MoBioTools: A Toolkit to Setup QM/MM Calculations

Gustavo Cárdenas,[*,†] Jesús Lucia-Tamudo,[†] Henar Mateo-delaFuente,[†] Vito F. Palmisano,[†] Nuria Anguita-Ortiz,[†] Lorena Ruano,[†] Álvaro Pérez-Barcia,[‡] Sergio Díaz-Tendero,[†,¶,§] Marcos Mandado,[‡] and Juan J. Nogueira[*,†,¶]

†*Department of Chemistry, Universidad Autónoma de Madrid, 28049, Madrid, Spain*

‡*Department of Physical Chemistry, University of Vigo, Lagoas-Marcosende s\n, ES-36310-Vigo, Galicia, Spain*

¶*Institute for Advanced Research in Chemistry (IAdChem), Universidad Autónoma de Madrid, 28049 Madrid, Spain*

§*Condensed Matter Physics Center (IFIMAC), Universidad Autónoma de Madrid, 28049 Madrid, Spain*

E-mail: gustavo.cardenas@uam.es; juan.nogueira@uam.es

Abstract

We present a toolkit that allows for the preparation of QM/MM input files from a conformational ensemble of molecular geometries. The toolkit can be used in command line, so that no programming experience is required, although it presents some features that can also be employed as a python application programming interface. We apply the toolkit in four situations in which different electronic-structure properties of organic molecules in the presence of a solvent or a complex biological environment are computed: the reduction potential of the nucleobases in acetonitrile, an energy decomposition analysis of tyrosine interacting with water, the absorption spectrum of an

azobenzene derivative integrated into a voltage-gated ion channel, and the absorption and emission spectra of the luciferine/luciferase complex. These examples show that the toolkit can be employed in a manifold of situations for both the electronic ground state and electronically excited states. It also allows for the automatic correction of the active space in the case of CASSCF calculations on an ensemble of geometries, as it is shown for the azobenzene derivative photoswitch case.

Introduction

Multiscale hybrid quantum/classical methods have gradually gained popularity since their introduction,[1,2] and are nowadays used routinely for the description of chemical processes that involve complex media.[3–7] These methods consider the partitioning of the system under study into two (or more) fragments: a quantum mechanical (QM) subsystem, which is the region of chemical interest, and a region described classically, which interacts with the QM subsystem. These multiscale approaches have been classified into two different categories in a recent review[8] depending on the description of the classical subsystem. In the first category, the atoms of the classical region are explicitly represented as point-wise particles in a classical potential, for example, a molecular mechanics (MM) analytic potential, that polarizes the quantum mechanical subsystem. The second category uses an implicit continuum description,[9–12] where the classical region is represented as a mean potential that polarizes and is polarized by the quantum mechanical subsystem.

The classification of the paradigms described above can be further branched into different methodologies, depending on the way in which the interaction between the two subsystems is established.[13] For example, within the quantum mechanics/molecular mechanics (QM/MM) framework, one of the most widely used schemes is the electrostatic embedding, in which fixed-point charges of the classical atoms polarize the QM subsystem by entering the one electron part of the Hamiltonian as an electrostatic perturbation. This approach accounts for the polarization of the QM region due to the classical point charges, but does not consider

the polarization of the classical region by the QM one; however, it has been widely employed in a manifold of situations, such as the study of catalytic reactions,[4,14–20] thermodynamic quantities[21–25] and excited state properties[26–37] in biologically and technologically relevant systems. An embedding scheme which accounts for the mutual polarization of the QM and the MM subsystems is the polarizable embedding approach,[7] whereby the classical region is described in terms of fixed point charges plus higher order multipoles. For example, the polarization of the environment can be accounted for by assigning polarizability tensors to the polarizable sites that give rise to an induced charge distribution in the environment. This methodology has been employed in the study of ground state[38,39] and excited state[40] properties of complex biological systems. Its main drawback is the computational cost that arises as self consistency needs to be attained simultaneously for the wavefunction of the QM region and the polarizabilities of the MM region.

An issue that needs to be addressed in the study of complex biological media is that of conformational sampling. Indeed, when the system under study consists of several thousands of atoms, the potential energy landscape often presents several minima that can be populated at the temperature of interest, and conformations corresponding to these minima can contribute to a great extent to the property under study.[6] Thus, this situation is most accurately represented as a statistical distribution arising from the individual conformations, rather than a value due to a specific conformation. In this regard, the application of sampling methods, such as molecular dynamics (MD),[41] Monte Carlo (MC),[42] and Wigner sampling,[43] is crucial to obtain an accurate description of the system. In this contribution, we will mainly focus on sampling by means of MD. However, whenever we refer to a generic ensemble of geometries, and unless otherwise stated, it is implicit that it may be generated by any of the sampling methodologies mentioned above. The potential energy or the potential energy gradient of the system (or part of it) along the sampling process can be computed by a QM method, for example, in *ab initio* MD or QM/MM MD simulations.[44] However, in most cases, using a high level QM potential to perform the conformational sampling is unfeasible.

In these situations, a less sophisticated potential is used to perform the exploration of the potential energy surface, for example, a cheap QM/MM MD approach or MD using analytic force fields.[6] Then, once the sampling is completed, an ensemble of geometries is fetched from the trajectory obtained, and single point QM or QM/MM computations - using a high level QM method of interest - are performed on top of each of these geometries.[28,30] The setup of the hundreds, or even thousands, of single point QM/MM calculations that need to be performed to obtain converged values of the property of interest is an arduous task that has to be automatized.

Nowadays, the computational chemistry community can exploit a manifold of quantum chemistry software that include the electrostatic embedding scheme to perform QM/MM calculations on the one hand, and software aimed at performing sampling at different levels of theory on the other hand, in some cases interfaced with the QM software. There are also packages that act as interfaces between the QM programs and the MD or MC programs, which are either distributed as application programming interfaces (APIs),[45–49] or as executables that require little to no programming experience and solely depend on input files provided by the user. Some examples of the latter case are the Cobramm[50] software, whose main purpose is to interface the Gaussian[51] program and several other QM software with Amber, and the SHARC[52,53] and Newton-X[54] packages which act also as interfaces between QM and MD software but have the main purpose of performing nonadiabatic MD simulations. Each of the paradigms presented above bears several advantages. On the one hand, the usage of APIs in most cases provides a high degree of versatility in the setting up of the QM/MM calculations, but its main drawback is that some degree of programming experience is required. On the other hand, the executables mentioned above are either usable for very specific tasks or require some degree of training before the user gets fully acquainted with them.

Herein, MoBioTools is presented, a simple package that allows for setting up electrostatic embedding QM/MM calculations in an automatic and simple fashion from an ensemble of

geometries. Although it is by no means intended to substitute the interfaces cited above, it aims at presenting an interface that requires only familiarity of the user with a specific quantum mechanical software but, at the same time, attempts to be as general as possible in the input. In this way, specialized tasks, for example, the computation of the interaction energy between two monomers, can be performed with all the QM software MoBioTools is interfaced with, considering as less variation of the input format as possible for the different QM interfaces. Currently the package requires a trajectory and a topology file in an Amber[55] compatible format, and has the possibility to generate QM/MM input files for Gaussian (09 and 16), Orca ($\geq$ 4.0),[56] NWChem[57] and (Open)Molcas.[58,59] However, the interface is under constant development and further compabilities with additional QM software are yet to come. The main feature of the package is the generation of the input files for the QM/MM computations, and not the computations themselves. Therefore, in practice, neither of the QM software are required to execute MoBioTools - the only exception is for (Open)Molcas in the case that the active space of a Complete Active Space Self-Consistent Field (CASSCF) calculation[60,61] needs to be corrected on-the-fly using an algorithm proposed by some of us in the past.[62] In the following, the main structure of MoBioTools is explained and the versatility of the algorithm is shown by applying it to four different situations: the computation of the reduction potential of the nucleobases in acetonitrile, the interaction energy between tyrosine and water and its energy decomposition analysis, the absorption spectrum of an azobenzene derivative integrated into a voltage-gated ion channel, and the emission spectrum of the luciferine/luciferase complex.

Methodology

As stated above, the main purpose of the MoBioTools package is to extract an ensemble of geometries that have been previously sampled by means of MD, MC or any other sampling approach, and automatically generate a QM/MM input file for each of these geometries in

accordance with the QM software and methodology requested by the user. Figure 1 shows a schematic representation of the geometry extraction process and the applicability of the QM/MM single point computations, which are performed on top of the selected snapshots to obtain the desired properties. Some examples of these properties include absorption spectra, interaction energies, free energies, among others (Figure 1b-d). In the case in which the property of interest needs to be computed by means of CASSCF (or a related wavefunction method),[60,61,63] the toolkit performs the automatic correction of the active space for those geometries for which the active space differs from that of a reference structure (Figure 1e).

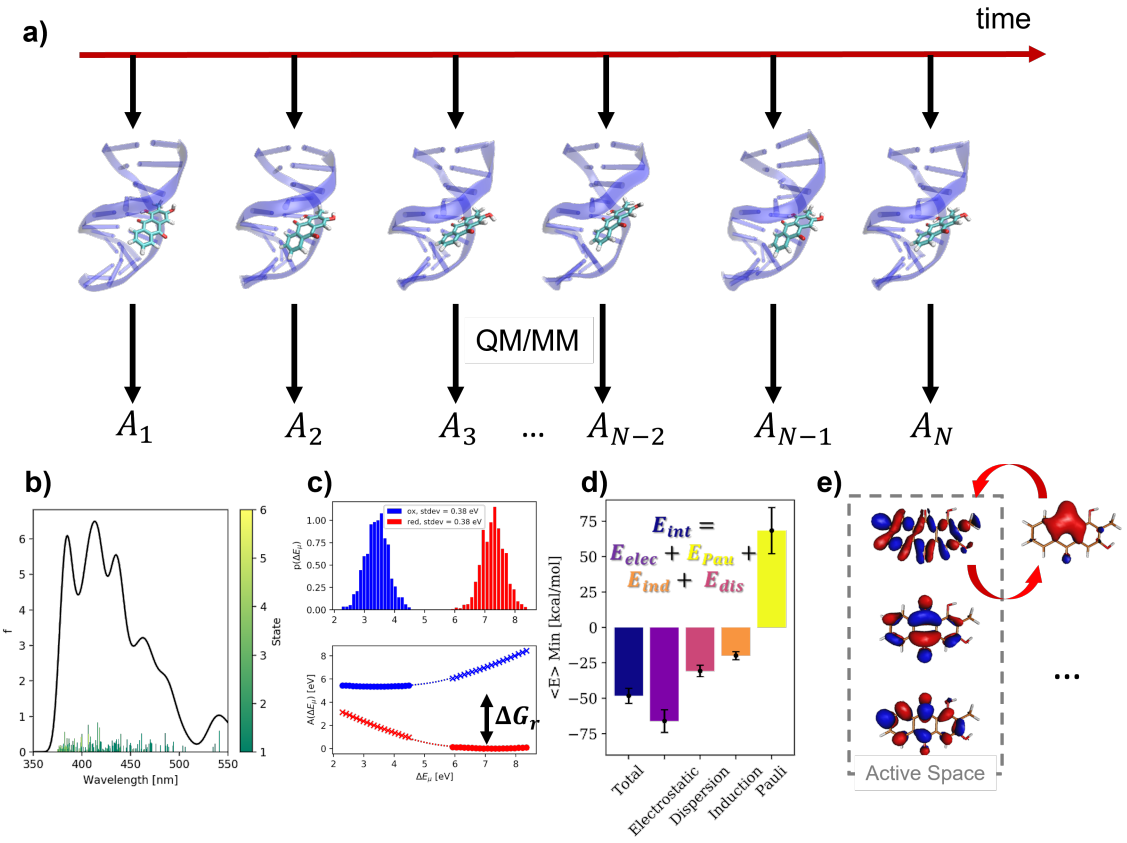


Figure 1: Schematic representation of the working principle of the MoBioTools toolkit and its applicability. a) The extraction of an ensemble of geometries (snapshots) from a trajectory obtained by means of MD, MC or other technique. The QM/MM single point calculations performed using the input files generated provide a distribution of the property of interest A_N, for example, b) the absorption spectrum, c) the reduction free energy, or d) an energy decomposition analysis of the interaction energy between two molecules of a complex. e) Representation of the automatic correction of the CASSCF active space for a specific sampled geometry.

