An Algorithm to Correct for the CASSCF Active Space in Multiscale QM/MM Calculations Based on Geometry Ensembles

G. Cárdenas[†] and Juan J. Nogueira[*,†,‡]

[†]*Department of Chemistry, Universidad Autónoma de Madrid, Calle Francisco Tomás y Valiente, 7, 28049, Madrid, Spain*

[‡]*IADCHEM, Institute for Advanced Research in Chemistry, Universidad Autónoma de Madrid, Calle Francisco Tomás y Valiente, 7, 28049 Madrid, Spain*

E-mail: juan.nogueira@uam.es

Abstract

The convolution of the excitation energies, computed by the complete active space self-consistent field (CASSCF) or other CAS-based methods, of an ensemble of geometries generated by molecular dynamic simulations is a usual recipe to obtain the absorption spectrum or the density of states of a chromophore. This approach requires that all the considered geometries have the same molecular orbitals within the active space. However, the different geometrical features and/or the different influence of the solvent or biological environments along the sample geometries makes the preservation of the active space a challenging task, which is usually ignored. In this work, we present an algorithm to correct for the active space of geometry ensembles in CASSCF calculations. The algorithm is based on the calculation of the molecular orbital overlap matrix between a previously selected reference geometry, with the desired active space, and each of the sampled geometries. Depending on the value of the overlap matrix

elements, the algorithm determines whether one or more pairs of molecular orbitals of the sampled geometry have to be swapped for a subsequent CASSCF calculation. We have applied the developed algorithm to quantum mechanics/molecular mechanics CASSCF/MM and CASPT2/MM calculations for sets of geometries of the five canonical nucleobases in aqueous solution obtained from classical molecular dynamics simulations. The algorithm shows a very good efficacy since it recovered the correct active space for 76% of the geometries which presented undesired molecular orbitals in the active space after the first CASSCF wavefunction optimization. In addition, the importance of having the same orbitals within the active space for all the geometries is discussed based on the computed density of states for the solvated nucleobases.

Introduction

Multiconfigurational self-consistent field (MCSCF)[1] methods represent nowadays a standard *ab initio* tool used by computational chemists to study systems in situations that cannot be properly described by a single Slater determinant (single reference methods), such as bond dissociations, vertical electronic excitations, and photochemical reaction pathways, among others. Among these methods, the complete active space self-consistent field (CASSCF)[2,3] is perhaps the most popular method due to its conceptual simplicity and the development that it has undergone since its introduction. The CASSCF wavefunction is constructed by first subdividing the molecular orbital (MO) space into three subspaces: inactive, active and secondary orbitals. During the optimization of the wavefunction the inactive and secondary orbitals are assumed to be doubly occupied and completely empty, respectively, whereas the active orbitals are allowed to assume all possible occupations, as long as the overall spin and spatial symmetry of the wavefunction is conserved. Thus, the problem of selecting the most suitable configurations for the chemical problem at hand reduces to that of choosing the set of MOs that compose the active space. This is mainly a problem that involves a strong chemical intuition; however, the size of the set of orbitals is also limited by the size of the

system, and in some cases not all suitable orbitals can be incorporated in the active space due to the huge amount of obtainable configurations. Once the CASSCF wavefunction has been obtained, it can be used as the reference wavefunction for methods that recover the so-called electron dynamical correlation due to electrons not present in the active space, such as the complete active space second-order perturbation theory (CASPT2),[4,5] the restricted active space second-order perturbation theory (RASPT2),[6] or the n-electron valence second-order perturbation theory (NEVPT2).[7] In the last decades, these quantum mechanical methods that rely on a CASSCF reference wavefunction have been extensively employed to unravel a large list of photochemical and photophysical processes.[8–13]

When the photoinduced behaviour of a chromophore is theoretically investigated, for example, by computing the potential-energy curves that connect different stationary points[14–16] or by evolving dynamic trajectories,[17–20] it is usually necessary to compare excitation energies or other electronic properties for different geometries of the chromophore. This is also the case when the absorption spectrum is calculated by considering an ensemble of geometries which are generated by, *e.g.*, molecular dynamic (MD) simulations.[21–26] If the electronic-structure calculations are based on the CAS approach, it is advisable to use very similar orbitals from a qualitative perspective – although they present slightly different coefficients – in the active space for all the geometries that are considered. This is especially true when studying non-reactive processes such as nuclear dynamics at the Franck-Condon region, where there is no bond breaking or formation and, thus, new MOs are not formed. Although the vibrational motion of the chromophore will modify the coefficients of the MOs, and the active spaces of different geometries will be quantitatively different, the qualitative nature of the MOs should not be altered since a drastic change of the electronic wavefunction is not expected. If a chemical reaction or a non-adiabatic process occurs, where the electronic wavefunction suffers important alterations, the modification of the active space must be done smoothly along the reaction pathway to avoid discontinuities in the potential-energy surfaces and sudden changes of the electronic properties. However, to preserve the same or-

bitals inside the active space is a very challenging task in many cases, especially when the molecular motion and the solvent or biological environments have to be explicitly considered in the theoretical model. In this situation it might happen that for some of the geometries the optimized active space differs from that of the desirable reference wavefunction, which could be, for example, the active space of the equilibrium geometry in vacuum. In such cases, a second (or even further) CASSCF optimization needs to be performed using the previously optimized wavefunction as the initial guess but swapping the problematic MOs, in the hope that the desired reference active space is obtained along the different optimizations. The usual way to tackle the problem is to visually assess for each of the failed geometries which orbitals are within the active space and which of them need to be swapped by inactive or secondary orbitals.[14,22,27] However, this becomes a formidable task when considering a large amount of sampled geometries which, in many cases, is disregarded.

We propose an approach to compare the active spaces of a set of geometries with the one of a reference geometry and, when required, to swap the problematic MOs and perform further CASSCF calculations to correct for the wavefunctions of the sampled geometries. The method is based on the calculation of the MO overlap matrix between the CASSCF optimized orbitals of the reference geometry and each of the sampled geometries. The criterion for swapping the MOs will, thus, depend on the value of the overlap integrals of the above mentioned matrix. There are several efficient methods for calculating overlap integrals between many-electron wavefunctions in the literature,[28,29] as well as Density Matrix Renormalization Group methods that provide a better scaling than the standard CASSCF implementations[30] and allow for an automatic selection of the active space,[31] and the present work is by no means an alternative to these approaches. However, to our knowledge no such method has been applied to correct for the active space of an ensemble of geometries that, for example, is considered in the computation of absorption spectra. It should be pointed out that it is not expected that two different geometries of the same molecular species possess exactly the same wavefunction – the configuration-interaction (CI) coefficients and the MO

coefficients will inevitably be different. However, it is desirable that they be qualitatively similar in terms of the nature of their active spaces. However, as evidenced in the present work, for a given geometry different minima of the energy (as a function of the CI and MO coefficients) can be obtained for different initial guess wavefunctions. Being the CASSCF method a variational one, it is thus of utmost importance to tackle this potential ambiguity by attempting to attain the minimum-energy wavefunction, which is the most suitable CASSCF wavefunction for a given geometry.

In this work we present the theoretical insights and the implementation of our approach, followed by an application on an ensemble of force field MD sampled geometries for each of the five canonical nucleobases in explicit water. Nucleobases represent a suitable test case set as their photophysical and photochemical properties have been thoroughly studied both in vacuum and with implicit solvation models within several different MCSCF approaches.[32–35] Thus, our goal is not to reproduce accurate vertical excitation energies, but rather to show the effectiveness of our approach in recovering the appropriate CASSCF active space of an ensemble of geometries. Moreover, we seek to stress out the importance of describing in a proper manner the wavefunction for each of these geometries by comparing the uncorrected CASSCF calculations with their corrected counterparts.

Theory and Implementation

In the CASSCF (or in general MCSCF) framework, the wavefunction is represented as a superposition of Slater determinants that span the configuration space that is most adapted for the system under study

$$\Psi_0 = \sum_j C_j \Phi_j \tag{1}$$

where each Slater determinant Φ_j is constructed from a set of N molecular orbitals $\{\psi_{jk_i}\}_{i\in\{1..N\}}$

$$\Phi_j = |\psi_{jk_1}...\psi_{jk_N}\overline{\psi}_{jk_1}...\overline{\psi}_{jk_N}| \tag{2}$$

To construct the Ψ_0 wavefunction, both the expansion coefficients C_j and the MOs $\{\psi_{jk_i}\}_{i\in\{1..N\}}$ are variationally optimized.

