F. Izmaylov, J. L. Sonnenberg, D. Williams-Young, F. Ding, F. Lipparini, F. Egidi, J. Goings, B. Peng, A. Petrone, T. Henderson, D. Ranasinghe, V. G. Zakrzewski, J. Gao, N. Rega, G. Zheng, W. Liang, M. Hada, M. Ehara, K. Toyota, R. Fukuda, J. Hasegawa, M. Ishida, T. Nakajima, Y. Honda, O. Kitao, H. Nakai, T. Vreven, K. Throssell, J. A. Montgomery, Jr., J. E. Peralta, F. Ogliaro, M. J. Bearpark, J. J. Heyd, E. N. Brothers, K. N. Kudin, V. N. Staroverov, T. A. Keith, R. Kobayashi, J. Normand, K. Raghavachari, A. P. Rendell, J. C. Burant, S. S. Iyengar, J. Tomasi, M. Cossi, J. M. Millam, M. Klene, C. Adamo, R. Cammi, J. W. Ochterski, R. L. Martin, K. Morokuma, O. Farkas, J. B. Foresman and D. J. Fox, *Gaussian 16 Revision C.01*, 2016.

70 I. Leon, N. Tasinato, L. Spada, E. R. Alonso, S. Mata, A. Balbi, C. Puzzarini, J. L. Alonso and V. Barone, *Chem. Plus Chem.*, 2021.

71 W. Bell, T., Z. Hou, Y. Luo, B. Drew, G., E. Chapoteau, P. Czech, B. and A. Kumar, *Science*, 1995, **269**, 671.

72 S. Grimme, *J. Chem. Phys.*, 2006, **124**, 034108.

73 M. Biczysko, P. Panek, G. Scalmani, J. Bloino and V. Barone, *J. Chem. Theory Comput.*, 2010, **6**, 2115–2125.

74 S. Grimme, J. Antony, S. Ehrlich and H. Krieg, *J. Chem. Phys.*, 2010, **132**, 154104.

75 S. Grimme, S. Ehrlich and L. Goerigk, *J. Comput. Chem.*, 2011, **32**, 1456–1465.

76 C. Puzzarini and V. Barone, *Phys. Chem. Chem. Phys.*, 2011, **13**, 7189–7197.

77 S. Alessandrini, V. Barone and C. Puzzarini, *J. Chem. Theory Comput.*, 2019, **16**, 988–1006.

78 S. Craw, J., P. Greatbanks, S., H. Hillier, I., J. Harrison, M. and N. A. Burton, *J. Chem. Phys.*, 1997, **106**, 6612.

79 Y. Valadbeigi, V. Ilbeigi and M. Tabrizchi, *Comp. Theor. Chem.*, 2015, **1061**, 27.

80 S. Du Pre and H. Mendel, *Acta Cryst.*, 1955, **8**, 311.

81 K. Vikram, S. Mishra, S. K. Srivastava and R. K. Singh, *J. Mol. Struct.*, 2012, **1012**, 141–150.

82 R. Dovesi, A. Erba, R. Orlando, C. M. Zicovich-Wilson, B. Civalleri, L. Maschio, M. Rerat, S. Casassa, J. L. Baima, S. Salustro and B. Kirtman, *WIREs Comp. Mol. Sci.*, 2018, **8**, e1360.

83 M. Saunders, K. N. Houk, Y. D. Wu, W. C. Still, M. Lipton, G. Chang and W. C. Guida, *J. Am. Chem. Soc.*, 1990, **112**, 1419–1427.

84 M. J. Vainio and M. S. Johnson, *J. Chem. Inf. Mod.*, 2007, **47**, 2462–2474.

85 J. S. Puranen, M. J. Vainio and M. S. Johnson, *J. Comp. Chem.*, 2009, **31**, 1722–1732.

86 N. M. O'Boyle, T. Vandermeersch, C. J. Flynn, A. R. Maguire and G. R. Hutchison, *J. Cheminform.*, 2011, **3**, 8.

87 H. Goto, T. Takahashi, Y. Takata, K. Ohta and U. Nagashima, *Nanotech*, 2003, **1**, 1–9.

88 M. A. Miteva, F. Guyon and P. Tuffery, *Nucleic Acids Research*, 2010, **38**, W622–W627.

89 P. C. D. Hawkins, A. G. Skillman, G. L. Warren, B. A. Ellingson and M. T. Stahl, *J. Chem. Inf. Mod.*, 2010, **50**, 572–584.