The MoBioTools package is subdivided into two main drivers depending on the task to be performed: a general purpose QM/MM input generator (`main_qminputs.py`), which is executed in command line and requires two input files; and a program to carry out CASSCF computations with an on-the-fly correction of the active space (`pyoverlaps.py`), which works in command line. Although both drivers in principle work similarly, as in both cases a set of QM/MM inputs is generated from an ensemble of geometries, their functionalities and the QM programs they are interfaced with differ considerably, as will be explored in the following sections.

The General Purpose QM/MM Input Generator

The script `main_qminputs.py` is the driver that currently provides an interface to generate a sequence of input files for a set of geometries provided as input in the form of an Amber topology/parameters and an Amber trajectory files. The script works in command line and needs to be provided with two input files: a general input file, which does not depend on the QM software to be used and contains information about the set of geometries to be fetched and settings on how to perform the QM/MM partitioning, and a template file for the QM calculations to be performed. Figure 2 shows a schematic representation of the options featured by each of these files. Both files are subdivided in sections, introduced by an ampersand symbol (&) and enclosed by an &`end` statement.

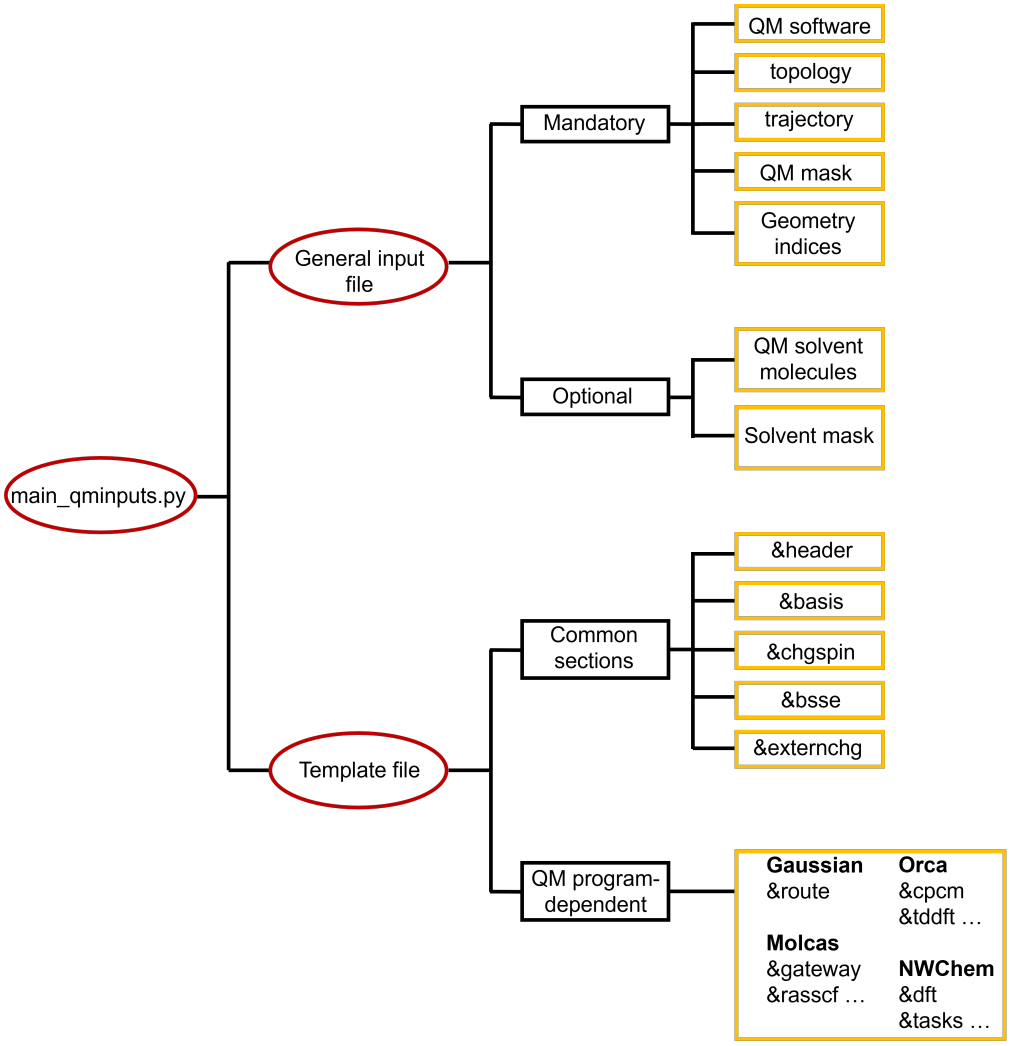


Figure 2: Schematic representation of the contents of the two files provided as input to the main_qminputs.py driver. Red ovals represent files and yellow rectangles represent input sections – introduced by an & symbol – and/or the options these sections enclose. See the text for a detailed description on each of these options.

The general input file (hereby and in listing1 referred to as `main.inp`) consists of a single section (&`main`...&`end`), in which up to seven options can be provided as input by the user on how to treat the ensemble of geometries at hand to generate the input files. In particular, five of these options are mandatory for the software to successfully generate the QM/MM input files: the QM software name (`tpl`), the trajectory (`traj`) and the topology (`top`) file relative to the ensemble of geometries at hand, the definition of the QM region (`qmmask`) and the geometry indices of the geometries for which to generate the input files. The options `solvmask` and `closest` are not mandatory, and refer to the mask of atoms or residues to be treated as solvent molecules, and the number of solvent molecules closest to the QM region, which also need to be treated quantum mechanically, respectively. This is useful whenever a part of the environment that surrounds the quantum mechanical part of the system needs to be included in the QM region, but the indices of the molecules of interest are unknown to the user or they change along the sampled geometries. Here, by mask it is intended the atom or molecule selection, as referred to in the Amber terminology.[55] For the selection mask, the Amber syntax is used so that, for example, `@` refers to atom selection or a colon refers to residue selection. This syntax is used in the general input file as well as in the template file (see listings 1 and 2). The discussion above may seem to imply that the system is treated as a solute-solvent system, where the solute is represented by the `qmmask` entry and the solvent is (if requested) introduced by means of the `solvmask` and the `closest` keywords. It should be emphasized that this is purely a notational convention and it by no means implies that only systems in solution can be studied. This is the notation employed by the cpptraj and the pytraj software,[64] whose machinery is employed by MoBioTools to perform operations on the trajectory objects. However, complex systems containing, for example, proteins and lipid membranes – and not only solvent – can be treated by MoBioTools, as will be evidenced by some of the examples below. If the `solvmask` and the `closest` keywords are not included, a QM/MM calculation (or even a gas phase QM calculation) can be set up for the geometries of interest, where only the `qmmask` entry is treated quantum mechanically.

The template file is also subdivided in sections, some of which are common to all the QM software interfaced with MoBioTools (`header`, `basis`, `chgspin`, `bsse` and `externchg`), whereas some others are exclusive to the QM software under consideration (see Figure 2 for reference). An example of the template for setting up an ensemble of (QM/MM) calculations using the Gaussian software is provided in listing 2. In this example, three common sections are featured: the `head` section, which corresponds to the `Link0` commands in Gaussian (general instructions); the `chgspin` section, which features the charge and the spin multiplicity of the system; and the `externchg` section, which does not take arguments and essentially informs the script that a QM/MM input is requested, so that the point charges that surround the QM region (defined by the `qmmask` keyword in the main input file) also need to be included in the input file. The `route` section is unique to the Gaussian software, and includes the route commands that define the methodology, the basis set and the type of calculation to be performed, among others. The execution of the script `main_qminputs.py` with the files `main.inp` and `template.inp` (listings 1 and 2, respectively) as arguments will generate a gaussian QM/MM input file for the geometry having index 3 in the `GUA_O2.inpcrd` trajectory file.

The example above describes a simple task of extracting a geometry from a coordinate file and generating a QM/MM input file with the desired methodology and settings. Its usefulness stems from the fact that the same operation can be performed for any of the QM software MoBioTools is interfaced with, and from the ease it is scaled to an ensemble of geometries. The keywords `basis` and `bsse` of the template file provide a further degree of flexibility: within the section `basis` the user can employ customized basis sets (for example, different basis sets for different atoms), and the `bsse` keyword provides a way to subdivide the system into two subsystems to request either a fragment calculation or the computation of a property that involves the two subsystems, such as the interaction energy, for which the basis set superposition error (BSSE)[65,66] needs to be corrected. This procedure has been extensively used for generating input files in the past,[67] and in the following section this

approach will be applied to a system of chemical interest.

Listing 1: Example of the general input file to generate a set of input files from an ensemble of geometries. The sections in square brackets refer to optional features. In this example, it is requested to generate an input file for the geometry having index 3 in the GUA_O2.inpcrd coordinate file. The qmmask ":1,2" indicates that molecules 1 and 2 are to be considered as QM molecules.

```
# Name file: main.inp
&main
tpl     = gaussian
traj    = GUA_O2.inpcrd
top     = GUA_O2.prmtop
qmmask = :1,2
geoms   = 3 [start, [stop, step]]
[solvmask = :WAT]
[closest   = 5]
&end
```

Listing 2: Example of a template file for a Gaussian QM/MM calculation. The common sections head, chgspin and externchg, are included, as well as the Gaussian specific route section.

```
# Name file: template.inp
&head
%NprocShared=4
%mem=4GB
&end
```

```
&route
#p M062X/cc-pVTZ Charge
&end

&chgspin
0,1
&end

&externchg
&end
```

The Active Space Automatic Preservation

When CASSCF-based calculations are performed for an ensemble of geometries, it is desirable that the computations of all the geometries include similar molecular orbitals within the active space.[62] If this is the case, the same reference wavefunction is employed along the ensemble of geometries and the different computations can be directly compared or convoluted to obtain, for example, the absorption or emission spectra. The script `pyoverlaps.py` is the driver of the software that allows for the automatic correction of the active space in a CASSCF computation on an ensemble of geometries by comparing, as will be explained below, the orbitals of the active space of the sampled geometries with the orbitals of the active space of a reference geometry. It performs a similar task to the `main_qminputs.py` of the previous section, in that it also generates a set of QM/MM input files from an ensemble of geometries. However, it also presents several differences with respect to the general input generator, perhaps the most important one being the fact that the script `pyoverlaps.py` does not only generate the input files, but it also controls the execution of the QM software, which in this case is the OpenMolcas program. The script works exclusively in command line interface, and requires the presence of four files in the working directory: the trajectory and

the topology/parameters files of the ensemble of geometries under study, a molden file bearing the molecular orbitals used as reference to preserve the active space on the single point CASSCF calculations throughout the selected geometries and a template file for the Molcas inputs. An example of the execution of the script is given by the command `pyoverlaps.py -h`, which prints a list of arguments it can be provided with, the most important of which are shown in listing 3. Apart from the four files that need to be present in the working directory, the user needs to provide the indices of the molecular orbitals (in the reference geometry) that need to be considered as the reference active space (`-ref` option), the mask of the QM atoms (`-qm`), and the index of the geometry for which to perform the CASSCF calculations and the active space correction, if needed. The execution is related to a single geometry, however, it can (and should) be easily executed in parallel.