In what follows we present a method to recover the active space in an ensemble of sampled geometries of a molecular species, using a specific geometry and the corresponding active space as the reference. For simplicity, we will drop the Slater determinant index j from the MOs (equation 2) as we will refer to the entire orbital space of a geometry, for which one index suffices. We will use the orbital indices p and q instead of k_i, and will represent the MO space of the reference geometry as $\{\psi_p\}$, and that of any sampled geometry for which to recover the active space as $\{\psi'_q\}$; thus, throughout the present work, MOs without a prime ($'$) will represent reference MOs, and those with a prime ($'$) will represent MOs of a sampled geometry. The method we propose consists of the computation of the MO overlap matrix $\langle\psi_p|\psi'_q\rangle$ (for simplicity $\mathbf{S}^{MO}$) between the reference geometry and any of the ensemble geometries. The molecular orbitals are expanded in an atomic orbital (AO) basis

$$\psi_p = \sum_\mu C_{\mu p}\phi_\mu \tag{3}$$

so that the MO overlap matrix is given by

$$\left\langle\psi_p\middle|\psi'_q\right\rangle = \sum_{\mu\nu} C_{\mu p}C'_{\nu q}\left\langle\phi_\mu\middle|\phi'_\nu\right\rangle \tag{4}$$

where $C_{\mu p}$ and $C'_{\nu q}$ are the MO coefficients of the reference and the sampled geometries, respectively, and $\langle\phi_\mu|\phi'_\nu\rangle$ is the corresponding AO overlap matrix $\mathbf{S}^{AO}$.

The $\mathbf{S}^{MO}$ matrix could be easily determined from the expansion coefficients of the reference $\{\psi_p\}$ and the sampled geometry $\{\psi'_p\}$ MO sets[36] ($C_{\mu p}$ and $C'_{\nu q}$ in equation 4, respectively), assuming that the AOs are equal for the different geometries. However, the AO basis

sets used in the present work comprises atom-centered functions, and since two different geometries of the same molecular system are compared, no such assumption can be made. The only assumption made – for the sake of comparing calculations at the same level of theory – is that the contraction schemes of both AO basis sets are the same. Therefore, the $\mathbf{S}^{AO}$ needs to be calculated explicitly.

To this end, we start by considering the basis set of atom-centered functions, which are linear combinations (contractions) of spherical harmonic Gaussian primitive functions of the type

$$\tilde{\phi}(\zeta, l, m, n, \mathbf{r}) = \tilde{N}(n, \zeta) Y_m^l r^n e^{-\zeta(\mathbf{r}-\mathbf{R})^2} \tag{5}$$

where $\tilde{N}$ denotes the normalization constant, ζ is the orbital exponent, n is the principal quantum number, Y_m^l is the spherical harmonic having orbital and magnetic angular momentum numbers l and m, respectively, $\mathbf{r}$ denotes the electron coordinates, and $\mathbf{R}$ denotes the atomic coordinates on which the Gaussian primitive is centered. However, several integral calculation algorithms rely on the usage of Cartesian Gaussian functions instead

$$\phi(\zeta, l_x, l_y, l_z, \mathbf{r}) = N(l_x, l_y, l_z, \zeta, n)(x - R_x)^{l_x}(y - R_y)^{l_y}(z - R_z)^{l_z} e^{-\zeta(\mathbf{r}-\mathbf{R})^2} \tag{6}$$

where N is the normalization constant, l_x, l_y, l_z are three non negative integers such that, for the orbital angular momentum number, $l = l_x + l_y + l_z$, whereas ζ, $\mathbf{R}$, $\mathbf{r}$ and n have the same meaning as before. In the present implementation, Cartesian Gaussian functions are used, so that instead of calculating the overlap matrix $\mathbf{S}^{AO}$ of spherical harmonic Gaussians directly, we compute the Cartesian overlap matrix $\mathbf{S}_{xyz}^{AO}$. In what follows, we will refer to a shell as a set of Cartesian Gaussian functions having the same principal (n) and orbital angular momentum (l) numbers, and the components of a shell will be the set of basis functions belonging to it. Each Cartesian Gaussian function will be represented as a vector $\mathbf{a} = (a_x, a_y, a_z)$ having as components the three numbers l_x, l_y, l_z mentioned above. Having implemented a vector description for the components of a shell, we will represent them using

vector notation. Thus, the single component of an s shell having angular momentum $l = 0$ is given by $\mathbf{s} = (0, 0, 0)$. To represent each of the three components of the p shell ($\mathbf{p}_x$, $\mathbf{p}_y$ and $\mathbf{p}_z$), and each of the six components of a d shell ($\mathbf{d}_{x^2}$, $\mathbf{d}_{xy}$, $\mathbf{d}_{xz}$, $\mathbf{d}_{y^2}$, $\mathbf{d}_{yz}$, $\mathbf{d}_{z^2}$) we will employ the unit vector $\mathbf{1}_i$ ($i = x, y, z$), given by

$$\mathbf{1}_i = (\delta_{ix}, \delta_{iy}, \delta_{iz}) \tag{7}$$

in terms of Kronecker deltas. In that case, the $\mathbf{p}_i$ and the $\mathbf{d}_{ij}$ components of the p and d shells, respectively, will be given by

$$\begin{aligned} \mathbf{p}_i &= \mathbf{1}_i \\ \mathbf{d}_{ij} &= \mathbf{1}_i + \mathbf{1}_j \end{aligned} \tag{8}$$

so for example, $\mathbf{p}_x$ = (1,0,0), $\mathbf{d}_{xy}$ = (1,1,0) and $\mathbf{d}_{z^2}$ = (0,0,2). Given two Cartesian Gaussian functions $\mathbf{a}$ and $\mathbf{b}$ centered at $\mathbf{A}$ and $\mathbf{B}$, respectively, there are several algorithms that allow for the calculation of the overlap integral $\langle \mathbf{a}|\mathbf{b}\rangle$ – that is, an arbitrary element of the $\mathbf{S}^{AO}_{xyz}$ matrix – in a recursive manner,[37–39] starting from integrals of low angular momentum functions to obtain higher angular momentum integrals. In all such cases, the recursion begins with the analytic expression for the overlap integral between two s functions:

$$\langle \mathbf{s}_a|\mathbf{s}_b\rangle = (\frac{\pi}{\zeta})^{\frac{3}{2}} e^{-\xi(\mathbf{A}-\mathbf{B})^2} \tag{9}$$

where

$$\xi = \frac{\zeta_a \zeta_b}{\zeta_a + \zeta_b} \tag{10}$$

In the present implementation, we use the Obara-Saika recursion relation[40] to compute the integral $\langle \mathbf{a} + \mathbf{1}_i|\mathbf{b}\rangle$ from the lower angular momentum integrals $\langle \mathbf{a}|\mathbf{b}\rangle$ and $\langle \mathbf{a} - \mathbf{1}_i|\mathbf{b}\rangle$

$$\langle \mathbf{a} + \mathbf{1}_i | \mathbf{b} \rangle = (P_i - A_i) \langle \mathbf{a} | \mathbf{b} \rangle + \frac{1}{2\zeta} N_i(\mathbf{a}) \langle \mathbf{a} - \mathbf{1}_i | \mathbf{b} \rangle + \frac{1}{2\zeta} N_i(\mathbf{b}) \langle \mathbf{a} | \mathbf{b} - \mathbf{1}_i \rangle \tag{11}$$

in which **a**, **b** and $\mathbf{1}_i$ have the same meaning as before and N_i is the projection operator on the ith component of a vector. **P** and ζ are defined as follows

$$\zeta = \zeta_a + \zeta_b \tag{12}$$

$$\mathbf{P} = \frac{\zeta_a \mathbf{A} + \zeta_b \mathbf{B}}{\zeta_a + \zeta_b} \tag{13}$$

so that P_i and A_i represent the ith components of **P** and **A** in equation 11, respectively. The recursion in equation 11 stops once a component of either **a** - $\mathbf{1}_i$ or **b** - $\mathbf{1}_i$ is negative (specifically -1), and the value of the corresponding integral is set to zero. All the ζ, ξ and **P** values, as well as the $\langle \mathbf{s}_a | \mathbf{s}_b \rangle$ matrix are computed at the beginning of the calculation. Once the $\mathbf{S}^{AO}_{xyz}$ matrix has been determined, it is transformed into the AO matrix $\mathbf{S}^{AO}$ of spherical harmonic Gaussians as follows

$$\mathbf{S}^{AO} = \mathbf{T}^T \mathbf{S}^{AO}_{xyz} \mathbf{T} \tag{14}$$

The transformation matrix **T** is constructed by using the expansion coefficients of spherical harmonic Gaussians in terms of Cartesian Gaussian functions reported elsewhere.[41] Finally the molecular orbital overlap matrix $\mathbf{S}^{MO}$ is computed using equation 4, which in matrix notation reads

$$\mathbf{S}^{MO} = \mathbf{C}^T \mathbf{S}^{AO} \mathbf{C}' \tag{15}$$

where **C** and $\mathbf{C}'$ are the expansion coefficient matrices of the $\{\psi_p\}$ and $\{\psi'_q\}$ spaces, respectively.

