Semiclassical Vibrational Spectroscopy of Biological Molecules using Force Fields

Fabio Gabas,[1] Riccardo Conte,[1] and Michele Ceotto[1, a]

*Dipartimento di Chimica, Università degli Studi di Milano, via Golgi 19,
20133 Milano, Italy.*

Semiclassical spectroscopy is a practical way to get an accurately approximate quantum description of spectral features starting from *ab initio* molecular dynamics simulations. The computational bottleneck for the method is represented by the cost of *ab initio* potential, gradient, and Hessian matrix estimates. This drawback is particularly severe for biological systems due to their unique complexity and large dimensionality. The main goal of this manuscript is to demonstrate that quantum dynamics and spectroscopy, at the level of semiclassical approximation, are doable even for sizable biological systems. To this end, we investigate the possibility of performing semiclassical spectroscopy simulations when *ab initio* calculations are replaced by computationally cheaper force field evaluations. Both polarizable (AMOEBABIO18) and non-polarizable (AMBER14SB) force fields are tested. Calculations of some particular vibrational frequencies of four nucleosides, i.e. uridine, thymidine, deoxyguanosine, and adenosine, show that *ab initio* simulations are accurate and widely applicable. Conversely, simulations based on AMBER14SB are limited to harmonic approximations, but those relying on AMOEBABIO18 yield acceptable semiclassical values if the investigated conformation has been included in the force field parametrization. The main conclusion is that AMOEBABIO18 may provide a viable route to assist semiclassical spectroscopy in the study of large biological molecules for which an *ab initio* approach is not computationally affordable.

[a] Electronic mail: michele.ceotto@unimi.it

I. INTRODUCTION

Semiclassical (SC) dynamics has recently demonstrated its important role in the field of theoretical vibrational spectroscopy. Exploiting information coming from classical dynamics, the SC approach provides zero point energies and the frequencies of quantum-mechanical vibrational transitions through a Fourier transform of the quantum wavepacket survival amplitude. Originating as a stationary phase approximation to the Feynman quantum propagator,[1] SC dynamics became popular thanks to the initial value representation (IVR) and the Herman-Kluk formulation of the semiclassical propagator.[2–5] Then, starting from the early 2000s, a sequence of theoretical advances has contributed to enlarge applicability and reliability of the theory. In 2003 Kaledin and Miller proposed a time averaging filtering technique, labeled TA SCIVR, that alleviated the convergence problem of the phase-space integral calculation.[6,7] Afterwards, in 2009, the computational cost of the semiclassical analysis was drastically decreased by the Multiple Coherent formulation (MC SCIVR), which introduced a tailored choice of a single or few classical trajectories to overcome the standard computationally-expensive Monte Carlo sampling.[8,9] Using such developments, the semiclassical method demonstrated the capability to study small and medium sized systems, up to the glycine molecule, efficiently and in full dimensionality.[10,11] Finally, a crucial leap forward has been performed as early as three years ago with the Divide-and-Conquer technique (DC SCIVR).[12,13] It consists of an efficient recipe to partition the system degrees of freedom, ensuring that the survival amplitude calculation leads to valuable information also in case of high dimensional systems. Exploiting these advances, the semiclassical theory has been successfully applied not only to the calculation of power spectra of medium-size isolated molecules, but also to the study of complex systems like water clusters, the Zundel cation, molecules adsorbed on TiO_2 surfaces, and solvation models.[13–20] SC calculations can be also performed to reproduce IR transition intensities,[21,22] while the most recent advances have focused on fundamental physics aspects like zero-point energy leakage and deterministic chaos. Specifically, it has been shown that SC calculations are free of zero-energy leakage at least when a full sampling of the phase space is performed,[23] and that the influence of chaotic classical trajectories can be largely reduced and sometimes completely avoided by adopting a preliminary adiabatic switching procedure to sample initial conditions.[24]

Semiclassical evaluation of the vibrational spectral density, i.e. calculation of SC power spectra, requires a phase space analysis, based on a short trajectory, together with calculation of the Hessian matrix of the potential energy along the dynamics. When a pre-calculated Potential Energy Surface (PES) is not available, the simulations are performed through *Ab Initio* Molecular Dynamics (AIMD), i.e. evaluating the potential energy step by step using an *ab initio* method. Therefore, any semiclassical approach is limited by the computational effort required and mostly due to the evaluation of the Hessian matrix at all trajectory steps. More than one strategy has been proposed to address this issue. Garashchuk and Light elaborated a method that approximates the Hessian calculation by generating classical trajectories with initial conditions close to the main reference trajectory.[25] Ceotto, Hase et al. proposed instead a compact finite difference (CFD) method to approximate the Hessian calculation at a certain time step by using the latest calculated one and extrapolating the new one.[26,27] Recently, we suggested the possibility to create a database of Hessian matrices during the dynamics that can be exploited to avoid the calculation step by step in favor of a re-use of already calculated Hessian matrices for similar geometries.[28] All these suggestions certainly help alleviate the computational overhead, but *ab initio* calculations remain a relevant time consuming factor which becomes less and less manageable as the system dimensionality increases. For this reason *ab initio* SC calculations are basically restricted to the DFT level of theory, which can limit the accuracy of results in certain instances but provides often quite satisfactorily estimates.[29]

Within this context, we wonder if a more efficient method for calculating trajectories and Hessians is viable. For instance, among all the available computational methods, classical molecular dynamics performed through force fields is implemented by means of fast potential energy calls. It is computationally cheap and commonly used to tackle huge biological systems, like solvated protein, nanotubes and DNA fragments.