90 P. Pracht, F. Bohle and S. Grimme, *Phys. Chem. Chem. Phys.*, 2020, **22**, 7169–7192.

91 D. K. Agrafiotis, A. C. Gibbs, F. Zhu, S. Izrailev and E. Martin, *J. Chem. Inf. Model.*, 2007, **47**, 1067–1086.

92 S. Riniker and G. A. Landrum, *J. Chem. Inf. Model.*, 2015, **55**, 2562–2574.

93 D. D. Beusen, E. Berkley Shands, S. Karasek, G. R. Marshall and R. A. Dammkoehler, *J. Mol. Struct. THEOCHEM*, 1996, **370**, 157–171.

94 C. M. Rienstra, L. Tucker-Kellogg, C. P. Jaroniec, M. Hohwy, B. Reif, M. T. McMahon, B. Tidor, T. Lozano-Perez and R. G. Griffin, *Proc. Nat. Acad. Sci.*, 2002, **99**, 10260–10265.

95 J. T. Ngo and M. Karplus, *J. Am. Chem. Soc.*, 1997, **119**, 5657–5667.

96 J. V. Kildgaard, K. V. Mikkelsen, M. Bilde and J. Elm, *J. Phys. Chem. A*, 2018, **122**, 5026–5036.

97 D. Ferro-Costas and A. Fernández-Ramos, *Frontiers in Chemistry*, 2020, **8**, 16.

98 J. Brownlee, *Clever algorithms: nature-inspired programming recipes*, LuLu.com, s.l., Revision 2 edn, 2012.

99 N. Nair and J. M. Goodman, *J. Chem. Inf. Comput. Sci.*, 1998, **38**, 317–320.

100 J. L. Llanio-Trujillo, J. M. C. Marques and F. B. Pereira, *J. Phys. Chem. A*, 2011, **115**, 2130–2138.

101 Y. Sakae, T. Hiroyasu, M. Miki and Y. Okamoto, *J. Comput. Chem.*, 2011, **32**, 1353–1360.

102 Z. E. Brain and M. A. Addicoat, *J. Chem. Phys.*, 2011, **135**, 174106.

103 J. Zhao, R. Shi, L. Sai, X. Huang and Y. Su, *Mol. Simul.*, 2016, **42**, 809–819.

104 H. A. A. Bahamish, R. Abdullah and R. A. Salam, Second Asia International Conference on Modelling and Simulation, 2008, pp. 911–916.

105 H. A. A. Bahamish, R. Abdullah and R. A. Salam, Third Asia International Conference on Modelling and Simulation, 2009, pp. 258–263.