It must be pointed out that to have a meaningful comparison between the MO sets of the reference and the sampled geometries, these need to be aligned prior to calculating the $\mathbf{S}^{MO}$ matrix. Therefore, the center of mass of the sampled geometry is first translated so that it

coincides with the center of mass of the reference geometry. Afterwards, the alignment is accomplished by determining the optimal rotation matrix $\boldsymbol{W}$ that minimizes the root mean square deviation between the two structures, using the algorithm developed by Kabsch[42] and implemented in the quaternion algebra reformulation due to Coutsias *et al.*[43] A comprehensive description of the alignment procedure for obtaining the matrix $\mathbf{W}$ is presented elsewhere;[43–45] here we present only the basic insights. Considering two geometries of the same molecular system of N_a atoms, let $\mathbf{X}$ and $\mathbf{Y}$ be the the two N_ax3 matrices containing the coordinates of the two geometries, with $\mathbf{X}$ being the structure to be superimposed (by means of a rotation $\mathbf{W}$) to the structure $\mathbf{Y}$. The rotation matrix $\mathbf{W}$, which provides the best superposition between the structures $\mathbf{X}$ and $\mathbf{Y}$, is the one that minimizes the following residual function:

$$\begin{aligned} E &= \frac{1}{N_a} tr((\mathbf{XW} - \mathbf{Y})^2) \\ &= \frac{1}{N_a}(G_X + G_Y - 2tr(\mathbf{MW})) \end{aligned} \tag{16}$$

where G_X is the inner product of structure $\mathbf{X}$

$$G_X = tr(\mathbf{X}^T\mathbf{X}) \quad = \sum_i^{N_a}(x_{X,i}^2 + y_{X,i}^2 + z_{X,i}^2) \tag{17}$$

with $x_{X,i}$ $y_{X,i}$ and $z_{X,i}$ being the x, y and z coordinates of the ith atom of structure $\mathbf{X}$, and $\mathbf{M}$ is the matrix product

$$\mathbf{M} = \mathbf{X}^T\mathbf{Y} = \begin{bmatrix} S_{xx} & S_{xy} & S_{xz} \\ S_{yx} & S_{yy} & S_{yz} \\ S_{zx} & S_{zy} & S_{zz} \end{bmatrix} \tag{18}$$

with

$$S_{xy} = \sum_i^N x_{Y,i} y_{X,i} \tag{19}$$

Thus, it can be shown[43] that the optimal rotation matrix $\mathbf{W}$, in the quaternion repre-

sentation,

corresponds to the eigenvector having the largest eigenvalue of the following matrix:

$$\begin{bmatrix} S_{xx}+S_{yy}+S_{zz} & S_{yz}-S_{zy} & S_{xz}-S_{zx} & S_{xy}-S_{yx} \\ S_{yz}-S_{zy} & S_{xx}-S_{yy}-S_{zz} & S_{xy}+S_{yx} & S_{xz}+S_{zx} \\ S_{zx}-S_{xz} & S_{xy}+S_{yx} & -S_{xx}+S_{yy}-S_{zz} & S_{yz}+S_{zy} \\ S_{xy}-S_{yx} & S_{xz}+S_{zx} & S_{yz}+S_{zy} & -S_{xx}-S_{yy}+S_{zz} \end{bmatrix} \tag{20}$$

so that the resulting 4-component eigenvector, which represents the above mentioned quaternion, needs to be transformed into the corresponding 3x3 matrix by a suitable transformation for quaternions.[43] In the case of a QM calculation including point charges (or in general an electrostatic embedding QM/MM calculation), the $\boldsymbol{W}$ matrix is determined for the QM part (which coincides in the number of atoms with the reference geometry), and the point charges are rotated using the same $\boldsymbol{W}$ matrix.

Once the $\mathbf{S}^{MO}$ matrix has been calculated, the comparison between the reference orbital space $\{\psi_p\}$ and the orbital space of the sampled geometry $\{\psi'_q\}$ proceeds as follows: for each column of the $\mathbf{S}^{MO}$, the absolute value of the maximum element is fetched and its matrix element indices recorded (the row indices of the $\mathbf{S}^{MO}$ matrix correspond to the MO labels of the reference structure and the column indices represent the MOs of the sampled geometry). This maximum value is chosen since it qualitatively represents which orbital of the reference space is the most similar to the MO of the sampled geometry under consideration. The orbitals to be swapped for the sampled geometry are determined by generating two lists of orbitals, one for orbitals that need to be removed from the active space (**lrem**) and the other one for orbitals to be added to the active space (**ladd**). These lists are generated by analyzing the above mentioned maximum value for each column of $\mathbf{S}^{MO}$ as follows:

(a) If the column index is outside of the range of the reference active space, but the row index is inside the range of the reference active space, the orbital corresponding to that column index is added to the **ladd** list (orange matrix element in Figure 1).

(b) If both the row and the column indices are within the range of the reference active space, the orbital is kept inside the active space (red matrix element in Figure 1).

(c) If the column index is inside the range of the reference active space, but the row index is outside, the orbital corresponding to that column index is added to the **lrem** list (green matrix element in Figure 1).

(d) Finally, if both the row and the column indices are outside the range of the reference active space, then the orbital corresponding to that column index is ignored, *i.e.*, it is kept outside the active space (blue matrix element in Figure 1).

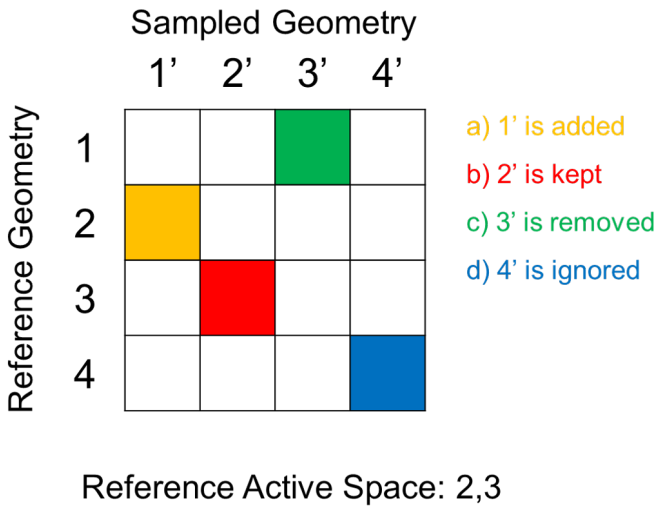

Figure 1: Schematic representation of the $\mathbf{S}^{MO}$ matrix, for the hypothetical situation of an orbital space of 4 orbitals. Here, the reference active space corresponds to orbitals 2 and 3, and the four situations described in the main text (a-d) are depicted by showing the maximum value of each column as a colored square, and what is to be done with the corresponding sampled orbitals, depending on the values of the row and column indices.

It should be emphasized that the lengths of **ladd** and **lrem** must be equal. For example in Figure 1, **ladd** and **lrem** contain the MOs 1' and 2', respectively. This means that the MOs 1' and 2' of the sampled geometry would need to be swapped before performing a second CASSCF calculation.

Computational Details

The force field MD simulations were performed using the AMBER18 package.[46] An octahedral water solvation box of 20 Å from the center of the box to either of the faces was built around each one of the nucleobases using the tleap program of AmberTools19,[46] and the potential of the water molecules was described by the TIP3P[47] solvation model. The intramolecular and van der Waals parameters for all nucleobases were taken from the general amber force field (GAFF) for organic molecules.[48] The geometries for all nucleobases were retrieved from Thiel's benchmark set[35] except for guanine, whose geometry was obtained from the work of Wiebeler *et al.*[34] These geometries were reoptimized at the MP2/6-31G* level of theory using the Gaussian16[49] software, and restrained electrostatic potential (RESP) charges of all the nucleobases were calculated at the B3LYP[50,51]/6-31G* level of theory using the same software. For each solvated nucleobase, the system was at first minimized for 5000 steps using the steepest descent algorithm, after which the conjugate gradient algorithm was used for another 5000 steps. Afterwards, a constant volume (NVT) heating to 300 K was performed for 300 ps using a time step of 2 fs, in which positional restraints were imposed to the geometry of the nucleobase by applying a force constant of 10 kcal/mol. Three consecutive MD simulations of 1 ns each were run in the NPT ensemble to equilibrate the density and to gradually remove the positional restraints previously applied, so that force constants of 10 kcal/mol, 5 kcal/mol and no force constant were applied correspondingly. Finally, a 100 ns production simulation was run in the NPT ensemble using a Langevin thermostat to keep the temperature constant, and the SHAKE[52] algorithm was used to maintain fixed the bond lengths of bonds involving hydrogen atoms as required when using the TIP3P solvation model; a time step of 2 fs was used for all the NPT simulations. For each nucleobase, 100 snapshots were fetched from the last 50 ns of the production run to perform the CASSCF molecular orbital analysis described above.