Since the release of the first versions of the most famous force fields, like AMBERff94, CHARMM22 or OPLS-AA, during the 1980s-1990s, the success of such an approach has been rapid and widespread.[30–33] In the following years, the growth in available computational power, the advent of multicore-CPU, GPU and specialized hardware, and the constant update of the potential energy function of each of these force fields contributed to the improvement of simulation accuracy.[34–41] Together with the advance of these pioneering versions, starting from the 2000s, a new class of force fields has been proposed by the scientific community. In fact, in the aforementioned force fields, the electrostatic term is described through fixed-point charge methods. To overcome this limitation, the new approach includes an additional term that effectively describes charge polarization. The resulting class of force fields is labeled "polarizable". A famous example is AMOEBA, recently versioned for proteins (AMOEBAPRO13) and nucleic acids (AMOEBABIO18).[42–45]

In this work we want to perform quantum dynamics simulations employing the of AMBER and AMOEBA force fields within the semiclassical approach, in comparison with the well established DFT *ab initio* method. To reach such a goal we selected some biological systems, specifically four nucleosides, for which a parametrization is available in both the chosen force fields. Nucleosides are molecules made of a nucleobase condensed with a five-membered furanose ring, i.e. ribose or deoxyribose. The importance of these molecules lies in the fact that even minor modifications in their structure can lead to different conformations, greatly affecting their biological functionality. Additionally, modified nucleosides are of great interest because they are often employed as new pharmaceuticals.[46,47] To have a representative sample of such biomolecules, we chose to study a couple of deoxy-nucleosides and a couple of nucleosides featuring the ribose sugar moiety, namely deoxyguanosine, thymidine, uridine and adenosine. All these systems have been experimentally studied in gas phase in the recent years. Specifically thymidine, uridine and adenosine have been investigated in argon matrices by the Ivanov group, while a comprehensive study of deoxyguanosine isolated and in mono and di-hydrated clusters has been performed by the Saigusa group.[48–51] The presence of such experimental data gives us a precise benchmark for our calculations, since an exact quantum theoretical estimation is out of reach for molecular systems of this size.

The paper is structured as follows: Section II describes the theoretical and computational details for both the *ab initio* and force field approaches here employed; section III presents all the vibrational frequencies obtained and a discussion of the results, while in section IV the conclusions are listed together with possible future developments.

II. THEORETICAL AND COMPUTATIONAL DETAILS

All DFT calculations were performed by means of the NWChem 6.6 suite of software.[52] We chose to adopt the B3LYP functional,[53] already employed in other semiclassical works focused on biological systems,[11,15,17] and the 6-31G* basis set. Force field calculations were implemented using two different software: Gromacs 5.0.4, in its double precision version, for AMBER simulations, and Tinker 8.6.1 for the AMOEBA counterparts.[54,55] The version of AMBER adopted is ff14SB while for AMOEBA we chose AMOEMABIO18.[39,45] The integration algorithms used in this work are the velocity-Verlet for NWChem simulations, the “md-vv-avek”, which is a more accurate version of velocity-Verlet, for Gromacs, and the Beeman integrator for Tinker. All the NVE trajectories were propagated for a total of 0.6 ps. Specifically, we ran 2500 steps of 10 a.u. (about 0.24fs) each for the DFT dynamics, and 3000 steps of 0.20 fs each for the force field ones. Such a short total propagation time is typical of an SC simulation and it is necessary for capturing all quantum-mechanical information within the survival amplitude calculation before the accuracy of the SC propagator starts to deteriorate.

As for the calculation of the Hessian matrices along the trajectory, we computed them step by step in the case of force field simulations while we adopted the already mentioned Hessian database strategy for DFT studies.[28] This approach allowed us to save about one order of magnitude in computational time. Hessians were analytically computed for calculations employing DFT or AMOEBA, while they were numerically estimated by means of a finite difference approach in the case of simulations based on the AMBER force field. The stability criterion for the monodromy matrix, required by the semiclassical method, has been enforced by means of the well-established regularization strategy.[56] This technique has been always applied choosing the threshold parameter in a way that the regularization is performed for a minimal number of times, in order to minimize the loss of accuracy.

More information regarding the semiclassical formulation and the force field energy functions is briefly reported hereafter.

Semiclassical DC-SCIVR Method

To clearly understand the working equation of the DC SCIVR approach here employed we start describing shortly the earlier MC SCIVR formulation, according to which the vibrational power spectrum for an N-dimensional system has the following formula

$$I\left(E\right) = \left(\frac{1}{2\pi\hbar}\right)^{N} \frac{1}{2\pi\hbar T}\frac{1}{N_{traj}} \sum_{j=1}^{N_{traj}} \left| \int_{0}^{T} e^{i[S_t(\mathbf{p}_j(0),\mathbf{q}_j(0))+Et+\phi_t]/\hbar} \left\langle g_t(\mathbf{p}_j\left(0\right), \mathbf{q}_j\left(0\right)) \middle| \Psi \right\rangle dt \right|^2, \tag{1}$$

where $(\mathbf{p}(0), \mathbf{q}(0))$ are the positions and momenta of the system degrees of freedom at the beginning of the trajectory, T is the total simulation time, S_t the instantaneous classical action at time t, E the Fourier-transform energy, ϕ_t the phase of the prefactor whose definition is

$$\phi_t = \text{phase}\left[\sqrt{\left|\frac{1}{2}\left(\frac{\partial \mathbf{q}(t)}{\partial \mathbf{q}(0)} + \Gamma^{-1}\frac{\partial \mathbf{p}(t)}{\partial \mathbf{p}(0)}\Gamma - i\hbar\frac{\partial \mathbf{q}(t)}{\partial \mathbf{p}(0)}\Gamma + \frac{i\Gamma^{-1}}{\hbar}\frac{\partial \mathbf{p}(t)}{\partial \mathbf{q}(0)}\right)\right|}\right], \tag{2}$$

and $\langle g_t(\mathbf{p}_j\left(0\right), \mathbf{q}_j\left(0\right))|\Psi\rangle$ is the quantum overlap between the coherent state $|g_t(\mathbf{p}_j\left(0\right), \mathbf{q}_j\left(0\right))\rangle$ and the reference state $|\Psi\rangle$.