106 G. Zhang, X.-G. Zhou, X.-F. Yu, X.-H. Hao and L. Yu, *IEEE/ACM Trans. Comput. Biol. Bioinf.*, 2016, 1.

107 Y. Guo and Y. Wang, Systems Biology (ISB), 2013 7th International Conference on, 2013, pp. 119–122.

108 F. Daecyaert, M. De Jonge, L. Koymans and M. Vinkers, *J. Comput. Chem.*, 2007, **28**, 890–898.

109 G.-F. Hao, W.-F. Xu, S.-G. Yang and G.-F. Yang, *Scientific Reports*, 2015, **5**,.

110 F. Glover and M. Laguna, *Tabu search*, Kluwer Academic Publishers, 1997.

111 D. H. Wolpert and W. G. Macready, *IEEE Trans. Evol. Comput.*, 1997, **1**, 67–82.

112 L. Chan, G. R. Hutchison and G. M. Morris, *J. Cheminform.*,

2019, **11**, 19.
113 L. Fang, E. Makkonen, M. Todorović, P. Rinke and X. Chen, *J. Chem. Theory Comput.*, 2021, acs.jctc.0c00648.
114 D. Polino and M. Parrinello, *J. Phys. Chem. A*, 2015, **119**, 978–989.
115 R. Galvelis and Y. Sugita, *J. Chem. Theory Comput.*, 2017, **13**, 2489–2500.
116 V. Barone, C. Adamo and F. Lelj, *J. Chem. Phys.*, 1995, **102**, 364–370.
117 V. Barone, M. Biczysko, J. Bloino and C. Puzzarini, *Phys. Chem. Chem. Phys.*, 2013, **15**, 1358–1363.
118 C. Shu, Z. Jiang and M. Biczysko, *J. Mol. Mod.*, 2020, **26**, 129.
119 D. Porezag, T. Frauenheim, T. Köhler, G. Seifert and R. Kaschner, *Phys. Rev. B*, 1995, **51**, 12947–12957.
120 J. J. P. Stewart, *J. Mol. Mod.*, 2007, **13**, 1173–1213.
121 J. J. P. Stewart, *J. Mol. Mod.*, 2013, **19**, 1–32.
122 C. Bannwarth, S. Ehlert and S. Grimme, *J. Chem. Theory Comput.*, 2019, **15**, 1652–1671.
123 P. Pracht, F. Bohle and S. Grimme, *Phys. Chem. Chem. Phys.*, 2020, 10.1039.C9CP06869D.
124 *Evolutionary computation*, ed. D. B. Fogel, T. Bäck and Z. Michalewicz, Institute of Physics Publishing, Bristol ; Philadelphia, 2000.
125 T. Szidarovszky, G. Czakó and A. G. Császár, *Mol. Phys.*, 2009, **107**, 761–775.
126 J. L. Alonso, C. Pérez, M. Eugenia Sanz, J. C. López and S. Blanco, *Phys. Chem. Chem. Phys.*, 2009, **11**, 617–627.
127 V. Barone, *Computational strategies for spectroscopy: from small molecules to nano systems*, Wiley, 2012, p. 594.
128 M. Rosa, M. Micciarelli, A. Laio and S. Baroni, *J. Chem. Theory Comput.*, 2016, **12**, 4385–4389.
129 U. N. Morzan, D. J. Alonso de Armiño, N. O. Foglia, F. Ramírez, M. C. González Lebrero, D. A. Scherlis and D. A. Estrin, *Chem. Rev.*, 2018, **118**, 4071–4113.
130 S. Grimme and P. R. Schreiner, *Angew. Chem. Int. Ed.*, 2018, **57**, 4170–4176.
131 G. A. Cisneros, M. Karttunen, P. Ren and C. Sagui, *Chem. Rev.*, 2014, **114**, 779–814.
132 B. Chandramouli, S. Del Galdo, G. Mancini, N. Tasinato and V. Barone, *Biopol.*, 2018, **109**, e23109.
133 I. Cacelli and G. Prampolini, *J. Chem. Theory Comput.*, 2007, **3**, 1803–1817.
134 V. Barone, I. Cacelli, N. De Mitri, D. Licari, S. Monti and G. Prampolini, *Phys. Chem. Chem. Phys.*, 2013, **15**, 3736–3751.
135 R. M. Betz and R. C. Walker, *J. Comput. Chem.*, 2015, **36**, 79–87.