The reference geometries for the MO overlap analysis correspond to the equilibrium geometries of the nucleobases in vacuum; the corresponding CASSCF MOs were optimized by

means of the state average (SA) CASSCF with the triple-ζ basis set developed by Alrichs,[53] using the OpenMolcas software.[54] For all nucleobases, no symmetry was used in the CASSCF calculations to have a better comparison with the sampled geometries, as it is expected that the C_s symmetry breaks down during the MD sampling. For most of the nucleobases, the same active spaces and number of roots as used in a previous study[34] were considered for the SA-CASSCF calculations. Specifically, for uracil we computed the first 10 roots using a (14,10) active space which included all π electrons plus the four nonbonding electrons of the oxygen atoms; for cytosine we computed the first 8 roots with a (14,10) active space. For adenine 11 roots were computed with a (18,13) active space and for guanine we considered a (20,14) active space and computed the first 9 roots. In the case of thymine, an active space smaller than the one reported before[34,35] was used, that is a (14,10) active space, in which we have excluded the molecular orbital localized on the methyl group (See Supporting Information for more details), and the first 10 roots were computed in the SA-CASSCF calculations. The MOs analyzed with the MO overlap algorithm correspond to the state-averaged natural orbitals, as provided by the OpenMolcas SA-CASSCF calculations. To further investigate the effect of the variation of the CASSCF active space on the ensemble of geometries we performed MS-CASPT2 on top of the SA-CASSCF optimized wavefunctions with the undesired active spaces, as well as on top of the SA-CASSCF wavefuctions with the corrected active spaces, using an imaginary level shift of 0.2 eV[55] and no IPEA correction.[56] All molecular orbital figures were made using the molden software.[57]

Results and Discussion

We start illustrating the presented algorithm in a visual and schematic way for an example geometry for uracil taken from the force field MD simulations. Figure 2a shows the orbitals of the active space –orbitals (23) to (32)– for the reference geometry of uracil, which corresponds to the optimized geometry in vacuum. The orbitals after the first CASSCF calculation for

the geometry labeled as geometry 29, and those corresponding to the same geometry, but after performing the $\mathbf{S}^{MO}$ analysis and the second CASSCF iteration are also shown. As can be seen, the active space resulted from the first CASSCF calculation for the sampled geometry contains two molecular orbitals, namely (24') and (31'), which should not be part of the active space. Consequently, there are two orbitals – (19') and (37') – which erroneously lie outside the active space. The developed algorithm is able to detect these two pairs of orbitals, which need to be swapped in the subsequent CASSCF calculation, by computing the overlap matrix between the MOs of the sampled and reference geometries, whose relevant rows and columns are shown in Figure 2b. The elements of the matrix columns (24') and (31'), which represent the overlap between the orbitals (24') and (31') of the geometry 29 and the orbitals of the active space of the reference geometry, have very small absolute values, indicating that the MOs (24') and (31') have to be removed from the active space. On the contrary, the matrix columns (19') and (37') present large overlap values with the reference MOs (23) and (31) and, therefore, orbitals (19') and (37') have to be included in the active space. For each of the geometries of the ensemble, the algorithm analyzes the MO overlap matrix after the CASSCF computation to determine whether there are orbitals that have to be swapped. If this is the case, a new CASSCF calculation is performed with the modified active space, and this procedure is repeated until the appropriated active space is obtained or until the maximum number of iterations is reached. In the example shown in Figure 2a the MOs inside the active space for the geometry 29 are the same as the MOs of the active space of the reference geometry after the second iteration and, therefore, the CASSCF optimization is converged.

The algorithm is not only useful to perform the swapping of the MOs in an automatic way along all the geometries of the ensemble, but also to identify the orbitals to be swapped in a quantitative and easy way. In some cases, the identification of the undesired and desired orbitals is straightforward and can be done just by visualisation. For example, as seen in Figure 3a, it is relatively easy to directly identify the MO (22') of the geometry 55, which lies

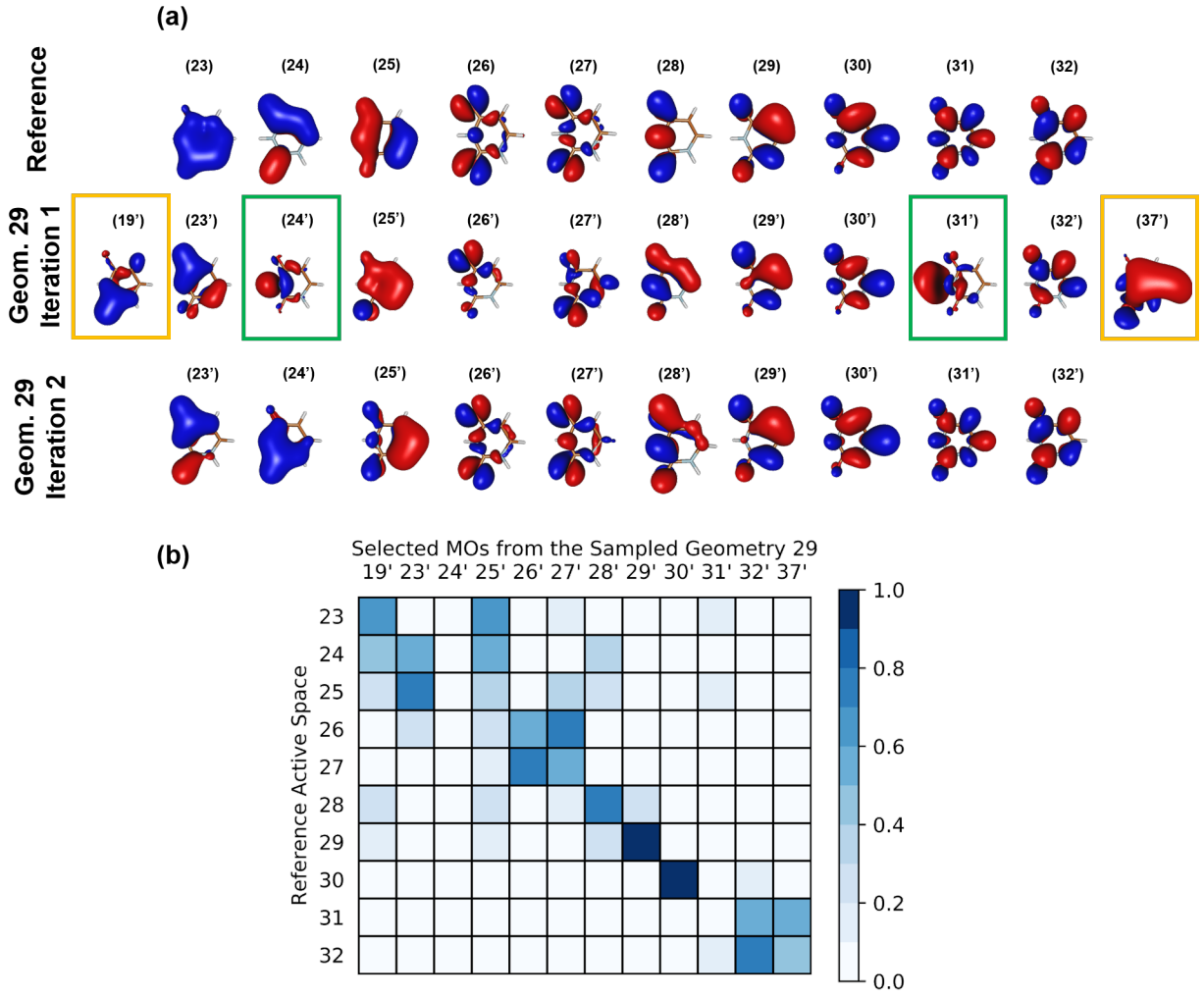

Figure 2: An illustration of the working principle of the $\mathbf{S}^{MO}$ analysis algorithm for sampled geometry 29 of uracil. (a) A comparison among the reference active space, the active space of geometry 29 after the first SA-CASSCF iteration - displaying the MOs that need to be removed (green) and added (orange) to the active space-, and the active space of geometry 29 after the second SA-CASSCF iteration. (b) A portion of the $\mathbf{S}^{MO}$ matrix showing the absolute values of the overlap integrals. It can be clearly evidenced that MOs 24' and 31' display negligible overlap integrals with the MOs of the reference active space, whereas MOs 19' and 37' show significant overlaps with at least one MO in the reference active space.

outside the active space after the first CASSCF calculation, as the equivalent counterpart of the orbital (26) of the reference geometry. Thus, it should be included in the active space. Moreover, the orbital (24'), included in the active space by the first CASSCF wavefunction optimization, is clearly different from all the active-space MOs of the reference geometry of uracil plotted in Figure 2a and, therefore, has to be removed from the active space. However, in many other cases, the visual characterization of the MOs is an impossible task if the optimization process did not reach the global minimum and the wavefunction is not converged with respect to the wavefunction of the reference geometry. For example, our algorithm has identified the orbital (37') of the geometry 27, shown in Figure 3b, as being equivalent to orbital (31) of the reference geometry. Therefore, orbital (37') needs to be

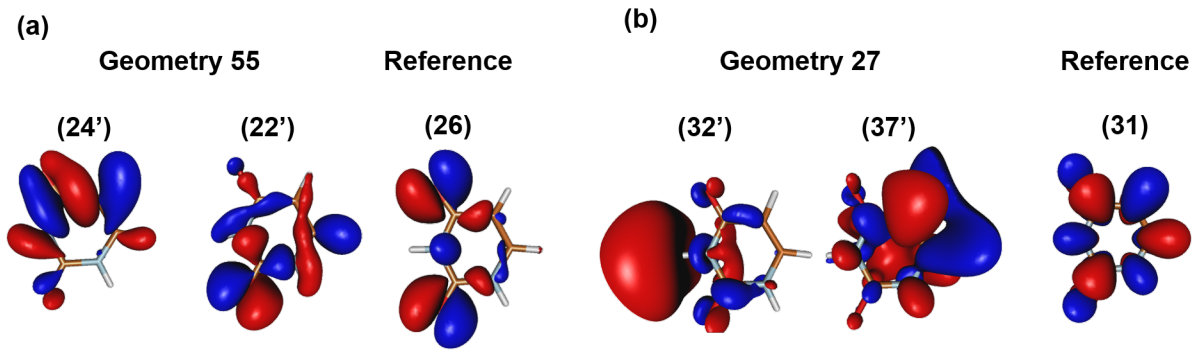

Figure 3: Two examples of geometries of uracil whose active spaces were recovered after a $\mathbf{S}^{MO}$ analysis. For geometry 55 (a), the algorithm correctly identifies MO 22' as being equivalent to the reference MO 26, as can be verified by simple inspection. For geometry 27 (b), the algorithm identifies MO 37' as being equivalent to the reference MO 31, although their similarity is not qualitatively straightforward.

included in the active space in replacement of the orbital (32'). This assignment would have been impossible to do by simple visualization since both orbitals (32') and (37') significantly differ from the reference orbital (31), and it would be hard to decide which one should be included within the active space.