Listing 3: Mandatory arguments for the execution of the pyoverlaps.py script.

```
-p  TOP               Topology file
-c CRD                Trajectory file
-r REFERENCE          Reference molden file
-tpl  TEMPLATE        Molcas template file
-rng RNG [RNG ...]    Range of MOs in the active space
                      (e.g: 30 43)
-qm QMMASK            QM mask
--ig IGEOM            Specific frame for which to generate
                      an input file. Default = 0
```

The driver `pyoverlaps.py` presents the implementation of a previously proposed algorithm[62] that automatically corrects the active space of a specific geometry. This is achieved by computing and analyzing the overlap matrix (S_{MO}) between the set of CASSCF molecular orbitals (MOs) of the reference geometry and that of the geometry of interest. Briefly, the script performs a CASSCF calculation on the sampled geometry, after which it computes

the overlap matrix S_{MO} and analyzes its column maxima. The column labels of S_{MO} represent the MOs of the sampled geometry, whereas the row labels represent the MOs of the reference geometry. In this framework, it is assumed that there exists a bijection between the set of reference MOs and the set of MOs of the sampled geometry (sampled MOs for simplicity). Thus, the maximum value of each column represents the reference orbital that is the most similar to the MO of the sampled geometry for that column. If there is a sampled MO outside of the active space, whose maximum value in S_{MO} coincides with a MO in the active space of the reference geometry, that MO should be included in the active space of the sampled geometry (see the matrix element $S_{MO}[4][5']$ on Figure 3). If, on the other hand, the maximum value of a sampled MO that belongs to the active space coincides with a reference MO outside the active space, that sampled MO should be removed from the active space. The correction occurs when at least two MOs are each in one of the two situations above, in which case the two MOs are swapped, thus, creating a new guess wavefunction, and a new CASSCF optimization is performed. This procedure is carried out iteratively by the program until the MOs in the active space of the sampled geometry are in a one-to-one correspondence with the MOs in the active space of the reference geometry. Figure 3 shows a schematic representation of the situation in which the positions of two MOs need to be swapped to correct the active space of sampled geometry 3.

Although the execution of `pyoverlaps.py` occurs in the command line, with little to no interaction of the user with the code, the script itself relies on a framework that can also be exploited by the user to some extent as an API as well. The general structure and dependencies of the `pyoverlaps.py` are shown in Figure 4. In synthesis, the driver carries out two main tasks: on the one hand, it is responsible for most of the input and output operations, as well as the operations that regard the reading and handling of the trajectory (and topology files), including the definition of the QM and MM regions. On the other hand, and as explained above, it bears the implementation of the active space correction algorithm.

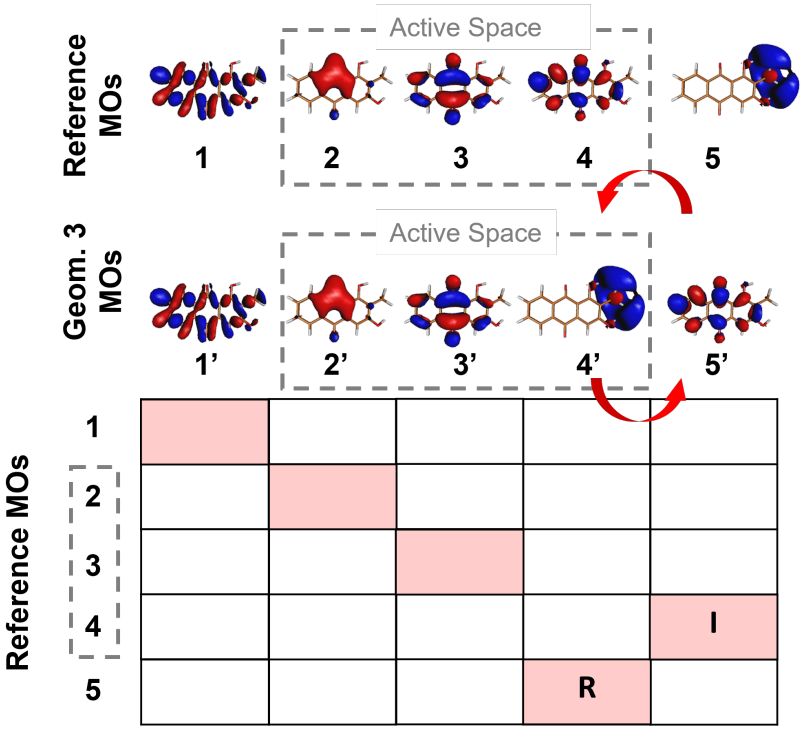


Figure 3: Schematic representation of the working principle of the algorithm to correct the active space on an ensemble of sampled geometries.[62] Top: the set of MOs of the reference geometry and the MOs of a sampled geometry (Geom. 3). Bottom: the overlap matrix S_{MO} between the two MO orbital sets. The column maxima are highlighted in pink. In this example, the maximum of column 4' (that is, MO 4' of geometry 3) coincides with MO 5 from the reference set of MOs, so that it has to be removed from the active space (hence the R label). On the other hand, the maximum of column 5' (MO 5' of geometry 3) coincides with reference MO 4, which is inside the reference active space, and it should be included in the active space (hence the I label). Thus, MOs 4' and 5' need to be swapped to correct the active space of geometry 3.

However, the machinery for reading and operating with the atomic orbital (AO) and MO information of the two geometries, whose orbital spaces are to be compared, is present in the `parse_molden.py` module. Its most important feature is the class `Mol()`, from which an object is created by providing it with a molden file as an argument. During the execution of `pyoverlaps.py`, two `Mol()` objects are created at each iteration: one for the reference geometry and one for the geometry under study, but no trace of these objects remains at the end of the execution. If the user is interested in analyzing the data of each molden file or computing the S_{MO} matrix without executing the `pyoverlaps.py`, the `parse_molden.py` module can be executed interactively on a python shell; this will create the two `Mol()` objects and perform the computation of the atomic orbital overlap matrix S_{AO} and the S_{MO} matrix. Alternatively, the user can import the `Mol()` class and compute these matrices by their own. An important feature is that the `Mol()` class is similar (although less sophisticated) to the `pyscf.gto.M` class from the PySCF[68] software, in particular it bears the attributes `_bas`, `_env` and `_atm` that can be employed to use the machinery of the PySCF to compute molecular integrals for a given molecular species.[68,69]

Figure 4 shows the dependencies of the `parse_molden.py` script/API. These dependencies are modules that provide a lower level framework for the treatment of a `Mol()` object and for the computation of the S_{AO} and S_{MO} matrices. The `Align.py` module aligns the sampled geometry with the reference one prior to the computation of the overlap matrices. The `permutations.py` module allows the reordering of the AO basis functions between the PySCF ordering and that from Molcas (and other QM software such as Gaussian or Orca), and the `ovlp_wrapper.py` module is a wrapper for the C++ implementation of the AO overlap integrals, present in the library `intwrap.so`. These AO integrals have been implemented formally using the Obara-Saika recursion relations[70] in an iterative fashion, a procedure which is routinely performed in the literature.[69,71] The reason for implementing the AO integrals stems from the fact that, to our knowledge, there are no APIs that allow

for the computation of AO overlap integrals between two different atom centered basis sets, a crucial feature to compare the MO sets of two different geometries of the same molecular species, as in principle the atom centers will not be equal for both structures. The current implementation is also 10 times faster than the one presented in a previous work.[62]

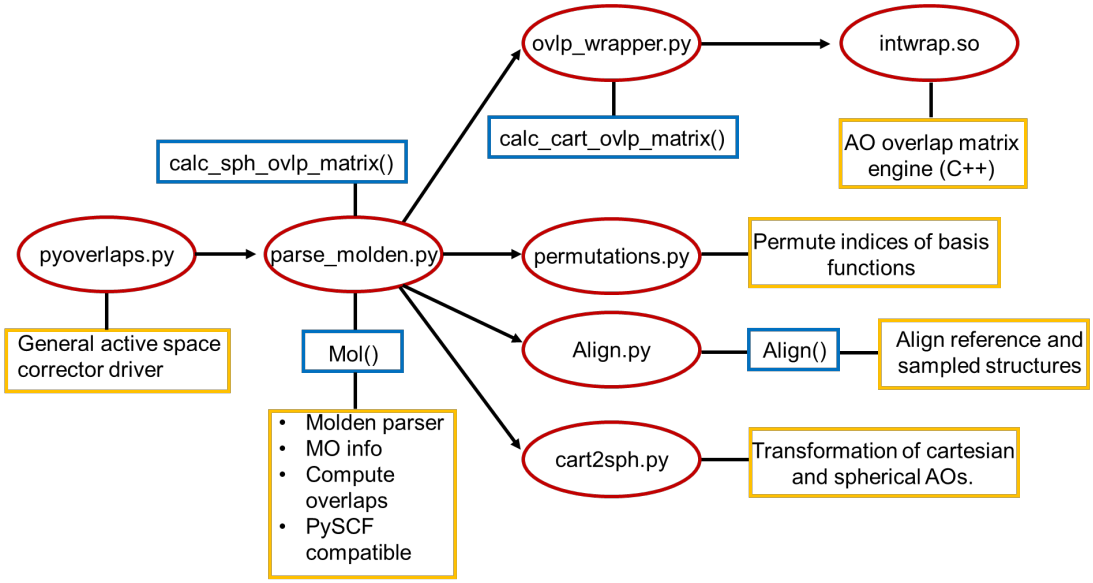


Figure 4: Schematic representation of the structure and main dependencies of the pyoverlaps.py script for the automatic correction of the active space on an ensemble of geometries. The red ovals represent a script or a program, the blue rectangles represent some functions and classes that can be used as an API, and the yellow boxes provide a brief description of the function of each script.

Applications

In this section, we illustrate the application of the MoBioTools toolkit to set up the QM/MM input files for four different systems. Specifically, the toolkit was employed to compute the reduction potential of the five nucleobases in acetonitrile, the interaction energy between tyrosine and water and its energy decomposition analysis, the absorption spectrum of an azobenzene derivative integrated into an voltage-gated ion channel, and the emission spectrum of the luciferine chomphore embedded in the luciferase protein. In the following subsections, for each of the four applications, first, the computational details are introduced, and then the results are briefly discussed.