136 F. Fracchia, G. Del Frate, G. Mancini, W. Rocchia and V. Barone, *J. Chem. Theory Comput.*, 2018, **14**, 255–273.
137 Y. Li, H. Li, F. C. Pickard, B. Narayanan, F. G. Sen, M. K. Y. Chan, S. K. R. S. Sankaranarayanan, B. R. Brooks and B. Roux, *J. Chem. Theory Comput.*, 2017, **13**, 4492–4503.
138 J. Behler and M. Parrinello, *Phys. Rev. Lett.*, 2007, **98**, 146401.
139 M. Gastegger, J. Behler and P. Marquetand, *Chem. Sci.*, 2017, **8**, 6924–6935.
140 J. S. Smith, O. Isayev and A. E. Roitberg, *Chem. Sci.*, 2017, **8**, 3192–3203.
141 T. Zubatiuk and O. Isayev, *Acc. Chem. Res.*, 2021, acs.accounts.0c00868.
142 J. A. Maier, C. Martinez, K. Kasavajhala, L. Wickstrom, K. E. Hauser and C. Simmerling, *J. Chem. Theory Comput.*, 2015, **11**, 3696–3713.
143 K. Vanommeslaeghe, E. Hatcher, C. Acharya, S. Kundu, S. Zhong, J. Shim, E. Darian, O. Guvench, P. Lopes, I. Vorobyov and A. D. Mackerell, *J. Comput. Chem.*, 2009, NA–NA.
144 R. T. Cygan, J.-J. Liang and A. G. Kalinichev, *J. Phys. Chem. B*, 2004, **108**, 1255–1266.
145 R. Galvelis, S. Doerr, J. M. Damas, M. J. Harvey and G. De Fabritiis, *J. Chem. Inf. Model.*, 2019, **59**, 3485–3493.
146 D. L. Mobley, C. C. Bannan, A. Rizzi, C. I. Bayly, J. D. Chodera, V. T. Lim, N. M. Lim, K. A. Beauchamp, M. R. Shirts, M. K. Gilson and P. K. Eastman, *bioRxiv*, 2018.
147 W. F. van Gunsteren, D. Bakowies, R. Baron, I. Chandrasekhar, M. Christen, X. Daura, P. Gee, D. P. Geerke, A. Glättli, P. H. Hünenberger, M. A. Kastenholz, C. Oostenbrink, M. Schenk, D. Trzesniak, N. F. A. van der Vegt and H. B. Yu, *Angew. Chem. Int. Ed.*, 2006, **45**, 4064–4092.
148 J. Wong-Ekkabut and M. Karttunen, *Biochim. Biophys. Acta*, 2016, **1858**, 2529–2538.
149 V. Lounnas, S. K. Lüdemann and R. C. Wade, *Biophys. Chem.*, 1999, **78**, 157–182.
150 P. H. Hunenberger and J. A. McCammon, *J. Chem. Phys.*, 1999, **110**, 1856.
151 I.-C. Yeh and G. Hummer, *J. Phys. Chem. B*, 2004, **108**, 15873–15879.
152 M. M. Reif, V. Kräutler, M. A. Kastenholz, X. Daura and P. H. Hünenberger, *J. Phys. Chem. B*, 2009, **113**, 3112–3128.
153 F. Lipparini and V. Barone, *J. Chem. Theory Comput.*, 2011, **7**, 3711–3724.
154 G. Mancini, G. Brancato and V. Barone, *J. Chem. Theory Comput.*, 2014, **10**, 1150–1163.
155 G. Petraglio, M. Ceccarelli and M. Parrinello, *J. Chem. Phys.*, 2005, **123**, 044103.
156 Z. G. Huang, Z. N. Guo, T. M. Yue and K. C. Chan, *EPL (Europhysics Letters)*, 2010, **92**, 50007.
157 G. Brancato, A. Di Nola, V. Barone and A. Amadei, *J. Chem. Phys.*, 2005, **122**, 154109.
158 G. Brancato, V. Barone and N. Rega, *Theor. Chem. Acc.*, 2007, **117**, 1001–1015.
159 G. Brancato, N. Rega and V. Barone, *Phys. Chem. Chem. Phys.*, 2010, **12**, 10736–10739.
160 M. Cossi, N. Rega, G. Scalmani and V. Barone, *J. Comp. Chem.*, 2003, **24**, 669–681.
161 N. Rega, G. Brancato, A. Petrone, P. Caruso and V. Barone, *J. Chem. Phys.*, 2011, **134**, 074504.