As explained above, we have performed CASSCF/MM calculations for 100 geometries for each of the five nucleobases in aqueous solution. The initial wavefunctions were taken from previous CASSCF calculations for the optimized nucleobases in vacuum. Since large geo-

metrical alterations with respect to the vacuum optimized geometries should not occur along the ground-state classical MD simulations, it could be expected that the wavefunction of the optimized nucleobases in vacuum is a good initial guess for the CASSCF/MM calculations. Therefore, it would be reasonable to assume that the vast majority of CASSCF/MM calculations finish correctly with the right active space. However, as it is shown in Table 1, this is partially true only for adenine, for which 82 out of 100 geometries present the right active space after the CASSCF calculation. However, for the other canonical nucleobases a large amount of geometries have undesirable MOs in the active space. Specifically, 39 geometries for uracil, 37 for cytosine, 34 for thymine and 77 for guanine have a wrong active space after the first CASSCF calculation. Therefore, for all these geometries it was necessary to swap some orbitals based on the analysis of the MOs overlap matrix. Our algorithm shows a very good efficacy since for most of the geometries initially presenting inappropriate MOs within the active spaces, the corresponding active spaces were successfully corrected as can be seen in Table 1. In particular, 92%, 86%, 71%, 67% and 66% of the wrong geometries have been corrected by the algorithm for uracil, cytosine, thymine, adenine, and guanine, respectively. For all the nucleobases but guanine, the right active space for the recovered geometries has been achieved, in average, after carrying out one additional CASSCF calculation, in which two pairs of orbitals have been swapped. The case of guanine is slightly more complicated since it was necessary to swap, in average, almost three pairs of orbitals in the additional CASSCF calculation. Moreover, it was not possible to recover the desired active space for 34 % of the failed geometries.

The preservation of the active space along all the geometries of the ensemble is very important. Different CASSCF – or in general CAS-based calculations – can be combined, for example, to obtain the absorption spectrum or the density of states, only when all the computed geometries have the same active space. Otherwise, different geometries would have electronic properties, such as excitation energies and oscillator strengths, which rely on qualitatively (in some cases drastically) different CASSCF wavefunctions. The convolution of

Table 1: Results of the $\mathbf{S}^{MO}$ analysis on the overall set of sampled geometries for all five canonical nucleobases. The wrong geometries represent those for which the active space differed from the reference active space. The recovered geometries are represented with the percentage over the whole set of geometries presenting the wrong active space after the first SA-CASSCF iteration.

Nucleobase	Correct Geometries[a]	Wrong Geometries[b]	Recovered Geometries[c]	Average Number of Iterations	Average Number of Swaps
Uracil	61	39	36 (92%)	1.06	2.03
Cytosine	63	37	32 (86%)	1.00	1.88
Thymine	66	34	24 (71%)	1.13	1.92
Adenine	82	18	12 (67%)	1.14	2.02
Guanine	23	77	51 (66%)	1.23	2.66

[a] Geometries presenting the desired active space after the first SA-CASSCF iteration
[b] Geometries presenting an undesired active space after the first SA-CASSCF iteration
[c] Geometries for which the desired active space was recovered after performing the $\mathbf{S}^{MO}$ analysis. They are recovered in the sense that they do not need to be discarded for studying the ensemble properties of the system.

these energies based on very different wavefunctions to obtain the spectrum would be equivalent to the convolution of energies computed at different levels of theory. The importance of preserving the active space is exemplified in the following analyses. Figure 4 displays, for each of the recovered geometries for the five nucleobases, the difference in excitation energies for the S_1 state between the wrong and corrected active-space CASSCF calculations, against the variation of the CI coefficient of the most representative determinant in the corrected active-space calculation. Many geometries present large energy deviations – in some cases larger than 0.5 eV – after the MO swapping to attain the desired active space, especially for uracil, cytosine, and guanine.

It is also interesting to see in Figure 4 that there is no correlation between the difference in excitation energy and the difference in the CI coefficients. This means that large excitation energy variations are not always originated by a significant alteration of the electronic wavefunction. In other words, an important error in the electronic properties can be obtained even when the undesired orbitals in the active space are not orbitals involved in the main electronic transitions of the electronic state under study.

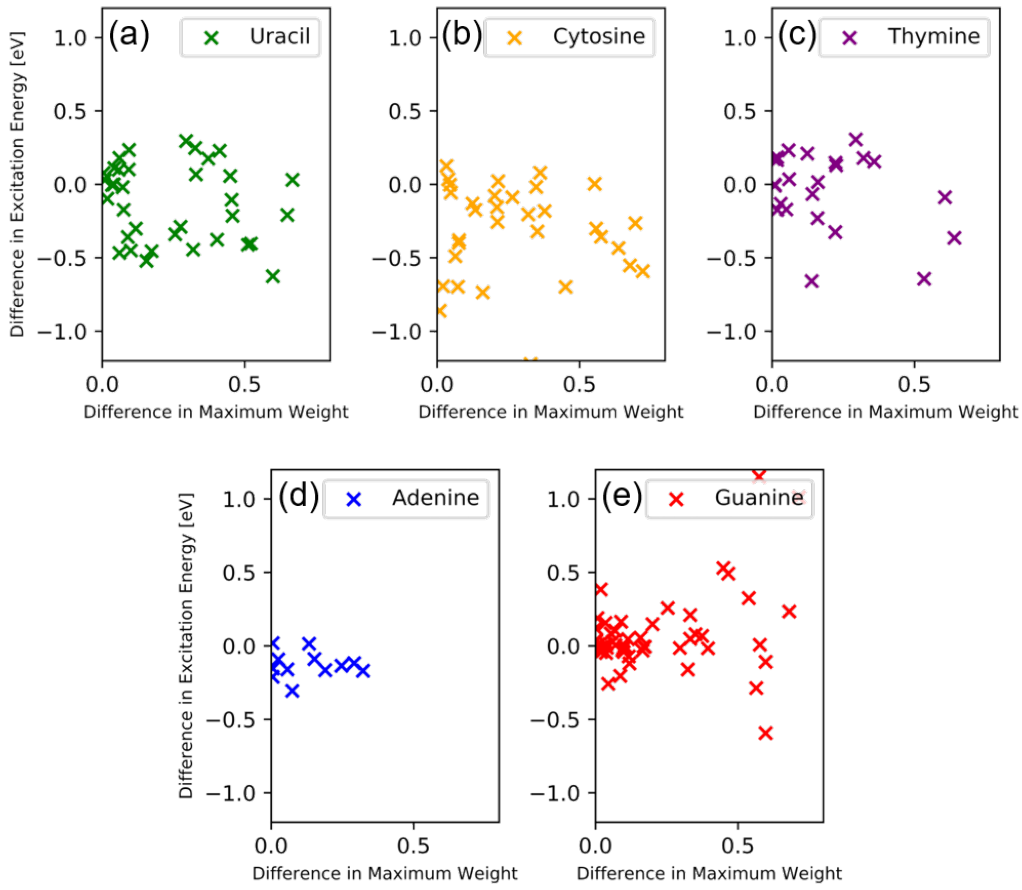

Figure 4: Difference in Excitation Energies towards the S_1 state between SA-CASSCF calculations with the corrected and the uncorrected active spaces for uracil (a), cytosine (b), thymine (c), adenine (d) and guanine (e), vs the difference between the coefficient of the most representative determinant in the corrected active space calculation with its corresponding coefficient in the uncorrected calculation

Although it is evident that swapping the problematic MOs to preserve the active space of the reference geometry provides significantly different results than the first SA-CASSCF wavefunction optimization, this does not directly imply that the result with the swapped orbitals is better. However, this can be easily investigated by comparing the SA-CASSCF energies of the electronic states obtained with the corrected and uncorrected active spaces. The active space that provides the lowest energies, and thus that approaches more to the full-CI result, contains the most suitable MOs. Figure 5a-e shows the energy difference for the six lowest-lying electronic states between the corrected and uncorrected active-space SA-CASSCF calculations for the geometries whose active space were successfully recovered by the algorithm. As can be seen, the energies of the vast majority of electronic states obtained by employing the corrected active space are lower than the energies of the states obtained after the first SA-CASSCF wavefunction optimization. Specifically, the use of the corrected orbitals gives electronic energies that are, in average, 0.56, 0.69, 0.79, 0.51, and 0.61 eV lower than those of the uncorrected calculations for uracil, cytosine, thymine, adenine, and guanine, respectively. Therefore, the wavefunction corrected by the algorithm, which is composed by the same MOs of the active space of the reference geometry, is more accurate than the wavefunction obtained after the first CASSCF calculation. Since the energies between the corrected and uncorrected calculations are clearly different, the density of states presents also important differences as is shown in Figure 5f-j. The band structure of both calculations clearly differs with different number of peaks located at different positions for the five solvated nucleobases.