The coherent state with a Gaussian width matrix Γ has the following formulation

$$\langle \mathbf{q}|g_t(\mathbf{p}\left(0\right), \mathbf{q}\left(0\right))\rangle = \left(\frac{\det(\Gamma)}{\pi^N}\right)^{\frac{1}{4}} e^{-(\mathbf{q}-\mathbf{q}(t))^T\Gamma(\mathbf{q}-\mathbf{q}(t))/2 + i\mathbf{p}^T(t)(\mathbf{q}-\mathbf{q}(t))/\hbar}. \tag{3}$$

The summation runs over a handful of trajectories (N_{traj}) selected according to the MC-SCIVR recipe: The initial conditions should be such that the trajectory explores a region of the phase space close in energy to the real quantum-mechanical vibrational levels. This choice has its foundation in a crucial work by De Leon and Heller who demonstrated that even a single trajectory can effectively lead to a correct quantum eigenvalue estimate, if properly chosen.[57] The statement has been confirmed by several semiclassical studies, remarkably also in the case of neutral glycine.[11] In that work the power spectrum was obtained in two ways, either by means of a single trajectory or using one trajectory per signal. In the case of a single trajectory calculation, the initial conditions were equilibrium positions and velocities derived from the harmonic zero-point vibrational energy estimate of each normal mode. In the case of the multi-trajectory calculations, instead, an additional quantum of excitation was given to the normal mode under consideration. This last strategy is called a “refined” analysis, while the study made in single trajectory is labeled “ZPE” that stands for “Zero Point Energy”. Both these approaches were employed in the present work too.

The DC-SCIVR formula is similar to the already presented MC-SCIVR one, with the difference that all involved quantities are projected onto appropriate subspaces:

$$\begin{aligned} \widetilde{I}_M\left(E\right) = \left(\frac{1}{2\pi\hbar}\right)^M \frac{1}{2\pi\hbar T}\frac{1}{N_{traj}} \sum_{j=1}^{N_{traj}} \\ \left|\int_0^T e^{i\left[\tilde{S}_t(\tilde{\mathbf{p}}_j(0),\tilde{\mathbf{q}}_j(0))+Et+\tilde{\phi}_t\right]/\hbar} \left\langle\tilde{g}_t(\tilde{\boldsymbol{p}}_j\left(0\right),\tilde{\boldsymbol{q}}_j\left(0\right))|\Psi\right\rangle dt\right|^2, \end{aligned} \tag{4}$$

where $\sim$ indicates projection onto an M-dimensional subset, with $M < N$.

All terms are trivially separable except for the potential energy, for which an ad hoc expression modeled on the separable case has been proposed:

$$V_S\left(\tilde{\mathbf{q}}\left(t\right)\right) = V\left(\tilde{\mathbf{q}}\left(t\right);\mathbf{q}_{N-M}\left(t\right)\right) - V\left(\mathbf{q}_M^{eq};\mathbf{q}_{N-M}\left(t\right)\right). \tag{5}$$

In few words, out of the full dimensional dynamics only information coming from a subset of degrees of freedom is considered for the semiclassical analysis. In this way the survival probability calculations return clear signals for the spectrum even for systems made of a large number of degrees of freedom. Such an approach works correctly if the subspace is a good approximation of an isolated system. For this reason more than one strategy has been proposed to partition the totality of the normal modes composing the whole system. Among the proposed methods, the least computationally expensive is the one involving the average Hessian matrix. Following this strategy, the grouping of normal modes is done according to the off diagonal elements of a single matrix obtained by averaging all the Hessian matrices computed along the trajectory. Once the threshold value is fixed, all combinations of normal modes that have off diagonal terms bigger than the threshold value are deemed to interact significantly and hence enrolled in the same subspace.[58]

All spectra presented in this work have been calculated by means of Equation (4), and all subspaces have been determined by means of the average Hessian matrix criterion.

Amber and Amoeba Potential Energy Function

The structure of the AMBER14SB potential energy function is simple and it has been kept in later versions almost unchanged with respect to the one published in 2000.[36] It is composed by four pair terms plus a specific component describing the electrostatic contribution, which is based on the calculation of fixed charges obtained with the restrained electrostatic potential (RESP) procedure.[36,59,60] During the development of AMBER, the major changes have involved different re-parameterizations based on more accurate theoretical quantum mechanical calculations or wider and more precise experimental databases. For example, AMBERff14SB, here employed and published in 2015, overcomes some limitations of the former version in the description of the protein backbones through MP2 calculations in vacuum and some empirical corrections based on recent experiments.[39]

The AMOEBA energy function presents five principal terms for short range interactions: bond stretchings, angle bendings, bond-angles cross terms, out-of-plane bendings and torsional rotations, plus three other terms for non-bonded van der Waals and electrostatic contributions.[43] The polarization term is modeled through dipole and quadrupole moments. Furthermore, a damping scheme for local polarization effects accounts for a consistent treatment of intra- and intermolecular polarization. The major difference with AMBER lies in the electrostatic description that in AMOEBA is evaluated by means of dipole and quadrupole moments. This permits a more precise reproduction of the actual electrostatic contribution to the potential as in the case of the directional hydrogen bond interactions.

We remark that both the Gromacs and Tinker software packages give the possibility to change the functional form of some of the terms mentioned above. For example, it is possible to choose the well-known Morse function to model the bond term. Indeed, in this work, we adopt this choice for all our force field calculations, ensuring a more realistic description for such contribution.

III. RESULTS AND DISCUSSION

Differently from the simpler nucleobases, the molecular structure of nucleosides is more complex and flexible. For this reason nucleosides present a great variety of possible conformations. The global minima adopted in this work are the ones described in the papers by Ivanov. This is true for all molecules with the exception of deoxyguanosine, which is reported in one paper by Saigusa.[48–51] In Figure 1 the minimum geometries predicted by DFT B3LYP/6-31G* *ab initio* calculations are reported. For brevity, we report here only the DFT minimum structures while the AMBER and AMOEBA ones can be found in the SI. However, it is important to point out that they are very similar to the DFT ones, with root mean square deviation (RMSD) values significantly below 1 Angstrom, which is commonly considered the upper limit for a good structural resemblance. Internal hydrogen bonds are