Reduction Potential: Canonical Nucleobases in Acetonitrile

Protocol. In this section the application of the toolkit to compute reduction potentials from an ensemble of geometries is discussed. We have computed the reduction potential associated with the half reaction of oxidation of each one of the five canonical nucleobases (adenine, guanine, thymine, cytosine and uracil), in a solution of acetonitrile. The reduction potential is associated with the free energy of the reduction half reaction ΔG_{red} by means of:

$$\Delta E^0_{red} = \frac{\Delta G_{red}}{nF} - E^0_{red,SHE} \tag{1}$$

where F is the Faraday constant, n is the number of exchanged electrons and $E^0_{red,SHE}$ is the potential of the standard hydrogen electrode (SHE).[72–75] A value of 4.28 V determined in an earlier work is used as $E^0_{red,SHE}$. In the present work we have employed the Marcus theory[76–79] in a framework first introduced by Warshel,[80,81] in which it was shown that within a linear response regime, the vertical energy difference between the oxidized and the reduced states could be defined as the reaction coordinate of the redox process to compute the free energy. In this context, two MD simulations were performed: one for the reduced state ($\mathbf{r}_R$) and one for the oxidized state ($\mathbf{r}_O$). In what follows, the energy differences between the

two states computed on top of geometries of the phase space of $\mathbf{r}_R$ ($\Delta E_{R\rightarrow O}\{\mathbf{r_R}\}$) will be referred to as the vertical ionization energy (VIE) and the energy differences computed on top of geometries of the phase space of $\mathbf{r}_O$ ($\Delta E_{O\rightarrow R}\{\mathbf{r_O}\}$) will be referred to as the vertical attachment energy (VAE). If the linear response conditions are satisfied, it can be shown that the free energy in Equation 1 can be determined from the averages of the VIE and the VAE, as follows:

$$\begin{aligned}\Delta G_{red} =& \frac{1}{2}(\langle \Delta E_{R\rightarrow O}\{\mathbf{r}_R\}\rangle - \langle \Delta E_{O\rightarrow R}\{\mathbf{r}_O\}\rangle) - G_{e(g)} \\ =& \frac{1}{2}\Big(\langle VIE\rangle_R - \langle VAE\rangle_O\Big) - G_{e(g)}\end{aligned} \tag{2}$$

where the subscripts on the averages refer to the phase space from which the geometries were extracted to compute the vertical energy differences. The term $G_{e(g)} = -0.867$ kcal/mol[82,83] is the free energy of the electron in the gas phase according to the Fermi-Dirac statistics. It has to be included in the equations because the reference SHE potential value of 4.28 V[84] also includes this contribution. This methodology has been widely used to compute redox potentials of redox half reactions in the past.[22,23,25,85]

For each of the five nucleobases, a classical MD simulation was performed on each of the two oxidations states under study. The system setup for each MD simulation was done using the AmberTools20[55] package and a set of different homemade scripts. For both the oxidized and the reduced forms a geometry optimization was performed at the PBEOP/6-311G(d)[86–88] level of theory using the NWChem[57] software. Implicit solvation effects were introduced by means of the COSMO[89] solvation model, using acetonitrile as solvent. Electrostatic potential (ESP) charges were obtained from the same DFT calculation. Bond and bond angle parameters were obtained using the Hessian matrix of the optimized geometry by means of the Semiario method.[90] Parameters for dihedral angles, improper torsions and van der Waals non-bonded terms were taken from the generalized amber force field (GAFF).[91] Each nucleobase was solvated in a truncated octahedron with a buffer of 25.0 Å with approximately 1200 acetonitrile molecules by using the packmol[92] software. For the oxidized

form, a chloride anion was also added to neutralize the system. The force field parameters for acetonitrile were obtained in the same manner as in the case of the five nucleobases.

Each system (for both oxidation states) was minimized for 10000 steps using the steepest descent algorithm[93] for the 5000 first steps and the conjugate gradient algorithm[94] for the last 5000 steps. Afterwards, a progressive heating to 300 K was performed for 1 ns at constant volume (NVT). The first 500 ps were employed to drive the system to the desired temperature and the last 500 ps were destined to equilibrate the structure of the system. The Langevin thermostat[95] was applied to control the temperature taking into account a collision frequency of 2 ps^{-1}. After that, the volume of the system and the appropriate density were equilibrated carrying out a 1 ns simulation in the NPT ensemble. Finally, an additional 500 ns production simulation was performed in the NPT ensemble. The pressure was maintained constant at 1 bar employing the Berendsen barostat[96] with isotropic position scaling and a pressure relaxation time of 2 ps. The electrostatic interactions were computed during the full protocol using the particle-mesh Ewald method[97] with a grid spacing of 1.0 Å. In the case of the nonbonded interactions a 10 Å cutoff was chosen. The SHAKE[98] algorithm restrained the bonds involving hydrogen atoms and a time step of 2 fs was used during the heating, equilibration and production stages.

For each of the trajectories in the oxidized and in the reduced states, 200 snapshots were fetched randomly from the last 450 ns of the production trajectories. The script `main_qminputs.py` was employed to select the snapshots and to automatically generate the input files for the sampled geometries in two situations: one to compute the vertical energy differences within an electrostatic embedding QM/MM framework, in which the nucleobases represented the QM part and the acetonitrile molecules were present as fixed-point charges, and one in which the point charges were removed and replaced by COSMO to compute the vertical energy differences by QM/COSMO computations. In what follows the former and the latter situations will be referred to as the dynamic Marcus explicit (DME) and the dynamic Marcus implicit (DMI) approaches, respectively, to compute redox potentials. In

both cases the QM subsystem was described at the PBEOP/6-311G(d) level of theory.

Results. We now analyze the computed reduction potentials of the five canonical nucleobases in a solution of acetonitrile. As said above, we have performed the computations within the framework of the Marcus theory using two different solvation models for the acetonitrile solvent: explicit solvation (direct Marcus explicit - DME) and implicit solvation (direct Marcus implicit - DMI). Figures 5a,b show the main input and the template files used to generate the QM/MM input files for the DME calculations. It should be emphasized that the same ensemble of geometries was employed for the DMI apporach, whereby the explicit point charges were replaced by an implicit solvation cavity. To generate the inputs for the DMI approach, it suffices to replace the

```
&externchg
&end
```

entry with the following section:

```
&cosmo
 dielec  37.5
&end
```

a)

```
&main
tpl   = nwchem
top   = nucleobase.prmtop
traj  = nucleobase.nc
qmmask = :1
geoms =  N
&end
```

b)

```
&header
start geom0
title "geom0.in“
memory total 12 gb
&end

&basis
 * library aug-cc-pvdz
&end

&chgspin
0,1
&end

&externchg
&end

&dft
 xc pbeop
 convergence nolevelshifting
 odft
 mult 1
 odft
&end

&tasks
task dft
&end
```

c)

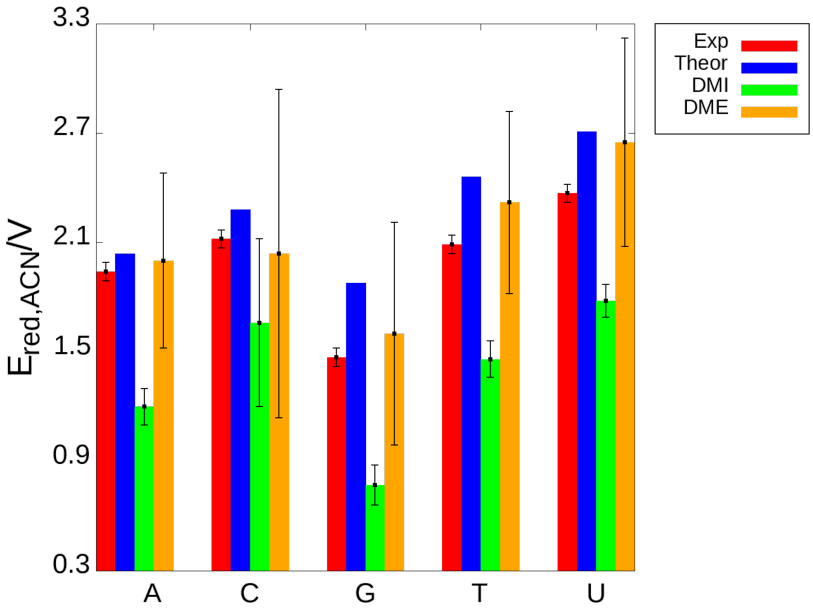


Figure 5: a) Main input file for the MoBioTools input generator. b) Template file for the NWChem software used to generate a set of 200 QM/MM input files for each of the nucleobases. The QM/COSMO inputs are generated using the &cosmo section instead of the &externchg entry. c) One electron reduction potentials obtained using the direct Marcus explicit (orange) and the direct Marcus implicit (green) approaches. The experimental results by Seidel[99] and the computational results by Crespo-Hernández[100] (blue) are also reported for comparison. The black lines represent standard deviations.

Figure 5 c shows the reduction potentials of an ensemble of 200 geometries for each one of the nucleobases computed using the DME and the DMI approaches. For comparison, we have also reported the experimental redox potentials determined by Seidel and co-workers,[99] and the redox potentials computed at the B3LYP/6-31++(d,p) using a thermodynamic cicle by Crespo-Hernández and co-workers.[100] It can be evidenced that the DMI approach underestimates the oxidation potentials for all five nucleobases, whereas the DME computations provide the most accurate results, with values that differ at most by 0.1 V from the experimental values. Interestingly, the opposite situation had been previously evidenced by some of us[25] in the case of the reduction potentials computed in water, for which the DMI approach had outperformed the DME results. This could indicate that the implicit solvation model performs better for aqueous solvent than for acetonitrile. However, additional calculations and analyses would be necessary to corroborate it.

Energy Decomposition Analysis: Tyrosine in Water

Protocol. This section aims at illustrating the ability of the MoBioTools toolkit to easily process MD trajectories into input files for the calculation of interaction energies and their different quantum mechanical energy contributions of a system consisting of two fragments, at the QM/MM level of theory. The chosen test system consists of the aminoacid tyrosine in a water solution, which undergoes configurational sampling by means of classical MD, followed by the Energy Decomposition Analysis (EDA) of the resulting geometry ensemble. An electron density based QM/MM-EDA scheme[67,101,102] is used for such purpose, allowing the decomposition of the total interaction energy in its electrostatic, polarization (induction and dispersion), and Pauli or exchange-repulsion components.

The geometry of tyrosine was taken from reference 103. Then, the antechamber and tleap modules from AmberTools20[55] were employed to solvate the aminoacid in a truncated octahedron simulation box with a buffer of 22.0 Å. The tyrosine was described by the GAFF[91] force field, and the water molecules were described by the TIP3P[104] model. Once the system

was set up, the next step was to minimize, heat and equilibrate the structure and density of the solvated tyrosine by classical MD using the AMBER20[55] package. At first, a minimization was carried out using the steepest descent method for 5000 steps and the conjugate gradient method for another 5000 steps. Then, the system was heated in the NVT ensemble applying a Langevin thermostat[95] with 1.0 ps^{-1} of collision frequency from 0 to 303.15 K for 1 ns. After the heating process, a production simulation of 100 ns was run in the NPT ensemble applying the Monte Carlo barostat to keep the pressure at 1.0 bar. For all these steps, the cutoff and switching distances for nonbonded interactions were limited to 12.0 and 10.0 Å, respectively. During the whole protocol, a time step of 2 fs was employed and the bonds involving hydrogen atoms were constrained by using the SHAKE[98] algorithm.