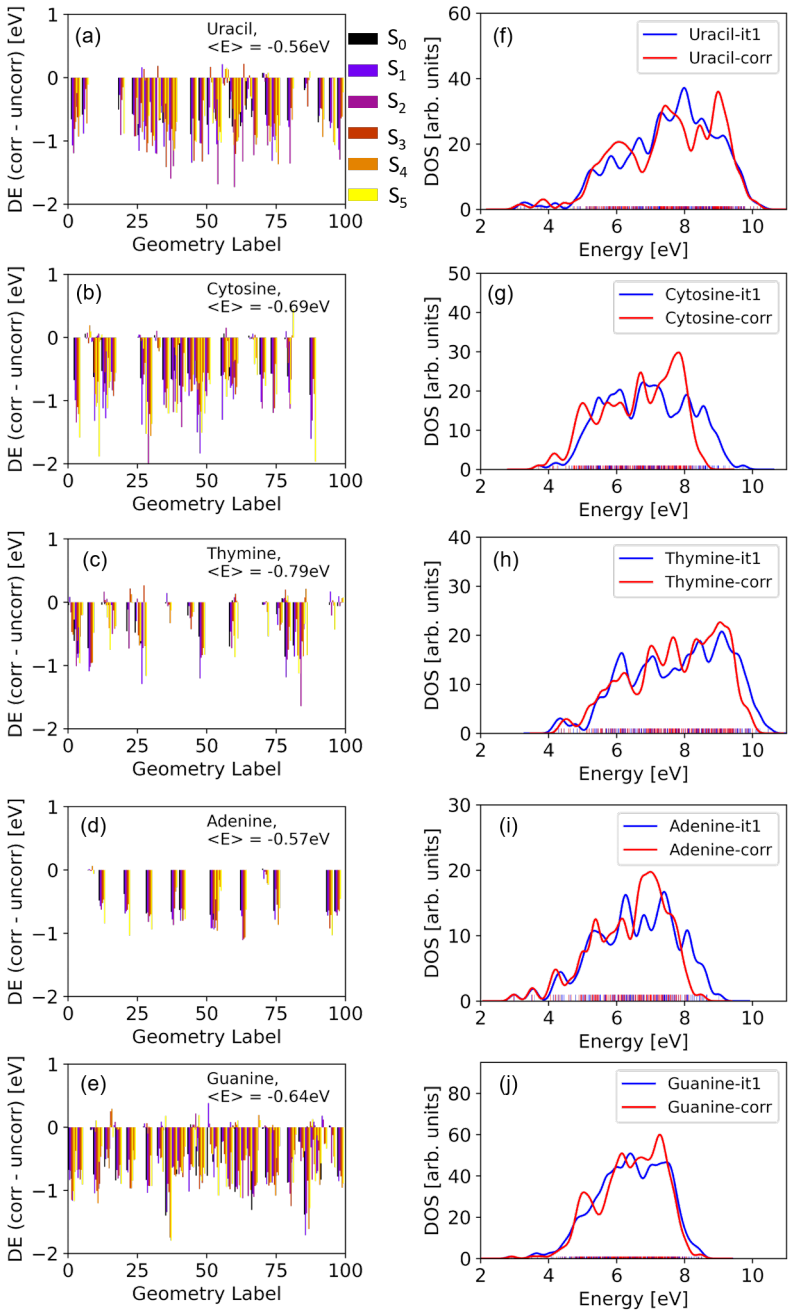

Figure 5: (Left). Absolute energy differences DE for the ground state and the first five singlet excited states between the SA-CASSCF calculations for the uncorrected and the corrected active spaces for uracil (a), cytosine (b), thymine (c), adenine (d) and guanine (e). (Right) Comparison of the DOS (SA-CASSCF) obtained with the active space stemming from the first SA-CASSCF iteration (blue), and the DOS (SA-CASSCF) obtained using the target active space (red) for uracil (f), cytosine (g), thymine (h), adenine (i) and guanine (j). Only those geometries for which the active space was successfully recovered are considered.

It is interesting to see that the large energy differences displayed between the corrected and uncorrected SA-CASSCF calculations partially disappear when MS-CASPT2 calculations are performed on top of the SA-CASSCF wavefunctions. Figure 6a-e shows that the MS-CASPT2 energies based on the corrected SA-CASSCF wavefunction can be lower or higher than the MS-CASPT2 energies based on the uncorrected SA-CASSCF wavefunction depending on the geometry considered. This is an expected result due to the non-variational nature of the MS-CASPT2 approach, which can provide a better result even when a worse reference wavefunction is employed or *vice versa*. In average, the absolute energy difference between the corrected and uncorrected MS-CASPT2 calculations, taking into account only the geometries for which the active space was recovered, is lower than 0.10 eV for the five nucleobases. Despite this fortuitous error cancellation observed when a large number of geometries is considered, the band structure of the density of states suffers important modifications after the swapping of the MOs to keep the same orbitals of CASSCF wavefunction of the reference geometry, as can be seen in Figure 6f-j. In addition, when individual geometries are considered, the error in the CASPT2 excitation energies is larger than 0.50 eV for several geometries. These errors in the electronic properties introduced by the presence of undesired MOs in the active space are not only relevant when investigating the photoabsorption of the chromophores at the Franck-Condon region by means of static calculations. It is evident that these errors can also be present in excited-state dynamics simulations and, thus, artificially alter the conclusions about the photophysics and photochemistry of the chromophores extracted from the dynamics.

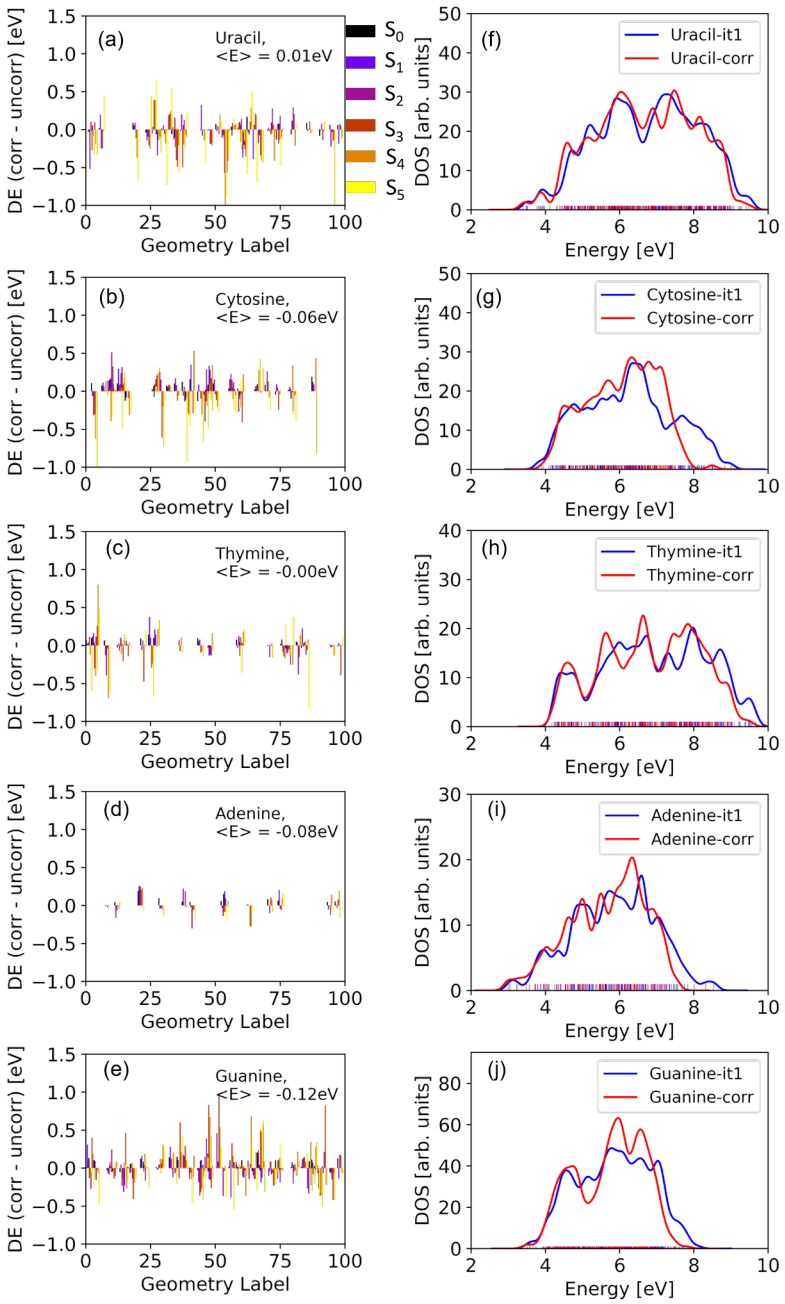

Figure 6: (Left). Absolute energy differences for the ground state and the first five singlet excited states between the MS-CASPT2 calculations for the uncorrected and the corrected active spaces for uracil (a), cytosine (b), thymine (c), adenine (d) and guanine (e). (Right) Comparison of the DOS (MS-CASPT2) obtained with the active space stemming from the first SA-CASSCF iteration (blue), and the DOS (MS-CASPT2) obtained using the target active space (red) for uracil (f), cytosine (g), thymine (h), adenine (i) and guanine (j). Only those geometries for which the active space was successfully recovered are considered.