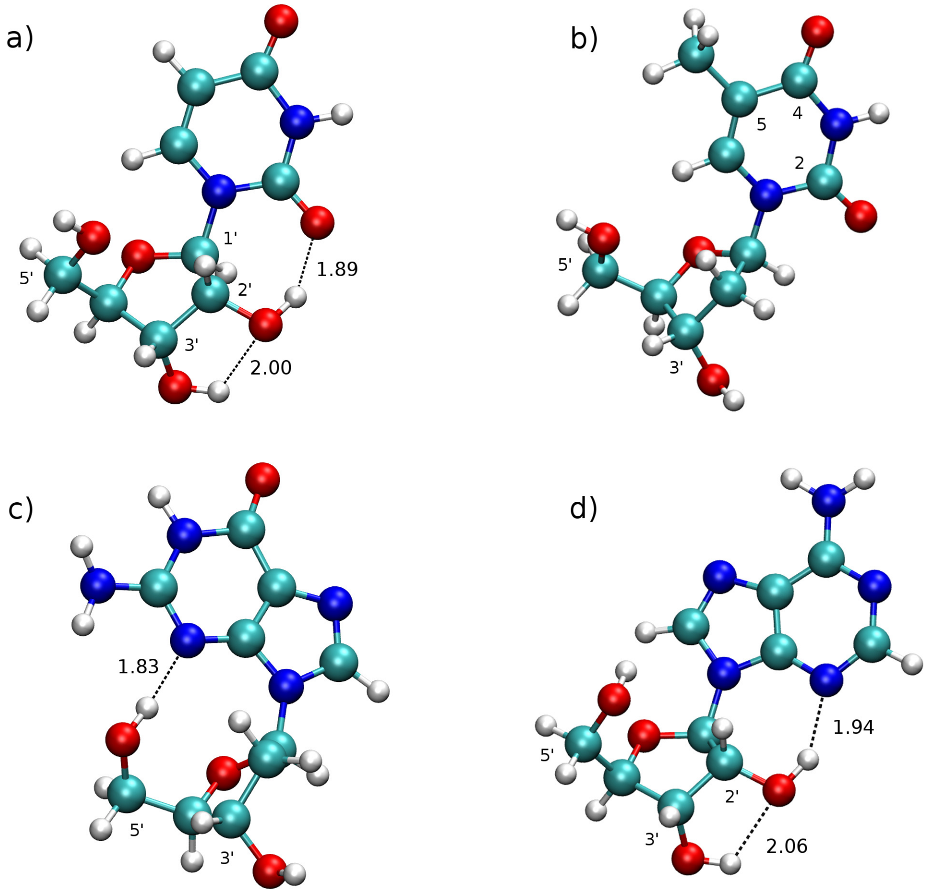

Figure 1. Global minimum structures of uridine (a), thymidine (b), deoxyguanosine, c) adenosine, (d) as predicted by DFT B3LYP *ab initio* calculations. Colors stand for: oxygen(red); hydrogen (grey); carbon (light blue); nitrogen (dark blue). The positions of some relevant carbon atoms are labeled according to the standard numbering and all the internal hydrogen bonds are reported, together with the corresponding distances, in Angstrom.

present in all the nucleosides here studied with the only exception of thymidine. They are displayed in Figure 1 as dashed black lines.

On the global minimum structures, after performing a minimization calculation, the first method we applied to evaluate the vibrational frequencies was the simple harmonic calculation. Results for all the levels of theory employed together with experimental data are reported in Table S1 of the Supplementary Information (SI). The difference between calculated and experimentally measured values is instead pictorially represented in Figure 2. The investigated portion of the vibrational spectra is the characteristic interval in the mid-infrared that spans the interval from 3000 to 4000 cm^{-1}. As already mentioned in the Introduction, the experimental results come from the analysis performed by the group of Ivanov, with the exception of deoxyguanosine. Similarly to its corresponding nucleobase,

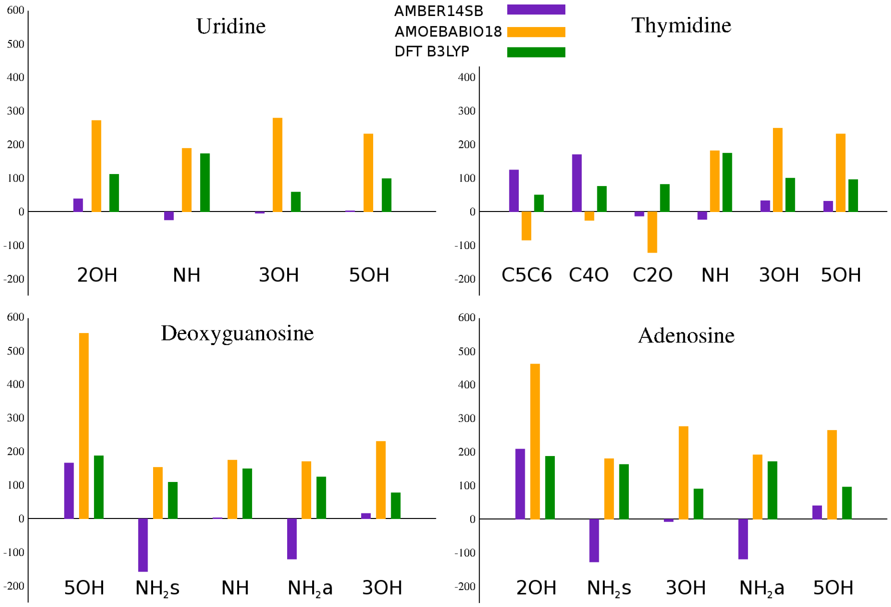

Figure 2. Differences (cm^{-1}) between all calculated harmonic frequencies and experimental values for each nucleoside and theoretical method.

this latter nucleoside exists in both ketonic and enolic conformations due to tautomerism. Unfortunately, in both force fields here employed the parametrized conformation is the ketonic one, while the experimental signals come from the enolic structure, as clearly stated in the Saigusa work.[51] Consequently, we did not have experimental data for the ketonic form but we could still reliably estimate them. In fact, for the OH stretching frequencies we maintained the frequency values coming from the sugar moiety (5'OH and 3'OH), considering negligible for these two OH stretchings the presence of a ketonic group instead of an enolic one in the nucleobase ring. As for the other three frequencies (the NH and the NH_2 symmetric and antisymmetric stretches) we took instead the values reported in the work by Choi and Miller for the ketonic form of the guanine molecules, once again ignoring the interaction effect between the sugar and these three modes in the nucleobase ring moiety.[61] The rationale for these choices comes from a work by Nir et al., in which the similarity between spectra of enolic guanosine and deoxyguanosine is highlighted.[62]

By just looking at the harmonic results, we can already draw some important considerations. As expected, the DFT harmonic estimates are usually higher than the experimental findings. A standard procedure to fix this deviation consists in applying *ad hoc* scaling factors to shift the harmonic predictions near the experimental bands. Conversely, a method like our DC SCIVR can, by construction, account for the actual anharmonicity of the system within a quantum mechanical framework. For this reason, DFT harmonic frequencies higher than the experimental counterpart are suggesting a promising prediction after the inclusion of the anharmonic contributions given by the semiclassical calculation. Such consideration is also valid for the AMOEBABIO18 harmonic results. Even if almost all the values are higher than DFT ones, they are still promising for application of the semiclassical procedure. An exception is the group of three modes of thymidine in the range between 1350 and 1950 cm^{-1}, namely the C5-C6, C4-O, and C2-O stretchings, whose frequencies are lower than the experimental ones. The same reasoning cannot be applied to the AMBER14SB harmonic estimates. In some cases the harmonic frequencies of AMBER14SB are already a good approximation to the real frequencies, while in other instances the estimated frequencies are way too low.