After the equilibration of the system, single point QM/MM calculations for 100 equally spaced snapshots selected from the last 50 ns of the simulation were performed with Gaussian16[51] using the M062X[105] functional and the cc-pVDZ[106] basis set. The QM region included the tyrosine molecule and the 10 water molecules closest to tyrosine, with the remaining solvent included as point charges assigned by the TIP3P model. To perform the EDA, three single point QM/MM calculations were carried out for each snapshot, corresponding to the full system (QM region and MM point charges) and the fragments on which to compute the interaction energy. The first fragment consisted of the QM solute in the presence of the basis set of the QM water molecules, whilst the second fragment included all QM water molecules as well as the basis set from the solute surrounded by the MM electrostatic embedding. Basis functions of absent atoms are aimed at accounting for the basis set superposition error (BSSE) by means of the Counterpoise correction (CPC).[65,66] The script `main_qminputs.py` was used for the selection of the snapshots from the MD trajectory, the definition of the QM and MM regions, and the automatic generation of the three input files required by the EDA calculation.

Results. We now proceed to discuss the energy decomposition analysis of the interaction

energy of tyrosine in water. Figures 6a,b show the main input and the template files used to generate the QM/MM input files from an ensemble of 100 MD geometries with the size of the QM region set to fit 10 water molecules. In this case, the two subsystems for which the interaction energy was computed are the tyrosine molecule on the one hand, and the 10 QM water molecules plus the water molecules defined in the MM region on the other hand. Thus, three input files are generated for each geometry to correct for the BSSE: one for each subsystem (with the basis set of the other subsystem) and one for the complex. This option is enabled by including the `&bsse` section in the template file (Figure 6b), where the two subsystems can be defined by using Amber masks. In the present case no arguments were provided, so that the toolkit will assume that the two subsystems consist of the `qmmask` argument on the one hand, and the combination of the `solvmask` and the `closest` arguments on the other hand.

Figure 6c shows the normal probability distribution of each interaction energy component for the ensemble of 100 MD geometries. The Kolmogórov-Smirnov test[107] (5% significance level) was used to test for normality. In regard with the width of the energy distributions, the relatively large standard deviations observed (electrostatic 11.94 kcal/mol, Pauli 16.08 kcal/mol, polarization 7.92 kcal/mol, dispersion 4.68 kcal/mol, induction 3.60 kcal/mol, total 7.46 kcal/mol) restate the importance of the carried conformational sampling even for simple systems like the one studied here. In addition, as can be seen from the averaged values, the Pauli component is the main energy contribution for tyrosine in water, whilst the electrostatic energy is the most relevant among the attractive components, followed by dispersion and induction. This qualitative distribution of the interaction energy components is in consonance with the structural features of tyrosine and their interaction with a highly polar solvent like water.

a)

```
&main
tpl   = gaussian
top   = tyr_solv.prmtop
traj  = tyr_solv.nc
qmmask = :1
geoms =  0 99
solvmask = :WAT
closest = 10
&end
```

b)

```
&header
%chk=gau.chk
&end

&route
#P M062X/6-31g* Charge NoSym integral=ultrafinegrid
&end

&externchg
&end

&bsse
&end

&chgspin
0,1,0,1,0,1
&end
```

c)

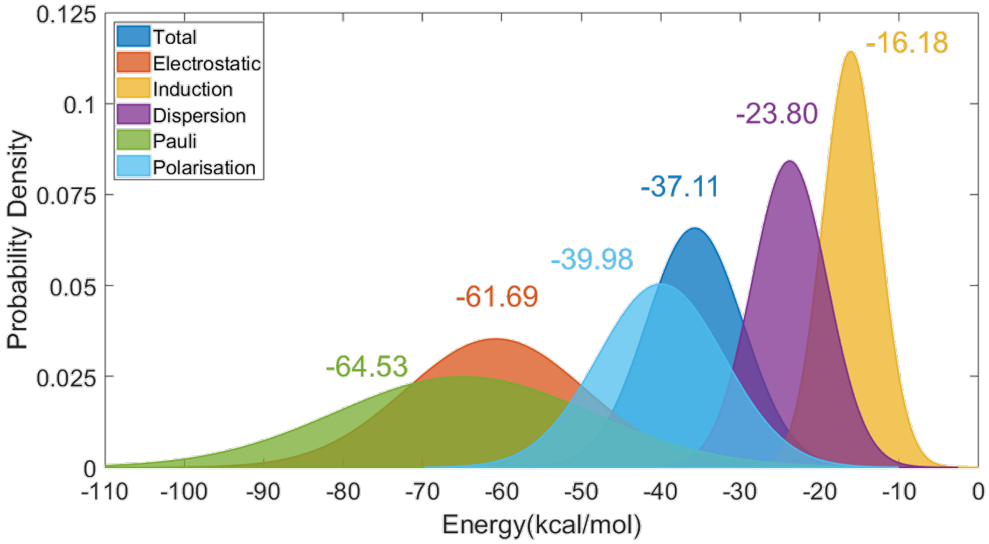


Figure 6: a) Main input file for the MoBioTools input generator. b) Template file for the Gaussian software used to generate a set of 100 QM/MM input files for the tyrosine molecule in water. c) Normal probability distributions for each interaction energy component in kcal/mol obtained from a sample of 100 geometries from a MD trajectory. The Pauli exchange-repulsion is represented with opposite sign. The average value of each distribution is reported in kcal/mol

Absorption Spectrum: p-Diaminoazobenzene Integrated into the Human $Na_V1.4$ Channel

Protocol. In this section we used the machinery of the MoBioTools toolkit to generate the input files for the calculation of the absorption spectrum of a photoswitch bound to a protein ion channel at the QM/MM multistate complete active space second-order perturbation theory (MS-CASPT2)[108–110] level for an ensemble of geometries. In this regard, the script `pyoverlaps.py` was used to perform the generation of the CASSCF input files for the different snapshots selected, the CASSCF calculations themselves *via* its interface with the OpenMolcas[59] software, and the correction of the active space for the ensemble of geometries. The script `main_qminputs.py` was employed to generate the input files for the CASPT2 calculations for those geometries for which the active space was successfully recovered.

The system under study consists of the p-diaminoazobenzene (p-DAZ) molecule inside the human's brain voltage gated ion channel $Na_V1.4$. We obtained an ensemble of geometries by means of a 100 ns classical MD simulation, for which the initial geometry was retrieved from a previous work.[111] For completeness, we report the preparation of the initial geometry and the obtainment of the force field parameters in what follows. A truncated pore model of the $Na_V1.4$ was constructed in VMD[112] based on the cryo-electron microscopic structure file of the channel in complex with its β_1 subunit (PDB ID: 6AGF).[113] The simulation comprised only the S5-S6 subunits of each domain, that form the central pore. The protein was aligned along the z-axis using the Positioning Proteins in Membrane (PPM) web server,[114] placed inside a 1-palmitoyl-2-oleoyl-sn-glycero-3-phosphocholine (POPC) lipid bilayer (xy length: 100 x 100 lipid components) and solvated in a rectangular box with aqueous solvent and NaCl at a concentration of 0.15 mol/L using CHARMM-GUI Bilayer Builder.[115] The potential parameters for the protein and lipids were taken from the CHARMM36m force field,[116] and the TIP3P model was employed for the water molecules.[104] The ligand p-DAZ was constructed with the IQmol molecular viewer,[117] the geometry was optimized at MP2/6-

31G* level of theory and used for the calculation of the ESP charges at the Hartree-Fock/6-31G* level of theory. The ligand force constants for the dihedral angles –C–N=N–C– and -C–C–N=N– were taken from a previous work,[118] while the rest of the parameters were taken from CGenFF.[119] After the system set up, in a previous work Gaussian accelerated MD simulations were run to identify the possible binding pockets of p-DAZ into the $Na_V1.4$ channel.[111] In the present work, one snapshot from the most stable binding pocket was selected as initial snapshot for a 100 ns MD simulation.

The 100 ns MD simulation of this work was run in the isothermal–isobaric ensemble (NPT) with a Monte Carlo barostat and a semiisotropic pressure scaling at 303.15 K and 1.01 bar. A timestep of 2 fs was used and the bond lengths involving hydrogen atoms were constrained using the SHAKE algorithm. The nonbonded interactions cutoff radius and switching distance were set to 12.0 Å and 10.0 Å, respectively. A force constant of 20 kcal/(mol $Å^2$) was imposed on top of the –C–N=N–C– and -C–C–N=N– dihedral angles to avoid unrealistic torsion motions and to keep the photoswitch on the trans configuration. The script `pyoverlaps.py` was employed to fetch 100 equidistant geometries from the MD trajectory to perform the state average (SA) CASSCF computations with an eventual correction of the active space. For each geometry, equal weights of the first 10 states has been used to generate the reference wavefunction for the MS-CASPT2 single point calculations to compute the absorption spectrum. The active space included a total of 12 electrons in 10 orbitals (12,10), which comprises two doubly occupied nitrogen lone pairs, four doubly occupied π orbitals and four unoccupied π^* orbitals, as shown in Figure 7.

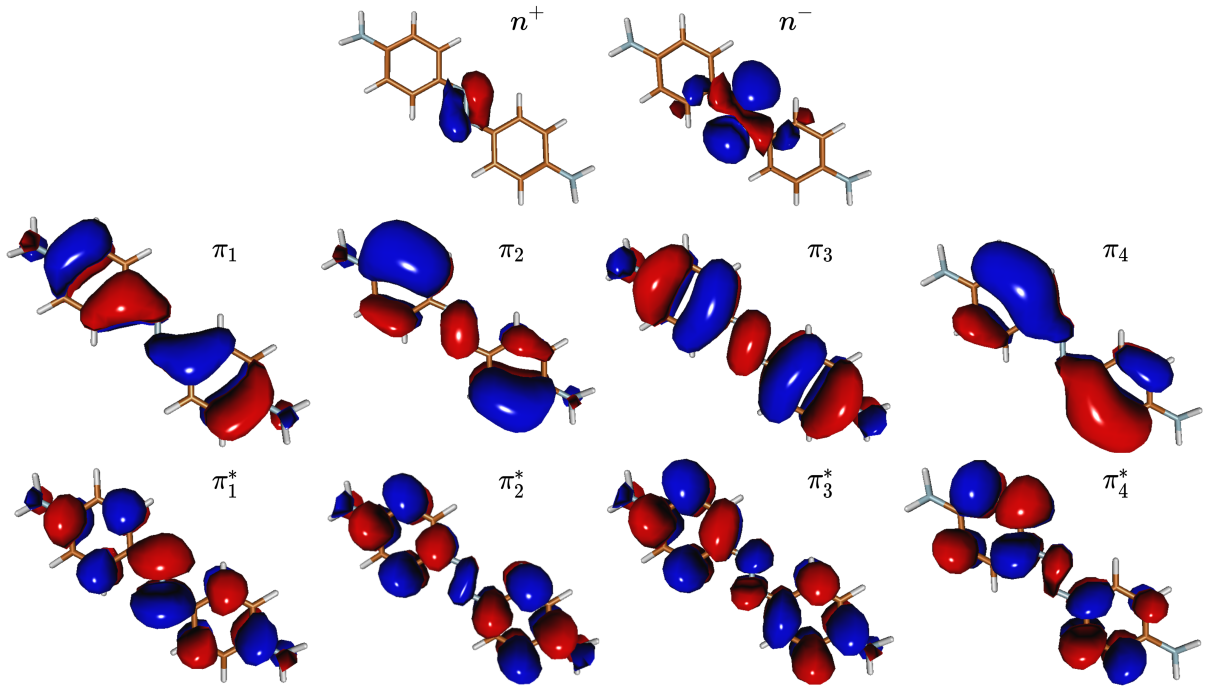


Figure 7: Representation of the molecular orbitals included in the active space. The n^+ and n^- orbitals are the lone pairs of the central diazine bond, π_{1-4} are four occupied π orbitals and π^*_{1-4} are four unoccupied π^* orbitals.