162 G. Mancini, G. Brancato, B. Chandramouli and V. Barone, *Chem. Phys. Lett.*, 2015, **625**, 186–192.

163 G. Mancini, S. Del Galdo, B. Chandramouli, M. Pagliai and V. Barone, *J. Chem. Theory Comput.*, 2020, **16**, 5747–5761.

164 D. Rozmanov and P. G. Kusalik, *Phys. Rev. E*, 2010, **81**, 056706.

165 D. J. Evans, *Mol. Phys.*, 1977, **34**, 317–325.

166 D. C. Rapaport, *J. Comp. Phys.*, 1985, **60**, 306–314.

167 C. F. Karney, *J. Mol. Graph. Mod.*, 2007, **25**, 595–604.

168 I. P. Omelyan, *Comp. Phys.*, 1998, **12**, 97–103.

169 D. Licari, M. Fusè, A. Salvadori, N. Tasinato, M. Mendolicchio, G. Mancini and V. Barone, *Phys. Chem. Chem. Phys.*, 2018, **7**, 3711–3724.

170 J. Han and M. Kamber, *Data mining: concepts and techniques*, Elsevier, Burlington, MA, 3rd edn, 2011.

171 D. Fraccalvieri, A. Pandini, F. Stella and L. Bonati, *BMC bioinformatics*, 2011, **12**, 158.

172 T. A. Feo and M. G. C. Resende, *J. Glob. Opt.*, 1995, **6**, 109–133.

173 A. E. Torda and W. F. van Gunsteren, *J. Comp. Chem.*, 1994, **15**, 1331–1340.

174 J. Shao, S. W. Tanner, N. Thompson and T. E. Cheatham, *J. Chem. Theory Comput.*, 2007, **3**, 2312–2334.

175 G. Mancini and C. Zazza, *PLOS ONE*, 2015, **10**, e0137075.

176 M. Macchiagodena, G. Mancini, M. Pagliai, G. Cardini and V. Barone, *Int. J. Quantum Chem.*, 2017, e25554.

177 R. J. G. B. Campello, D. Moulavi and J. Sander, *Advances in Knowledge Discovery and Data Mining: 17th Pacific-Asia Conference, PAKDD 2013, Gold Coast, Australia, April 14-17, 2013, Proceedings, Part II*, Springer Berlin Heidelberg, Berlin, Heidelberg, 2013, pp. 160–172.

178 T. Hastie, R. Tibshirani and J. Friedman, *The Elements of Statistical Learning*, Springer New York, 2009.

179 S. Del Galdo, B. Chandramouli, G. Mancini and V. Barone, *J. Chem. Theory Comput.*, 2019, **15**, 3170–3184.

180 L. Kaufmann and P. Rousseeuw, *Data Analysis based on the L1-Norm and Related Methods*, 1987, 405–416.

181 J. Bloino, M. Biczysko and V. Barone, *J. Chem. Theory Comput.*, 2012, **8**, 1015–1036.

182 V. Barone, M. Biczysko and J. Bloino, *Phys. Chem. Chem. Phys.*, 2014, **16**, 1759–1787.

183 V. Barone, J. Bloino, M. Biczysko and F. Santoro, *J. Chem. Theory Comput.*, 2009, **5**, 540–554.

184 J. Bloino, M. Biczysko, F. Santoro and V. Barone, *J. Chem. Theory Comput.*, 2010, **6**, 1256–1274.

185 J. Bloino, A. Baiardi and M. Biczysko, *Int. J. Quant. Chem.*, 2016, **116**, 1543–1574.

186 A. Baiardi, J. Bloino and V. Barone, *J. Chem. Phys.*, 2016, **144**, 084114.

187 T. T. Giovannini, P. Lafiosca, B. Chandramouli, V. Barone and C. Cappelli, *J. Chem. Phys.*, 2019, **150**, 124102.

188 S. A. Katsyuba, S. Spicher, T. P. Gerasimova and S. Grimme, *J. Phys. Chem. B*, 2020, **124**, 6664–6670.

189 S. Del Galdo, M. Fusé and V. Barone, *J. Chem. Theory Comput.*, 2020, **16**, 3294–3306.

190 I. Carnimeo, C. Cappelli and V. Barone, *J. Comput. Chem.*, 2015, **36**, 2271–2290.

191 T. Vreven, K. S. Byun, I. Komàromi, S. Dapprich, J. A. Montgomery Jr., K. Morokuma, and M. J. Frisch, *J. Chem. Theory Comput.*, 2006, **2**, 815–826.

192 L. Zanetti-Polzi, , S. Del Galdo, I. Daidone, M. D'Abramo, V. Barone, M. Aschi and A. Amadei, *Phys. Chem. Chem. Phys.*, 2018, **20**, 24369–24378.

193 G. Longhi, E. Castiglioni, J. Koshoubu, G. Mazzeo and S. Abbate, *Chirality*, 2016, **28**, 696–707.

194 S. Del Galdo, M. Fusé and V. Barone, *Frontiers in Chemistry*, 2020, **8**, 584.