We now apply the DC SCIVR technique employing the two force fields in addition to the B3LYP DFT functional and compare results to the available experimental fundamentals of vibration. As described in section II, the "refined" DC SCIVR analysis requires to run a trajectory per vibrational mode, instead of deriving all spectral signals from a single "ZPE" trajectory. In some cases, in the refined approach, it was also necessary to remove the initial kinetic energy associated with a few modes that might induce internal rotations. This is mandatory to avoid spurious signal splittings in the power spectrum. Owing to the size of the molecules under study, the computational effort required by *ab initio* DFT dynamics and Hessian matrix calculations has limited the possibility to apply the refined procedure to each normal mode for all the molecules. Therefore, we adopted the "ZPE" approach when performing the *ab initio* simulations, limiting the refinement to a single normal mode per molecule, where the ZPE estimate was not satisfactory. On the contrary, force field simulations are extremely cheap, permitting a refined analysis for all the target vibrations of all the nucleosides.

All the semiclassical spectra are displayed in Figures 3, 4, 5, 6, and 7, while the frequency values are listed in the SI. In Figure 7 we report the three peaks belonging to the lower IR frequency region of the thymidine spectrum, which is the only molecule for which we found experimental data also in that spectral region (1350-1950 cm^{-1}).

From these figures we can conclude that we obtained a very good agreement between DFT semiclassical spectra and experimental findings, even if the original harmonic estimates were quite off the mark. The Mean Absolute Error (MAE) calculated for each nucleoside is 40,

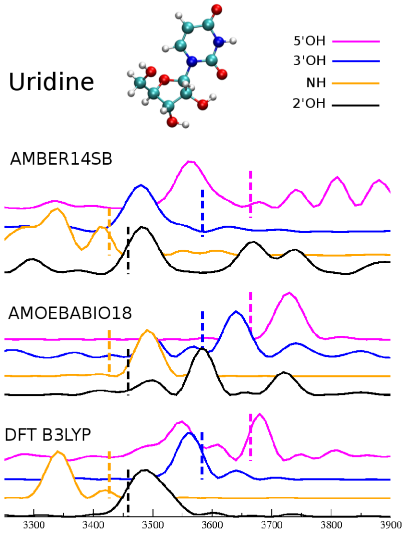

Figure 3. Some DC-SCIVR fundamental frequencies of vibration calculated for the Uridine nucleoside with AMBER14SB, AMOEBABIO18 and *ab initio* DFT B3LYP functional. The experimental values are reported as vertical dashed lines.[48]

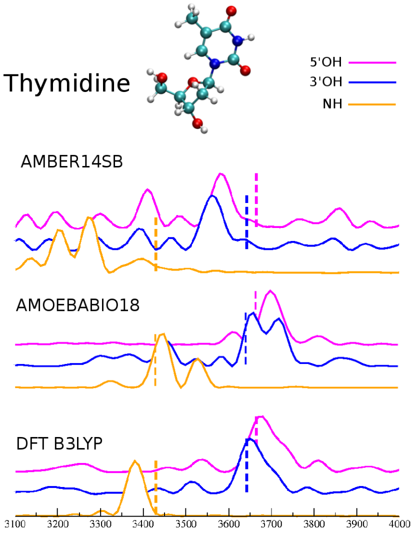

Figure 4. The same of Figure 3 but for the Thymidine molecule. Experimental values are taken from Ref. 49.

33, 25, and 26 cm^{-1} for uridine, thymidine, deoxyguanosine and adenosine respectively. These are reasonable deviations for semiclassical simulations, given that the basis set here employed, 6-31G*, is quite small. For example, Ivanov's work on thymidine presents a VPT2 calculation, performed with the same B3LYP functional but in conjunction with the triple zeta 6-311++G** basis set. Such a theoretical estimate leads to a MAE equal to 7 cm^{-1}. Unfortunately we could not employ that basis set for the semiclassical calculations due to

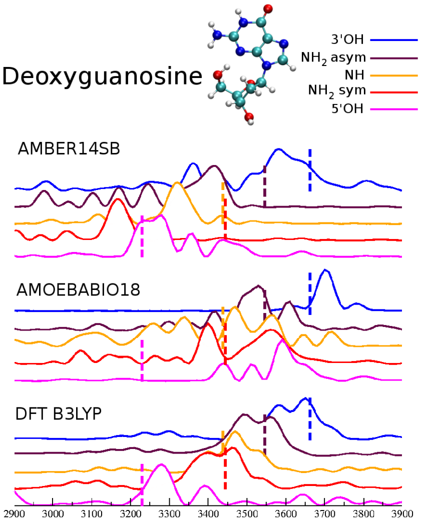

Figure 5. The same of Figure 3 but for the Deoxyguanosine molecule. Experimental values are taken from Refs. 51,61.

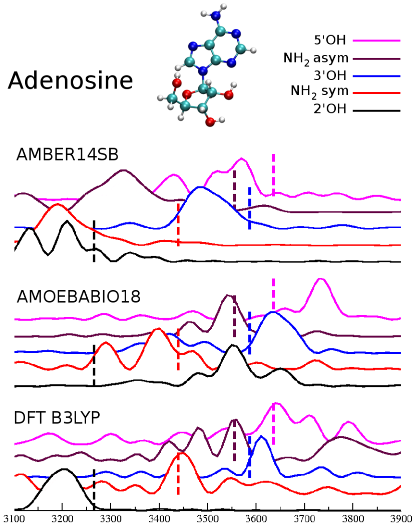

Figure 6. The same of Figure 3 but for the Adenosine molecule. Experimental values are taken from Ref. 50.

its computational overhead, and we had to settle for a faster calculation but slightly lower accuracy.