The cc-pVDZ[106] basis set was used, and the Resolution of Identity approach was used to treat the two electron integrals.[120] In the MS-CASPT2 calculations, an imaginary shift of 0.2 Ha was added to the zero order Hamiltonian to preclude the inclusion of intruder states,[121] with an IPEA value of 0.0 Ha as previously recommended.[122] The restricted active space interaction (RASSI) protocol was used to compute the transition oscillator strengths to compare transition probabilities among the states studied.

Results. spAs described above, the software was used to calculate the absorption spectrum of an ensemble of geometries obtained from 100 ns of a classical MD trajectory of the photoswitch embedded in the protein. The script `pyoverlaps.py` allowed to correct the active space in correspondence with a reference vacuum calculation. Figure 8a,b shows the main input and template files used to generate the QM/MM input files from an ensemble of 100 MD geometries with the size of the QM region set to contain only p-diaminoazobenzene. The absorption spectrum displays a strong UV band ($\lambda_{\max} = 383$ nm) which arises from a symmetry allowed $\pi \rightarrow \pi^*$ transition. Intensity borrowing is observed for higher-energy lying $\pi\pi^*$ states at 220-330 nm as well as for the symmetry forbidden n $\rightarrow \pi^*$ transition appearing at 550-550 nm. This coupling between different electronic states of different symmetry shows the importance of sampling an ensemble of geometries around the equilibrium geometry, instead of computing the absorption spectrum only at the minimum-energy geometry.

a)

```
&main
tpl   = molcas
top   = system.parm7
traj  = 100f.nc
qmmask = :587
geoms = 0 100 1
solvmask =
&end
```

b)

```
&header
title = p-AZO molecule
group = C1
basis set = cc-pVDZ spherical all
RICD
&end

&seward
&end

&chgspin
0,1,0,1,0,1
&end

&guess
jobiph
&end

&caspt2
IPEA= 0.0
SHIFT= 0.0
IMAGinary= 0.2
maxiter=100
multistate = 10 1 2 3 4 5 6 7 8 9 10
PROP
&end

&rassi
Nr of JobIph = 1 10; 1 2 3 4 5 6 7 8 9 10
Ejob
&end
```

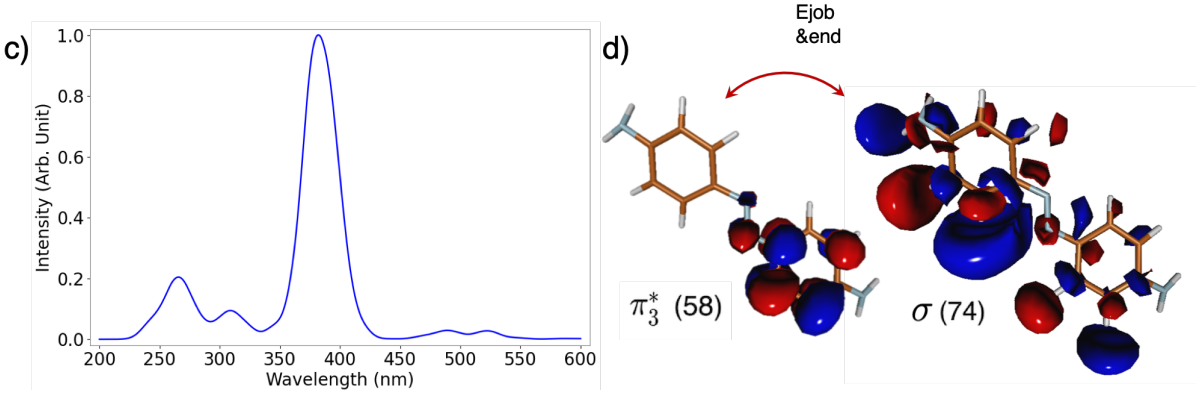


Figure 8: a) Main input file for the MoBioTools input generator. b) Template file for the OpenMolcas software used to generate a set of 100 QM/MM input files for the p-diaminoazobenzene molecule embedded in the $Na_v1.4$ ion channel. c) Absorption spectrum of p-diaminoazobenzene molecule embedded in the $Na_v1.4$ ion channel. d) Example of a σ orbital swapped for a π orbital corresponding to the reference wavefunction.

Emission and Absorption Spectra: Luciferine/Luciferase Complex

Protocol. As evidenced in the previous sections, the MoBioTools toolkit can be employed to compute electronically excited states for an ensemble of geometries of a molecular system. As a final example, we illustrate its application to compute the absorption and the emission spectra of the oxyluciferin chromophore in the luciferase enzyme system. In this case, two classical MD simulations were evolved using the AMBER20[55] software to generate the ensemble of geometries for the electronic ground state and for the first excited state, as explained below. These trajectories were proccesed by the `main_qminputs.py` script to generated the QM/MM input files for the subsequent excited state calculations.

The luciferine/luciferase complex was constructed using PyMol[123] based on the X-ray diffraction structure file of firefly luciferase crosslinked in the second catalytic form (PDB ID: 4G37[124]). The complex was solvated with the tleap module of AmberTools20[55] by TIP3P[104] water molecules within a truncated octahedral box, ensuring a solvent shell of at least 12 Å from any solute molecule. The luciferase enzyme was described using the ff19SB force field,[125] whereas the GAFF force filed was used to describe the dihedral and improper torsions and Lennard-Jones parameters of the chromophore. Bond and bond angle parameters for both the S_0 and S_1 electronic states of the oxyluciferin molecule were obtained by the Seminario method. The optimized geometries and the Hessian matrices were obtained using the Gaussian16[51] software by means of DFT and TD-DFT calculations, respectively. In particular, for both ground and excited state calculations the B3LYP[126–128]/6-311G(2d,p)[129–131] level of theory was employed. The ESP charges for the two electronic states were computed using the same level of theory.

A minimization was performed for 5000 steps using the steepest descent algorithm, and for another 5000 steps using the conjugate gradient algorithm. Afterwards, a progressive heating was performed from 0 to 300 K for 500 ps in the NVT ensemble using a timestep of 2 fs , followed by a constant temperature simulation for another 500 ps in the same ensemble, using a Langevin thermostat[95] with 1.0 ps^{-1} of collision frequency. Then, a production simulation of 100 ns was performed in the NPT ensemble, again with a timestep of 2 fs. A Berendsen barostat[96] and thermostat were used throughout the simulation to keep the pressure at 1.0 bar and the temperature at 300 K, respectively. The electrostatic interactions were computed using the particle-mesh Ewald[97] method, with a grid spacing of 1.0 Å. For the non-bonded interactions cutoff and switching distances of 9.0 and 7.0 Å were chosen, respectively. Bonds involving hydrogen atoms were constrained using the SHAKE[98] algorithm.

For both MD simulations in the S_0 and S_1 states, the first 25 ns of the production run were discarded as the equilibration time of the protein structure. The script `main_qminputs.py`

was used to randomly extract 200 geometries from the last 75 ns of the MD trajectory, and to generate the QM/MM input files to compute the absorption spectrum (for the S_0 trajectory) and the emission spectrum (for the S_1 trajectory) of the chromophore. The QM/MM calculations were performed by Gaussian16,[51] where 10 excited states were computed at the B3LYP/6-311G(2d,p) level of theory for each of the snapshots.

Results. As stated above, the toolkit was used to generate 200 Gaussian16 input files from the last 75 ns of the classical MD trajectory in the S_0 state (for the absorption spectrum) and in the S_1 state (for the emission spectrum). As the structure of the input files for the toolkit is the same in both situations, for simplicity we only report the main and the template files for the computation of the absorption spectrum of the oxyluciferin molecule in Figures 9a,b. Figure 9c shows the computed absorption and the emission spectra of the oxyluciferin/oxiluciferase complex. In the case of the absorption spectrum, two bands can be observed, which correspond to transitions to the S_1 and S_3 electronic states. The orbitals involved in the first one are shown in Figure 9d. The emission spectrum present a band centered at 567 nm, which is in good agreement with the experimental value of 553 nm.[132]

a)

```
&main
tpl   = gaussian
top   = luci.prmtop
traj  = luci.nc
qmmask = :1
geoms =  0 199
&end
```

b)

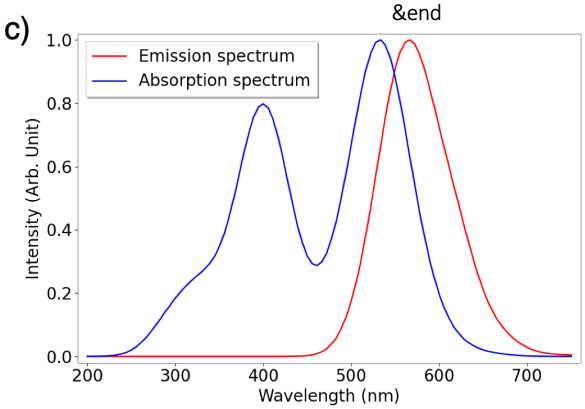
```
&header
%chk=geom.chk
&end

&route
#P B3LYP/6-311(2d,p) Charge NoSym TD(singlets,nstates=10)
&end

&externchg
&end

&bsse
&end

&chgspin
-1,1
&end
```



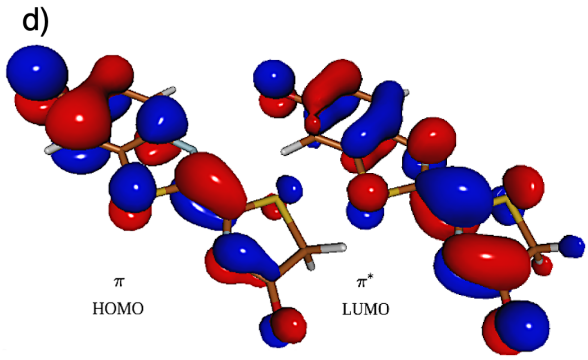


Figure 9: a) Main input file for the MoBioTools input generator. b) Template file for the Gaussian16 software used to generate a set of 200 QM/MM input files for the oxyluciferin molecule embedded in the luciferase enzyme system. c) Emission and absroption spectra of the oxyluciferin/luciferase solvated complex. d) Orbitals that participate in the transition from the S_1 and to the S_1 state.