Conclusions

The analysis and comparison of electronic properties computed by CAS-based approaches for different geometries of the same chromophore requires the use of the same MOs in the active space for all the geometries. This issue is often ignored when, for example, the absorption spectrum or the density of states is computed by selecting an ensemble of geometries which was generated by a sampling technique, such as classical MD. It is generally assumed that the sampled geometries keep the same active space as the reference active space provided as initial guess in the CASSCF calculation. However, we have demonstrated that this is not necessarily true by performing CASSCF/MM and CASPT2/MM computations for the five canonical nucleobases in water. Specifically, 205 out of the 500 geometries (41%) employed in our calculations presented undesired orbitals inside the active space after the CASSCF wavefunction optimization. The correction of the active space by swapping the undesired MOs by the desired ones would be a tremendous task if one aims to perform it manually after visual inspection of the orbitals for the 205 failed geometries.

In this work we have presented an algorithm which automatically assesses whether the sampled geometries present the desired active space and, when this is not the case, identifies the MOs that have to be swapped in subsequent CASSCF wavefunction optimizations. The MOs to be swapped are selected based on the computation of the overlap matrix between the MOs of a reference geometry and the MOs of the sampled geometries. The algorithm was able to correct, in average, the orbitals of 76% of the geometries that presented a undesired active space after the first CASSCF calculation. The efficacy of the algorithm was higher for the pyrimidine (83%) than for the purine (67%) nucleobases, indicating that the preservation of the active space along the ensemble is more difficult to achieve for large chromophores.

The importance of having the correct active space was also demonstrated. Specifically, it has been shown that the SA-CASSCF energy of the electronic states is lower, and therefore more accurate, when the orbitals of the active space of the reference geometry are employed than when undesired orbitals appears in the active space. It has been also discussed that

the excitation energies can suffer artificial and large shifts when undesired orbitals enter the active space during the wavefunction optimization, even when these orbitals are not involved in the excitation that dominates the electronic state. These energy shifts have induced important modifications on the shape of the DOS bands at both SA-CASSCF and MS-CASPT2 levels. Therefore, the amendment of the active space for all the geometries of the ensemble is highly advisable.

The algorithm has been applied to the calculation of the density of states of the five canonical nucleobases in aqueous solution showing a good performance. However, more challenging situations need to be investigated in the future to assess the robustness of the implemented algorithm. For example, chromophores that undergo larger geometrical alterations within the Franck-Condon region, biological environments that strongly interact with the chromophore, or chemical reactions that modify the relevant orbitals that should be included in the active space along the evolution of a reaction coordinate are more difficult scenarios that could be addressed. In addition, the current algorithm will allow the systematic investigation of the factors that can potentially mess up the active space along a set of sampled geometries, such as the size of the chromophore, the nature of its chemical substituents, the presence and description of environments, and the sampling approach, among others. The current implementation can also be adapted to be used in excited-state molecular dynamic codes. If the electronic structure is described by a CAS-based method along the excited-state dynamics, it is important to keep the same active space along the trajectories to avoid discontinuities on the potential-energy surfaces and on the electronic properties of the chromophore. This can be achieved by comparing the orbitals of the active space of the current geometry with those of the previous geometry, and perform orbital swapping based on the analysis of the MO overlap matrix when required.

Acknowledgements

This work is funded by the Comunidad de Madrid through the Attraction of Talent Program 2018 (Grant Ref. 2018-T1/BMD-10261). The Centro de Computación Científica (CCC) of Universidad Autónoma de Madrid is thanked for generous computational resources.

Supporting Information Available

The workflow of the algorithm implementation, the active spaces of the five nucleobases, and the explanation for the choice of the active space of thymine.

References

(1) Shepard, R. *Advances in Chemical Physics*; John Wiley & Sons, Ltd, 2007; pp 63–200.

(2) Roos, B. O. *Advances in Chemical Physics*; John Wiley & Sons, Ltd, 2007; pp 399–445.

(3) Roos, B. O.; Taylor, P. R.; Sigbahn, P. E. A complete active space SCF method (CASSCF) using a density matrix formulated super-CI approach. *Chemical Physics* **1980**, *48*, 157 – 173.

(4) Andersson, K.; Malmqvist, P. A.; Roos, B. O.; Sadlej, A. J.; Wolinski, K. Second-order perturbation theory with a CASSCF reference function. *The Journal of Physical Chemistry* **1990**, *94*, 5483–5488.

(5) Andersson, K.; Malmqvist, P.; Roos, B. O. Second-order perturbation theory with a complete active space self-consistent field reference function. *The Journal of Chemical Physics* **1992**, *96*, 1218–1226.

(6) Malmqvist, P. A.; Pierloot, K.; Shahi, A. R. M.; Cramer, C. J.; Gagliardi, L. The restricted active space followed by second-order perturbation theory method: Theory

and application to the study of CuO2 and Cu2O2 systems. *The Journal of Chemical Physics* **2008**, *128*, 204109.

(7) Guo, Y.; Sivalingam, K.; Valeev, E. F.; Neese, F. SparseMaps—A systematic infrastructure for reduced-scaling electronic structure methods. III. Linear-scaling multireference domain-based pair natural orbital N-electron valence perturbation theory. *The Journal of Chemical Physics* **2016**, *144*, 094111.

(8) Montero, R.; Longarte, A.; Conde, A. P.; Redondo, C.; Castaño, F.; González−Ramírez, I.; Giussani, A.; Serrano-Andrés, L.; Merchán, M. Photophysics of 1-Aminonaphthalene: A Theoretical and Time-Resolved Experimental Study. *The Journal of Physical Chemistry A* **2009**, *113*, 13509–13518.

(9) Josefsson, I.; Kunnus, K.; Schreck, S.; Föhlisch, A.; de Groot, F.; Wernet, P.; Odelius, M. Ab Initio Calculations of X-ray Spectra: Atomic Multiplet and Molecular Orbital Effects in a Multiconfigurational SCF Approach to the L-Edge Spectra of Transition Metal Complexes. *The Journal of Physical Chemistry Letters* **2012**, *3*, 3565–3570.

(10) Bokarev, S. I.; Dantz, M.; Suljoti, E.; Kühn, O.; Aziz, E. F. State-Dependent Electron Delocalization Dynamics at the Solute-Solvent Interface: Soft-X-Ray Absorption Spectroscopy and Ab Initio Calculations. *Phys. Rev. Lett.* **2013**, *111*, 083002.

(11) Schapiro, I.; Sivalingam, K.; Neese, F. Assessment of n-Electron Valence State Perturbation Theory for Vertical Excitation Energies. *Journal of Chemical Theory and Computation* **2013**, *9*, 3567–3580.

(12) Gendron, F.; Fleischauer, V. E.; Duignan, T. J.; Scott, B. L.; Löble, M. W.; Cary, S. K.; Kozimor, S. A.; Bolvin, H.; Neidig, M. L.; Autschbach, J. Magnetic circular dichroism of UCl6^{-} in the ligand-to-metal charge-transfer spectral region. *Phys. Chem. Chem. Phys.* **2017**, *19*, 17300–17313.

(13) Aguilera-Porta, N.; Granucci, G.; Munoz-Muriedas, J.; Corral, I. Unveiling the photophysics of thiourea from CASPT2/CASSCF potential energy surfaces and singlet/triplet excited state molecular dynamics simulations. *Computational and Theoretical Chemistry* **2019**, *1151*, 36 – 42.

(14) Nogueira, J. J.; Oppel, M.; González, L. Enhancing Intersystem Crossing in Phenotiazinium Dyes by Intercalation into DNA. *Angewandte Chemie International Edition* **2015**, *54*, 4375–4378.

(15) Martínez-Fernández, L.; Arslancan, S.; Ivashchenko, D.; Crespo-Hernández, C. E.; Corral, I. Tracking the origin of photostability in purine nucleobases: the photophysics of 2-oxopurine. *Phys. Chem. Chem. Phys.* **2019**, *21*, 13467–13473.

(16) Aleotti, F.; Soprani, L.; Nenov, A.; Berardi, R.; Arcioni, A.; Zannoni, C.; Garavelli, M. Multidimensional Potential Energy Surfaces Resolved at the RASPT2 Level for Accurate Photoinduced Isomerization Dynamics of Azobenzene. *Journal of Chemical Theory and Computation* **2019**, *15*, 6813–6823.