Moving to AMOEBABIO18 results, we notice that the general agreement with the experiment is quite good for all the investigated frequencies, with the exception of some OH stretching modes. In particular, we obtained very large deviations from the dashed vertical experimental sticks for the 2'OH stretching in uridine and in adenosine and for the 5'OH stretching in deoxyguanosine. The discrepancies are equal to 122, 285 and 360 cm^{-1} respectively. These modes are all involved in internal hydrogen bond interactions. Our explanation

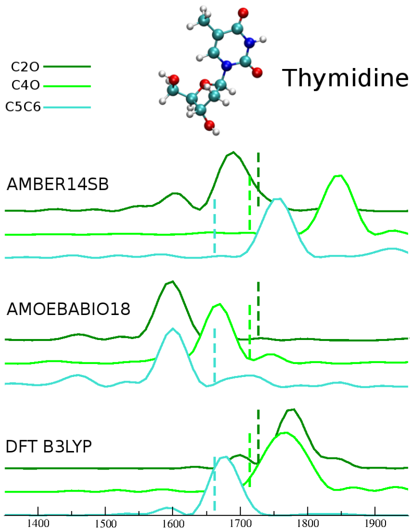

Figure 7. The same as Figure 3, for the 1350-1950 cm^{-1} frequency region of thymidine. The experimental results come from Ref. 49.

for these significant deviations is that AMOEBABIO18 has been probably parametrized on different nucleoside conformations, resulting in a set of atom types that could not predict these internal hydrogen bond interactions. This consideration is reasonable if we think that nucleosides are involved in the double helix formation, where the sugar moiety is perpendicular to the nucleobase and interactions between these two components are minimal. The original parametrization of the force field, hence, refers to this conformation, which is the one biologically active, rather than the one investigated in this paper. Indeed, when the internal hydrogen bond is formed within the sugar moiety, as for example in the 3'OH stretching of uridine and adenosine, the agreement with the experiment turns out to be acceptable (43 and 56 cm^{-1}). A very similar situation has already been detected in AMOEBA by Marx, Head-Gordon, and collaborators. Specifically, in 2017, they published a work of comparison between AIMD and AMOEBA, studying the THz spectra of solvated glycine and valine.[63] They noticed that the zwitterionic form of glycine needed a re-parametrization in order to correctly reproduce the hydrogen bond network and hence the correct signal position and intensity in the THz spectrum. Another example can be found in the SAMPL4 challenge event, which consisted in a blind comparison between various theoretical methods on a set of experimental hydration free energies. In that occasion, a series of papers highlighted that the worst performance of AMOEBA with respect to GAFF (Generalized Amber Force Field) was due to the great sensitivity of AMOEBA to the conformations used for the parametrization.[64–67] Both these examples indeed confirm our AMOEBA semiclassical spectra interpretations. The semiclassical AMOEBABIO18 MAE, calculated considering all the investigated signals, is 77 cm^{-1} for uridine, 51 cm^{-1} for thymidine, 98 cm^{-1} for deoxyguanosine, and 95 cm^{-1} for adenosine. If we remove the three aforementioned erroneous estimates, related to the unparametrized hydrogen bonds, the MAEs become 61, 51, 33 and 48 cm^{-1}, respectively. The deviations from the experiment are higher than those obtained with the DFT simulations, but they are still acceptable, and most importantly the approach is promising for bigger molecular systems, which cannot be treated with *ab initio* DFT.

Conversely AMBER14SB results were, not surprisingly, largely inadequate. As expected, in nearly all the cases in which the harmonic calculation was already a good estimate, the semiclassical analysis deteriorated the accuracy of frequency evaluations. More precisely, almost all harmonic frequencies are closer to the experimental value than the semiclassical ones. This fact suggests us that the AMBER force field parametrization was set to give the best frequency at a harmonic level. For this reason, we do not encourage the reader to employ advanced anharmonic methodologies with the AMBER force field.

To complete the comparison between these three theoretical approaches, we look at the computational effort, expressed in terms of cpu time. It was not surprising to ascertain that the most accurate method required more computational resources than the others. Specifically, DFT B3LYP/6-31G* simulations required about 50 hours on 20 2.4 GHz cpus for the 0.6 ps trajectory. This is an average time for the variously sized nucleosides studied in this work. Additionally, the Hessian matrices took about 30 minutes each to be computed, again on 20 2.4GHz cpus. This time had to be multiplied by the number of Hessian matrices required, which thanks to the adoption of the Hessian database approach was reduced from 2500 to just around 250. A completely different picture was offered by both force fields. The trajectory took a handful of seconds to be evolved , while 3000 Hessian matrices were computed in less than an hour, employing a single cpu. Furthermore, although it is known that AMOEBA is a bit more computationally expensive than AMBER, the difference could not be appreciated at these molecular sizes. The huge advantage of AMOEBA in terms of cpu times over DFT calculations at the cost of a moderate loss in accuracy opens up the route to the semiclassical vibrational study of sizeable biological systems.

IV. SUMMARY AND CONCLUSIONS

In this paper quantum molecular dynamics simulations in semiclassical approximation for AMBER14SB, AMOEBABIO18, and ab initio DFT have been performed for the calculation of the vibrational frequencies of four nucleosides through the DC-SCIVR method. The good agreement with experimental data obtained using DFT demonstrates that the DC-SCIVR method is an adequate approach for medium sized systems and that the B3LYP functional may be appropriate for studying biological systems in spite of the small basis set here employed. AMBER14SB best estimates are harmonic ones, while application of the semiclassical recipe worsens the prediction for almost all the simulated signals. Conversely, we obtained a reasonable set of vibrational frequencies when AMOEBABIO18 was used for semiclassical analysis. In fact, we achieved a comprehensive MAE of about 50 cm^{-1}, with the caveat that frequencies calculated for normal modes involving atoms parametrized for different conformations must be neglected. In our opinion, this aspect represents the real limitation of the AMOEBABIO18 force field: It is necessary to study molecular systems in their biological active conformation, the one for which the force field has been correctly parametrized. In this regard, our semiclassical method can be employed to validate new force fields. Currently, geometrical parameters are the main terms of comparison with experiments for assessing the quality of force fields. Here we propose an additional tool for force field validation, which is based on an anharmonic spectroscopic comparison.