Conclusions

In this work we have introduced MoBioTools, a toolkit that allows for a straightforward setting up of a set of QM/MM calculations from an ensemble of geometries, so as to compute various physical properties at a high level of accuracy, while accounting for conformational sampling. The toolkit has been applied to four different situations, showing a large degree of versatility. Specifically, we have used it to set up the calculations of the reduction potentials of the five canonical nucleobases in acetonitrile and the interaction energies (and corresponding energy decomposition analysis) of a system consisting of tyrosine solvated by water. In the latter application, we have evidenced how the toolkit defines the environment for cases in which part of it needs to be included within the quantum mechanical subsystem. MoBioTools has also been used to compute electronically excited state properties, such as the emission and absorption spectra of some chromophores embedded in complex biological media. In the case that a multiconfigurational method, such as CASSCF, is employed to compute the excited states of the chromophore, we have shown that the toolkit can also be used to automatically correct the active space for those geometries of the ensemble in consideration for which a first CASSCF wavefunction optimization has provided a different active space from the one used as reference. The toolkit developed and presented here will be useful for the computational chemistry community, as it attempts to carry out in an automatic way tasks that need to be performed within the framework of QM/MM, but that otherwise would be tedious and time consuming.

Acknowledgement

We acknowledge the generous allocation of computer time at the Centro de Computación Científica at the Universidad Autónoma de Madrid (CCC-UAM). This work was supported by the Spanish Ministry of Science and Innovation through the projects PID2019-110091GB-I00 and PID2020-117806GA-I00 funded by MCIN/AEI/10.13039/501100011033, the María

de Maeztu (CEX2018-000805-M) Program for Centers of Excellence in R&D, Comunidad de Madrid through the Attraction of Talent Program (Grant ref 2018-T1/BMD-10261), Xunta de Galicia through the project GRC2019/24, the European Social Fund through the Programa Operativo de Empleo Juvenil y la Iniciativa de Empleo Juvenil, the Spanish Ministry of Education and Vocational Training through the FPU19/02292 grant, the Universidade de Vigo through the PREUVIGO-21 grant, and the Universidad Autónoma de Madrid through the Ayudas para el Fomento de la Investigación en Estudios de Máster program.

Data and Software Availability

The authors declare no conflict of interest nor competing financial interest. The software CHARMM-GUI (`https://www.charmm-gui.org/`) was used[111] to set up the human voltage-gated $Na_V1.4$ ion channel. The package packmol[92] (`http://leandro.iqm.unicamp.br/m3g/packmol/home.shtml`) was used to prepare the systems consisting of the five canonical nucleobases in acetonitrile. The package tleap and antechamber from the AmberTools20[55] toolkit were used to generate the topology and the coordinate files for the MD simulations. The CUDA version[133,134] of the AMBER20[55] (`https://ambermd.org/`) software was used to perform the classical MD simulations. The MoBioTools toolkit (`https://github.com/mobiochem/MoBioTools`) presented in this work was used to automatically generate the QM/MM input files. The QM/MM computations themselves were performed using the Gaussian16[51] (`https://gaussian.com/`), NWChem[57] (`https://www.nwchem-sw.org/`) and OpenMolcas[58,59] (`https://gitlab.com/Molcas/OpenMolcas`) packages. The trajectories were visualized with the Visual Molecular Dynamics (VMD, `https://www.ks.uiuc.edu/Research/vmd/`) package, and the molecular orbitals images were generated with the molden software (`https://www3.cmbi.umcn.nl/molden/`).

References

(1) Warshel, A.; Levitt, M. Theoretical studies of enzymic reactions: Dielectric, electrostatic and steric stabilization of the carbonium ion in the reaction of lysozyme. *J. Mol. Biol.* **1976**, *103*, 227–249.

(2) Rivail, J.-L.; Rinaldi, D. A quantum chemical approach to dielectric solvent effects in molecular liquids. *Chem. Phys.* **1976**, *18*, 233–242.

(3) Senn, H. M.; Thiel, W. QM/MM Methods for Biomolecular Systems. *Angew. Chem. Int. Ed.* **2009**, *48*, 1198–1229.

(4) van der Kamp, M. W.; Mulholland, A. J. Combined Quantum Mechanics/Molecular Mechanics (QM/MM) Methods in Computational Enzymology. *Biochemistry* **2013**, *52*, 2708–2728.

(5) Warshel, A. Multiscale Modeling of Biological Functions: From Enzymes to Molecular Machines (Nobel Lecture). *Angew. Chem. Int. Ed.* **2014**, *53*, 10020–10031.

(6) Sousa, S. F.; Ribeiro, A. J. M.; Neves, R. P. P.; Brás, N. F.; Cerqueira, N. M. F. S. A.; Fernandes, P. A.; Ramos, M. J. Application of quantum mechanics/molecular mechanics methods in the study of enzymatic reaction mechanisms. *Wiley Interdiscip. Rev. Comput. Mol. Sci.* **2017**, *7*, e1281.

(7) Bondanza, M.; Nottoli, M.; Cupellini, L.; Lipparini, F.; Mennucci, B. Polarizable embedding QM/MM: the future gold standard for complex (bio)systems? *Phys. Chem. Chem. Phys.* **2020**, *22*, 14433–14448.

(8) Lipparini, F.; Mennucci, B. Hybrid QM/classical models: Methodological advances and new applications. *Chem. Phys. Rev.* **2021**, *2*, 041303.

(9) Cramer, C. J.; Truhlar, D. G. Implicit Solvation Models: Equilibria, Structure, Spectra, and Dynamics. *Chem. Rev.* **1999**, *99*, 2161–2200.

(10) Orozco, M.; Luque, F. J. Theoretical Methods for the Description of the Solvent Effect in Biomolecular Systems. *Chem. Rev.* **2000**, *100*, 4187–4226.

(11) Tomasi, J.; Mennucci, B.; Cammi, R. Quantum Mechanical Continuum Solvation Models. *Chem. Rev.* **2005**, *105*, 2999–3094.

(12) Klamt, A. The COSMO and COSMO-RS solvation models. *Wiley Interdiscip. Rev. Comput. Mol. Sci.* **2011**, *1*, 699–709.

(13) Bakowies, D.; Thiel, W. Hybrid Models for Combined Quantum Mechanical and Molecular Mechanical Approaches. *J. Phys. Chem.* **1996**, *100*, 10580–10594.

(14) Roca, M.; Andrés, J.; Moliner, V.; Tuñón, I.; Bertrán, J. On the Nature of the Transition State in Catechol O-Methyltransferase. A Complementary Study Based on Molecular Dynamics and Potential Energy Surface Explorations. *J. Am. Chem. Soc.* **2005**, *127*, 10648–10655.

(15) Senn, H. M.; Thiel, W. QM/MM studies of enzymes. *Curr. Opin. Chem.* **2007**, *11*, 182–187.

(16) Brás, N. F.; Moura-Tamames, S. A.; Fernandes, P. A.; Ramos, M. J. Mechanistic studies on the formation of glycosidase-substrate and glycosidase-inhibitor covalent intermediates. *J. Comput. Chem.* **2008**, *29*, 2565–2574.

(17) Bowman, A. L.; Grant, I. M.; Mulholland, A. J. QM/MM simulations predict a covalent intermediate in the hen egg white lysozyme reaction with its natural substrate. *Chem. Commun.* **2008**, 4425–4427.

(18) Adamczyk, A. J.; Cao, J.; Kamerlin, S. C. L.; Warshel, A. Catalysis by dihydrofolate reductase and other enzymes arises from electrostatic preorganization, not conformational motions. *Proc. Natl. Acad. Sci. USA* **2011**, *108*, 14115–14120.

(19) van der Kamp, M. W.; Żurek, J.; Manby, F. R.; Harvey, J. N.; Mulholland, A. J. Testing High-Level QM/MM Methods for Modeling Enzyme Reactions: Acetyl-CoA Deprotonation in Citrate Synthase. *J. Phys. Chem. B* **2010**, *114*, 11303–11314.

(20) Ranaghan, K. E.; Morris, W. G.; Masgrau, L.; Senthilkumar, K.; Johannissen, L. O.; Scrutton, N. S.; Harvey, J. N.; Manby, F. R.; Mulholland, A. J. Ab Initio QM/MM Modeling of the Rate-Limiting Proton Transfer Step in the Deamination of Tryptamine by Aromatic Amine Dehydrogenase. *J. Phys. Chem. B* **2017**, *121*, 9785–9798.

(21) Lu, X.; Fang, D.; Ito, S.; Okamoto, Y.; Ovchinnikov, V.; Cui, Q. QM/MM free energy simulations: recent progress and challenges. *Mol. Simul.* **2016**, *42*, 1056–1078.

(22) Diamantis, P.; Gonthier, J. F.; Tavernelli, I.; Rothlisberger, U. Study of the Redox Properties of Singlet and Triplet Tris(2,2'-bipyridine)ruthenium(II) ([Ru(bpy)3]2+) in Aqueous Solution by Full Quantum and Mixed Quantum/Classical Molecular Dynamics Simulations. *J. Phys. Chem. B* **2014**, *118*, 3950–3959.

(23) Diamantis, P.; Tavernelli, I.; Rothlisberger, U. Redox Properties of Native and Damaged DNA from Mixed Quantum Mechanical/Molecular Mechanics Molecular Dynamics Simulations. *J. Chem. Theory Comput.* **2020**, *16*, 6690–6701.

(24) Cárdenas, G.; Marquetand, P.; Mai, S.; González, L. A Force Field for a Manganese-Vanadium Water Oxidation Catalyst: Redox Potentials in Solution as Showcase. *Catalysts* **2021**, *11*.

(25) Lucia-Tamudo, J.; Cárdenas, G.; Anguita-Ortiz, N.; Díaz-Tendero, S.; Nogueira, J. J. Computation of Oxidation Potentials of Solvated Nucleobases by Static and Dynamic Multilayer Approaches. *ChemRxiv* **2022**,

(26) Eckert-Maksić, M.; Vazdar, M.; Ruckenbauer, M.; Barbatti, M.; Müller, T.; Lischka, H. Matrix-controlled photofragmentation of formamide: dynamics simulation in argon by nonadiabatic QM/MM method. *Phys. Chem. Chem. Phys.* **2010**, *12*, 12719–12726.

(27) Plasser, F.; Aquino, A. J. A.; Hase, W. L.; Lischka, H. UV Absorption Spectrum of Alternating DNA Duplexes. Analysis of Excitonic and Charge Transfer Interactions. *J. Phys. Chem. A* **2012**, *116*, 11151–11160.

(28) Nogueira, J. J.; Plasser, F.; González, L. Electronic delocalization, charge transfer and hypochromism in the UV absorption spectrum of polyadenine unravelled by multiscale computations and quantitative wavefunction analysis. *Chem. Sci.* **2017**, *8*, 5682–5691.

(29) Marquetand, P.; Nogueira, J. J.; Mai, S.; Plasser, F.; González, L. Challenges in Simulating Light-Induced Processes in DNA. *Molecules* **2017**, *22*.

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

(31) Nogueira, J. J.; Roßbach, S.; Ochsenfeld, C.; González, L. Effect of DNA Environment on Electronically Excited States of Methylene Blue Evaluated by a Three-Layered QM/QM/MM ONIOM Scheme. *J. Chem. Theory Comput.* **2018**, *14*, 4298–4308.