(17) Szymczak, J. J.; Barbatti, M.; Soo Hoo, J. T.; Adkins, J. A.; Windus, T. L.; Nachtigallová, D.; Lischka, H. Photodynamics Simulations of Thymine: Relaxation into the First Excited Singlet State. *The Journal of Physical Chemistry A* **2009**, *113*, 12686–12693.

(18) Barbatti, M.; Aquino, A. J. A.; Szymczak, J. J.; Nachtigallová, D.; Hobza, P.; Lischka, H. Relaxation mechanisms of UV-photoexcited DNA and RNA nucleobases. *Proceedings of the National Academy of Sciences* **2010**, *107*, 21453–21458.

(19) Martínez-Fernández, L.; González-Vázquez, J.; González, L.; Corral, I. Time-resolved insight into the photosensitized generation of singlet oxygen in endoperoxides. *Journal of chemical theory and computation* **2015**, *11*, 406–414.

(20) Mai, S.; Wolf, A.-P.; González, L. Curious Case of 2-Selenouracil: Efficient Population of Triplet States and Yet Photostable. *Journal of Chemical Theory and Computation* **2019**, *15*, 3730–3742.

(21) Rauer, C.; Nogueira, J. J.; Marquetand, P.; González, L. Cyclobutane Thymine Photodimerization Mechanism Revealed by Nonadiabatic Molecular Dynamics. *Journal of the American Chemical Society* **2016**, *138*, 15911–15916.

(22) Nogueira, J. J.; Meixner, M.; Bittermann, M.; González, L. Impact of Lipid Environment on Photodamage Activation of Methylene Blue. *ChemPhotoChem* **2017**, *1*, 178–182.

(23) Nogueira, J. J.; González, L. Computational Photophysics in the Presence of an Environment. *Ann. Rev. Phys. Chem.* **2018**, *69*, 473–497.

(24) Xu, Z.; Matsika, S. Combined Multireference Configuration Interaction/ Molecular Dynamics Approach for Calculating Solvatochromic Shifts: Application to the nO $\rightarrow \pi^*$ Electronic Transition of Formaldehyde. *The Journal of Physical Chemistry A* **2006**, *110*, 12035–12043.

(25) Conti, I.; Nenov, A.; Höfinger, S.; Flavio Altavilla, S.; Rivalta, I.; Dumont, E.; Orlandi, G.; Garavelli, M. Excited state evolution of DNA stacked adenines resolved at the CASPT2//CASSCF/Amber level: from the bright to the excimer state and back. *Phys. Chem. Chem. Phys.* **2015**, *17*, 7291–7302.

(26) Giussani, A.; Conti, I.; Nenov, A.; Garavelli, M. Photoinduced formation mechanism of the thymine–thymine (6–4) adduct in DNA; a QM(CASPT2//CASSCF):MM(AMBER) study. *Faraday Discuss.* **2018**, *207*, 375–387.

(27) Zobel, J. P.; Nogueira, J. J.; González, L. Quenching of Charge Transfer in Nitrobenzene Induced by Vibrational Motion. *The Journal of Physical Chemistry Letters* **2015**, *6*, 3006–3011.

(28) Åke Malmqvist, P.; Roos, B. O. The CASSCF state interaction method. *Chemical Physics Letters* **1989**, *155*, 189 – 194.

(29) Plasser, F.; Ruckenbauer, M.; Mai, S.; Oppel, M.; Marquetand, P.; González, L. Efficient and Flexible Computation of Many-Electron Wave Function Overlaps. *Journal of Chemical Theory and Computation* **2016**, *12*, 1207–1219.

(30) Zgid, D.; Nooijen, M. The density matrix renormalization group self-consistent field method: Orbital optimization with the density matrix renormalization group method in the active space. *The Journal of Chemical Physics* **2008**, *128*, 144116.

(31) Stein, C. J.; Reiher, M. Automated Selection of Active Orbital Spaces. *Journal of Chemical Theory and Computation* **2016**, *12*, 1760–1771, PMID: 26959891.

(32) Lorentzon, J.; Fuelscher, M. P.; Roos, B. O. Theoretical Study of the Electronic Spectra of Uracil and Thymine. *Journal of the American Chemical Society* **1995**, *117*, 9265–9273.

(33) Fülscher, M. P.; Serrano-Andrés, L.; Roos, B. O. A Theoretical Study of the Electronic Spectra of Adenine and Guanine. *Journal of the American Chemical Society* **1997**, *119*, 6168–6176.

(34) Wiebeler, C.; Borin, V.; Sanchez de Araújo, A. V.; Schapiro, I.; Borin, A. C. Excitation Energies of Canonical Nucleobases Computed by Multiconfigurational Perturbation Theories. *Photochemistry and Photobiology* **2017**, *93*, 888–902.

(35) Schreiber, M.; Silva-Junior, M. R.; Sauer, S. P. A.; Thiel, W. Benchmarks for electronically excited states: CASPT2, CC2, CCSD, and CC3. *The Journal of Chemical Physics* **2008**, *128*, 134110.

(36) Shepard, R. *Modern Electronic Structure Theory*; World Scientific Publishing Co Pte Ltd, 1995; pp 345–458.

(37) Boys, S. F.; Egerton, A. C. Electronic wave functions - I. A general method of calculation for the stationary states of any molecular system. *Proceedings of the Royal Society of London. Series A. Mathematical and Physical Sciences* **1950**, *200*, 542–554.

(38) Raffenetti, R. C. General contraction of Gaussian atomic orbitals: Core, valence, polarization, and diffuse basis sets; Molecular integral evaluation. *The Journal of Chemical Physics* **1973**, *58*, 4452–4458.

(39) McMurchie, L. E.; Davidson, E. R. One- and two-electron integrals over cartesian gaussian functions. *Journal of Computational Physics* **1978**, *26*, 218 – 231.

(40) Obara, S.; Saika, A. Efficient recursive computation of molecular integrals over Cartesian Gaussian functions. *The Journal of Chemical Physics* **1986**, *84*, 3963–3974.

(41) Schlegel, H. B.; Frisch, M. J. Transformation between Cartesian and pure spherical harmonic Gaussians. *International Journal of Quantum Chemistry* **1995**, *54*, 83–87.

(42) Kabsch, W. A solution for the best rotation to relate two sets of vectors. *Acta Crystallographica Section A* **1976**, *32*, 922–923.

(43) Coutsias, E. A.; Seok, C.; Dill, K. A. Using quaternions to calculate RMSD. *Journal of Computational Chemistry* **2004**, *25*, 1849–1857.

(44) Theobald, D. L. Rapid calculation of RMSDs using a quaternion-based characteristic polynomial. *Acta Crystallographica Section A* **2005**, *61*, 478–480.

(45) Liu, P.; Agrafiotis, D. K.; Theobald, D. L. Fast determination of the optimal rotational matrix for macromolecular superpositions. *Journal of Computational Chemistry* **2010**, *31*, 1561–1563.

(46) Case, D. et al. AMBER 2018. **2018**,

(47) Jorgensen, W. L.; Chandrasekhar, J.; Madura, J. D.; Impey, R. W.; Klein, M. L. Comparison of simple potential functions for simulating liquid water. *The Journal of Chemical Physics* **1983**, *79*, 926–935.

(48) Wang, J.; Wolf, R. M.; Caldwell, J. W.; Kollman, P. A.; Case, D. A. Development and testing of a general amber force field. *Journal of Computational Chemistry* **2004**, *25*, 1157–1174.

(49) Frisch, M. J. et al. Gaussian˜16 Revision C.01. 2016; Gaussian Inc. Wallingford CT.

(50) Lee, C.; Yang, W.; Parr, R. G. Development of the Colle-Salvetti correlation-energy formula into a functional of the electron density. *Phys. Rev. B* **1988**, *37*, 785–789.

(51) Becke, A. D. Density-functional thermochemistry. III. The role of exact exchange. *The Journal of Chemical Physics* **1993**, *98*, 5648–5652.

(52) Miyamoto, S.; Kollman, P. A. Settle: An analytical version of the SHAKE and RATTLE algorithm for rigid water models. *Journal of Computational Chemistry* **1992**, *13*, 952–962.

(53) Schäfer, A.; Huber, C.; Ahlrichs, R. Fully optimized contracted Gaussian basis sets of triple zeta valence quality for atoms Li to Kr. *The Journal of Chemical Physics* **1994**, *100*, 5829–5835.

(54) Fdez. Galván, I. et al. OpenMolcas: From Source Code to Insight. *Journal of Chemical Theory and Computation* **2019**, *15*, 5925–5964.

(55) Forsberg, N.; Åke Malmqvist, P. Multiconfiguration perturbation theory with imaginary level shift. *Chemical Physics Letters* **1997**, *274*, 196 – 204.

(56) Zobel, J. P.; Nogueira, J. J.; González, L. The IPEA dilemma in CASPT2. *Chem. Sci.* **2017**, *8*, 1482–1499.

(57) Schaftenaar, G.; Noordik, J. H. Molden: a pre- and post-processing program for molecular and electronic structures. *Journal of Computer-Aided Molecular Design* **2000**, *14*, 123–134.