The potential energy surface of the investigated nucleosides is characterized by many low-energy conformers in addition to the global minimum one.[49] We have presented semiclassical simulations based on a short-time dynamics (less than 1 ps long) initiated at the global minimum. Adoption of a much longer dynamics is not a viable route in semiclassical calculations not only because of computational costs, but also because the semiclassical propagator loses rather fast its unitarity and ability to reproduce quantum effects. Similarly to what we pointed out in our past study on glycine,[11] some secondary conformers may be visited during the dynamics in spite of its short duration, owing to the high energy of the trajectories (harmonic zero point energy or higher) compared to the interconversion barriers. On the other side, in such a short time it is not possible to sample the entire phase space including all conformers, and results may overweight the contribution of the global minimum conformer. The necessity to sample a larger portion of the phase space is certainly more compelling when experiments are performed at room temperature. For the investigated nucleosides the benchmark experimental values have been obtained at very low temperature (6-12 K)[48–51] and, even if we cannot rule out that several low-energy conformers might have been populated in the experiment, supported by results we deem that our semiclassical estimates derived from trajectories started at the global minimum provide accurate comparisons to the experiments.

Overall the study opens up the possibility to simulate the quantum dynamics and spectroscopy of very large biomolecules by means of semiclassical techniques assisted by an adequately parametrized force field. For instance, the negligible computational time required for an AMOEBABIO18 DC-SCIVR simulation is promising for future investigations on biological systems like couples of bases, single or double DNA strands, and solvated biomolecules.

ACKNOWLEDGMENTS

Authors acknowledge financial support from the European Research Council [Grant Agreement No. (647107)— SEMICOMPLEX—ERC-2014-CoG] under the European Union's Horizon 2020 research and innovation programme, and from the Italian Ministry of Education, University, and Research (MIUR) (FARE programme Grant No. R16KN7XBRB-project QURE).

SUPPORTING INFORMATION

Supporting material includes minimum structures for the force fields, tables of harmonic and semiclassical frequencies and details regarding the average Hessian criterion and Hessian database strategy.

REFERENCES

[1] R. P. Feynman and A. R. Hibbs, *Quantum mechanics and path integrals* (McGraw-Hill, 1965).

[2] M. F. Herman and E. Kluk, Chem. Phys. **91**, 27 (1984).

[3] W. H. Miller and T. F. George, J. Chem. Phys. **56**, 5637 (1972).

[4]W. H. Miller, J. Chem. Phys. **53**, 3578 (1970).

[5]K. G. Kay, J. Chem. Phys. **100**, 4377 (1994).

[6]A. L. Kaledin and W. H. Miller, J. Chem. Phys. **119**, 3078 (2003).

[7]A. L. Kaledin and W. H. Miller, J. Chem. Phys. **118**, 7174 (2003).

[8]M. Ceotto, G. F. Tantardini, and A. Aspuru-Guzik, J. Chem. Phys. **135**, 214108 (2011).

[9]M. Ceotto, S. Atahan, G. F. Tantardini, and A. Aspuru-Guzik, J. Chem. Phys. **130**, 234113 (2009).

[10]R. Conte, A. Aspuru-Guzik, and M. Ceotto, J. Phys. Chem. Lett. **4**, 3407 (2013).

[11]F. Gabas, R. Conte, and M. Ceotto, J. Chem. Theory Comput. **13**, 2378 (2017).

[12]M. Ceotto, G. Di Liberto, and R. Conte, Phys. Rev. Lett. **119**, 010401 (2017).

[13]G. Di Liberto, R. Conte, and M. Ceotto, J. Chem. Phys. **148**, 104302 (2018).

[14]G. Bertaina, G. Di Liberto, and M. Ceotto, J. Chem. Phys. **151**, 114307 (2019).

[15]F. Gabas, G. Di Liberto, R. Conte, and M. Ceotto, Chem. Sci. **9**, 7894 (2018).

[16]M. Buchholz, F. Grossmann, and M. Ceotto, J. Chem. Phys. **148**, 114107 (2018).

[17]F. Gabas, G. Di Liberto, and M. Ceotto, J. Chem. Phys. **150**, 224107 (2019).

[18]M. Buchholz, F. Grossmann, and M. Ceotto, J. Chem. Phys. **144**, 094102 (2016).

[19]M. Buchholz, F. Grossmann, and M. Ceotto, J. Chem. Phys. **147**, 164110 (2017).

[20]M. Cazzaniga, M. Micciarelli, F. Moriggi, A. Mahmoud, F. Gabas, and M. Ceotto, J. Chem. Phys. **152**, 104104 (2020).

[21]M. Micciarelli, R. Conte, J. Suarez, and M. Ceotto, J. Chem. Phys. **149**, 064115 (2018).

[22]M. Micciarelli, F. Gabas, R. Conte, and M. Ceotto, J. Chem. Phys. **150**, 184113 (2019).

[23]M. Buchholz, E. Fallacara, F. Gottwald, M. Ceotto, F. Grossmann, and S. D. Ivanov, Chem. Phys. **515**, 231 (2018).

[24]R. Conte, L. Parma, C. Aieta, A. Rognoni, and M. Ceotto, J. Chem. Phys. **151**, 214107 (2019).

[25]S. Garashchuk and J. C. Light, J. Chem. Phys. **113**, 9390 (2000).

[26]M. Ceotto, Y. Zhuang, and W. L. Hase, J. Chem. Phys. **138**, 054116 (2013).

[27]Y. Zhuang, M. R. Siebert, W. L. Hase, K. G. Kay, and M. Ceotto, J. Chem. Theory Comput. **9**, 54 (2012).

[28]R. Conte, F. Gabas, G. Botti, Y. Zhuang, and M. Ceotto, J. Chem. Phys. **150**, 244118 (2019).

[29]R. Conte, G. Botti, and M. Ceotto, Vib. Spectrosc. , 103015 (2019).

[30]W. D. Cornell, P. Cieplak, C. I. Bayly, I. R. Gould, K. M. Merz, D. M. Ferguson, D. C. Spellmeyer, T. Fox, J. W. Caldwell, and P. A. Kollman, J. Am. Chem. Soc. **117**, 5179 (1995).

[31]W. D. Cornell, P. Cieplak, C. I. Bayly, I. R. Gould, K. M. Merz, D. M. Ferguson, D. C. Spellmeyer, T. Fox, J. W. Caldwell, and P. A. Kollman, J. Am. Chem. Soc. **118**, 2309 (1996).

[32]B. R. Brooks, R. E. Bruccoleri, B. D. Olafson, D. J. States, S. a. Swaminathan, and M. Karplus, J. Comp. Chem. **4**, 187 (1983).

[33]W. L. Jorgensen and J. Tirado-Rives, J. Am. Chem. Soc. **110**, 1657 (1988).

[34]S. A. Adcock and J. A. McCammon, Chem. Rev. **106**, 1589 (2006).

[35]P. S. Nerenberg and T. Head-Gordon, Curr. Opin. Struc. Biol. **49**, 129 (2018).

[36]J. Wang, P. Cieplak, and P. A. Kollman, J. Comp. Chem. **21**, 1049 (2000).

[37]V. Hornak, R. Abel, A. Okur, B. Strockbine, A. Roitberg, and C. Simmerling, Proteins **65**, 712 (2006).

[38]K. Lindorff-Larsen, S. Piana, K. Palmo, P. Maragakis, J. L. Klepeis, R. O. Dror, and D. E. Shaw, Proteins **78**, 1950 (2010).

[39]J. A. Maier, C. Martinez, K. Kasavajhala, L. Wickstrom, K. E. Hauser, and C. Simmerling, J. Chem. Theory Comput. **11**, 3696 (2015).

[40]A. D. MacKerell Jr, M. Feig, and C. L. Brooks, J. Am. Chem. Soc. **126**, 698 (2003).

[41]R. B. Best, X. Zhu, J. Shim, P. E. Lopes, J. Mittal, M. Feig, and A. D. MacKerell Jr, J. Chem. Theory Comput. **8**, 3257 (2012).

[42]P. Ren and J. W. Ponder, J. Comput Chem. **23**, 1497 (2002).

[43]J. W. Ponder, C. Wu, P. Ren, V. S. Pande, J. D. Chodera, M. J. Schnieders, I. Haque, D. L. Mobley, D. S. Lambrecht, R. A. DiStasio Jr, *et al.*, J. Phys. Chem. B **114**, 2549 (2010).

[44]Y. Shi, Z. Xia, J. Zhang, R. Best, C. Wu, J. W. Ponder, and P. Ren, J. Chem. Theory Comput. **9**, 4046 (2013).

[45]C. Zhang, C. Lu, Z. Jing, C. Wu, J.-P. Piquemal, J. W. Ponder, and P. Ren, J. Chem. Theory Comput. **14**, 2084 (2018).

[46]W. Saenger, *Principles of nucleic acid structure* (Springer Science & Business Media, 2013).

[47]M. S. de Vries and P. Hobza, Annu. Rev. Phys. Chem. **58**, 585 (2007).

[48] A. Y. Ivanov, Low Temp. Phys. **36**, 458 (2010).

[49] A. Y. Ivanov, S. Stepanian, V. Karachevtsev, and L. Adamowicz, Low Temp. Phys. **45**, 1008 (2019).

[50] A. Y. Ivanov, Y. V. Rubin, S. Egupov, L. Belous, and V. Karachevtsev, Low Temp. Phys. **41**, 936 (2015).

[51] H. Asami, S.-h. Urashima, and H. Saigusa, Phys. Chem. Chem. Phys. **11**, 10466 (2009).

[52] M. Valiev, E. Bylaska, N. Govind, K. Kowalski, T. Straatsma, H. V. Dam, D. Wang, J. Nieplocha, E. Apra, T. Windus, and W. de Jong, Comput. Phys. Commun. **181**, 1477 (2010).

[53] C. Lee, W. Yang, and R. G. Parr, Phys. Rev. B **37**, 785 (1988).

[54] M. J. Abraham, T. Murtola, R. Schulz, S. Páll, J. C. Smith, B. Hess, and E. Lindahl, SoftwareX **1**, 19 (2015).

[55] J. A. Rackers, Z. Wang, C. Lu, M. L. Laury, L. Lagardere, M. J. Schnieders, J.-P. Piquemal, P. Ren, and J. W. Ponder, J. Chem. Theory Comput. **14**, 5273 (2018).

[56] G. Di Liberto and M. Ceotto, J. Chem. Phys. **145**, 144107 (2016).

[57] N. De Leon and E. J. Heller, J. Chem. Phys. **78**, 4005 (1983).

[58] G. Di Liberto, R. Conte, and M. Ceotto, J. Chem. Phys. **148**, 014307 (2018).

[59] C. I. Bayly, P. Cieplak, W. Cornell, and P. A. Kollman, J. Phys. Chem. **97**, 10269 (1993).

[60] P. Cieplak, W. D. Cornell, C. Bayly, and P. A. Kollman, J. Comput. Chem. **16**, 1357 (1995).

[61] M. Y. Choi and R. E. Miller, J. Am. Chem. Soc. **128**, 7320 (2006).

[62] E. Nir, C. Plützer, K. Kleinermanns, and M. De Vries, Eur. Phys. J. D **20**, 317 (2002).

[63] A. Esser, S. Belsare, D. Marx, and T. Head-Gordon, Phys. Chem. Chem. Phys. **19**, 5579 (2017).

[64] F. Manzoni and P. Söderhjelm, J. Comput. Aid. Mol. Des. **28**, 235 (2014).

[65] D. L. Mobley, K. L. Wymer, N. M. Lim, and J. P. Guthrie, J. Comput. Aid. Mol. Des. **28**, 135 (2014).

[66] R. T. Bradshaw and J. W. Essex, J. Chem. Theory Comput. **12**, 3871 (2016).

[67] N. A. Mohamed, R. T. Bradshaw, and J. W. Essex, J. Comput. Chem. **37**, 2749 (2016).