(32) De Vetta, M.; Menger, M. F. S. J.; Nogueira, J. J.; González, L. Solvent Effects on Electronically Excited States: QM/Continuum Versus QM/Explicit Models. *J. Phys. Chem. B* **2018**, *122*, 2975–2984.

(33) Boulanger, E.; Harvey, J. N. QM/MM methods for free energies and photochemistry. *Curr. Opin. Struct. Biol.* **2018**, *49*, 72–76.

(34) Marazzi, M.; Gattuso, H.; Fumanal, M.; Daniel, C.; Monari, A. Charge-Transfer versus Charge-Separated Triplet Excited States of [ReI(dmp)(CO)3(His124)(Trp122)]+ in Water and in Modified Pseudomonas aeruginosa Azurin Protein. *Chem. Eur. J.* **2019**, *25*, 2519–2526.

(35) Francés-Monerris, A.; Tuñón, I.; Monari, A. Hypoxia-Selective Dissociation Mechanism of a Nitroimidazole Nucleoside in a DNA Environment. *J. Phys. Chem. Lett.* **2019**, *10*, 6750–6754.

(36) Dokukina, I.; Nenov, A.; Garavelli, M.; Marian, C. M.; Weingart, O. QM/MM Photodynamics of Retinal in the Channelrhodopsin Chimera C1C2 with OM3/MRCI. *ChemPhotoChem* **2019**, *3*, 107–116.

(37) Avagliano, D.; Bonfanti, M.; Garavelli, M.; González, L. QM/MM Nonadiabatic Dynamics: the SHARC/COBRAMM Approach. *J. Chem. Theory Comput.* **2021**, *17*, 4639–4647.

(38) Boulanger, E.; Thiel, W. Solvent Boundary Potentials for Hybrid QM/MM Computations Using Classical Drude Oscillators: A Fully Polarizable Model. *J. Chem. Theory Comput.* **2012**, *8*, 4527–4538.

(39) Savelyev, A.; MacKerell Jr., A. D. All-atom polarizable force field for DNA based on the classical drude oscillator model. *J. Comput. Chem.* **2014**, *35*, 1219–1239.

(40) Curutchet, C.; Muñoz-Losa, A.; Monti, S.; Kongsted, J.; Scholes, G. D.; Mennucci, B. Electronic Energy Transfer in Condensed Phase Studied by a Polarizable QM/MM Model. *J. Chem. Theory Comput.* **2009**, *5*, 1838–1848.

(41) Adcock, S. A.; McCammon, J. A. Molecular Dynamics: Survey of Methods for Simulating the Activity of Proteins. *Chem. Rev.* **2006**, *106*, 1589–1615.

(42) Borrelli, K. W.; Vitalis, A.; Alcantara, R.; Guallar, V. PELE: Protein Energy Landscape Exploration. A Novel Monte Carlo Based Technique. *J. Chem. Theory Comput.* **2005**, *1*, 1304–1311.

(43) Dahl, J. P.; Springborg, M. The Morse oscillator in position space, momentum space, and phase space. *J. Chem. Phys.* **1988**, *88*, 4535–4547.

(44) Helgaker, T.; Uggerud, E.; Jensen, H. J. A. Integration of the classical equations of motion on ab initio molecular potential energy surfaces using gradients and Hessians: application to translational energy release upon fragmentation. *Chem. Phys. Lett.* **1990**, *173*, 145–150.

(45) Field, M. J. The pDynamo Program for Molecular Simulations using Hybrid Quantum Chemical and Molecular Mechanical Potentials. *J. Chem. Theory Comput.* **2008**, *4*, 1151–1161.

(46) Řezáč, J. Cuby: An integrative framework for computational chemistry. *J. Comput. Chem.* **2016**, *37*, 1230–1237.

(47) Larsen, A. H. et al. The atomic simulation environment—a Python library for working with atoms. *J. Phys. Condens. Matter* **2017**, *29*, 273002.

(48) Lu, Y.; Farrow, M. R.; Fayon, P.; Logsdail, A. J.; Sokol, A. A.; Catlow, C. R. A.; Sherwood, P.; Keal, T. W. Open-Source, Python-Based Redevelopment of the ChemShell Multiscale QM/MM Environment. *J. Chem. Theory Comput.* **2019**, *15*, 1317–1328.

(49) Martí, S. QMCube (QM3): An all-purpose suite for multiscale QM/MM calculations. *J. Comput. Chem.* **2021**, *42*, 447–457.

(50) Weingart, O.; Nenov, A.; Altoè, P.; Rivalta, I.; Segarra-Martí, J.; Dokukina, I.; Garavelli, M. COBRAMM 2.0 — A software interface for tailoring molecular electronic structure calculations and running nanoscale (QM/MM) simulations. *J. Mol. Model.* **2018**, *24*, 271.

(51) Frisch, M. J. et al. Gaussian 16 Revision C.01. **2016**, Gaussian Inc. Wallingford CT.

(52) Richter, M.; Marquetand, P.; González-Vázquez, J.; Sola, I.; González, L. SHARC: ab Initio Molecular Dynamics with Surface Hopping in the Adiabatic Representation Including Arbitrary Couplings. *J. Chem. Theory Comput.* **2011**, *7*, 1253–1258.

(53) Mai, S.; Marquetand, P.; González, L. Nonadiabatic dynamics: The SHARC approach. *Wiley Interdiscip. Rev. Comput. Mol. Sci.* **2018**, *8*, e1370.

(54) Barbatti, M.; Ruckenbauer, M.; Plasser, F.; Pittner, J.; Granucci, G.; Persico, M.; Lischka, H. Newton-X: a surface-hopping program for nonadiabatic molecular dynamics. *Wiley Interdiscip. Rev. Comput. Mol. Sci.* **2014**, *4*, 26–33.

(55) Case, D. et al. AMBER 2020. **2020**,

(56) Neese, F. Software update: the ORCA program system, version 4.0. *Wiley Interdiscip. Rev. Comput. Mol. Sci.* **2018**, *8*, e1327.

(57) Valiev, M.; Bylaska, E. J.; Govind, N.; Kowalski, K.; Straatsma, T. P.; Van Dam, H. J. J.; Wang, D.; Nieplocha, J.; Apra, E.; Windus, T. L.; de Jong, W. A. NWChem: A comprehensive and scalable open-source solution for large scale molecular simulations. *Comput. Phys. Commun.* **2010**, *181*, 1477–1489.

(58) Fdez. Galván, I. et al. OpenMolcas: From Source Code to Insight. *J. Chem. Theory Comput.* **2019**, *15*, 5925–5964.

(59) Aquilante, F. et al. Modern quantum chemistry with [Open]Molcas. *J. Chem. Phys.* **2020**, *152*, 214117.

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

(61) Roos, B. O. *Adv. Chem. Phys.*; John Wiley & Sons, Ltd, 2007; pp 399–445.

(62) Cárdenas, G.; Nogueira, J. J. An algorithm to correct for the CASSCF active space in multiscale QM/MM calculations based on geometry ensembles. *Int. J. Quantum Chem.* **2021**, *121*, e26533.

(63) Shepard, R. *Adv. Chem. Phys.*; John Wiley & Sons, Ltd, 2007; pp 63–200.

(64) Roe, D. R.; Cheatham, T. E. PTRAJ and CPPTRAJ: Software for Processing and Analysis of Molecular Dynamics Trajectory Data. *J. Chem. Theory Comput.* **2013**, *9*, 3084–3095.

(65) Boys, S.; Bernardi, F. The Calculation of Small Molecular Interactions by the Differences of Separate Total Energies. Some Procedures With Reduced Errors. *Mol. Phys.* **1970**, *19*, 553–566.

(66) Simon, S.; Duran, M.; Dannenberg, J. J. How does basis set superposition error change the potential surfaces for hydrogen-bonded dimers? *J. Chem. Phys.* **1996**, *105*, 11024–11031.

(67) Cárdenas, G.; Pérez-Barcia, A.; Mandado, M.; Nogueira, J. J. Characterization of cisplatin/membrane interactions by QM/MM energy decomposition analysis. *Phys. Chem. Chem. Phys.* **2021**, *23*, 20533.

(68) Sun, Q.; Berkelbach, T. C.; Blunt, N. S.; Booth, G. H.; Guo, S.; Li, Z.; Liu, J.; McClain, J. D.; Sayfutyarova, E. R.; Sharma, S.; Wouters, S.; Chan, G. K.-L. PySCF: the Python-based simulations of chemistry framework. *Wiley Interdiscip. Rev. Comput. Mol. Sci.* **2018**, *8*, e1340.

(69) Sun, Q. Libcint: An efficient general integral library for Gaussian basis functions. *J. Comput. Chem.* **2015**, *36*, 1664–1671.

(70) Obara, S.; Saika, A. Efficient recursive computation of molecular integrals over Cartesian Gaussian functions. *J. Chem. Phys.* **1986**, *84*, 3963–3974.

(71) Valeev, E. F. Libint: A library for the evaluation of molecular integrals of many-body operators over Gaussian functions. http://libint.valeyev.net/, 2021; version 2.7.1.

(72) Truhlar, D. G.; Cramer, C. J.; Lewis, A.; Bumpus, J. A. Molecular Modeling of Environmentally Important Processes: Reduction Potentials. *J. Chem. Educ.* **2004**, *81*, 596.

(73) Truhlar, D. G.; Cramer, C. J.; Lewis, A.; Bumpus, J. A. J. Chem. Educ. 2004, 81, 596–604. *J. Chem. Educ.* **2007**, *84*, 934.

(74) Isse, A. A.; Gennaro, A. Absolute Potential of the Standard Hydrogen Electrode and the Problem of Interconversion of Potentials in Different Solvents. *J. Phys. Chem. B* **2010**, *114*, 7894–7899.

(75) Kelly, C. P.; Cramer, C. J.; Truhlar, D. G. Aqueous Solvation Free Energies of Ions and Ion-Water Clusters Based on an Accurate Value for the Absolute Aqueous Solvation Free Energy of the Proton. *J. Phys. Chem. B* **2006**, *110*, 16066–16081.

(76) Marcus, R. A. On the Theory of Oxidation-Reduction Reactions Involving Electron Transfer. I. *J. Chem. Phys.* **1956**, *24*, 966–978.

(77) Marcus, R. A. On the Theory of Oxidation-Reduction Reactions Involving Electron Transfer. III. Applications to Data on the Rates of Organic Redox Reactions. *J. Chem. Phys.* **1957**, *26*, 872–877.

(78) Marcus, R. A. On the Theory of Oxidation-Reduction Reactions Involving Electron Transfer. V. Comparison and Properties of Electrochemical and Chemical Rate Constants. *J. Phys. Chem.* **1963**, *67*, 853–857.

(79) Marcus, R. A. On the theory of electron-transfer reactions. VI. Unified treatment for homogeneous and electrode reactions. *J. Chem. Phys.* **1965**, *43*, 679–701.

(80) Warshel, A. Dynamics of reactions in polar solvents. Semiclassical trajectory studies of electron-transfer and proton-transfer reactions. *J. Phys. Chem.* **1982**, *86*, 2218–2224.

(81) King, G.; Warshel, A. Investigation of the free energy functions for electron transfer reactions. *J. Chem. Phys.* **1990**, *93*, 8682–8692.

(82) Bartmess, J. E. Thermodynamics of the Electron and the Proton. *J. Phys. Chem.* **1994**, *98*, 6420–6424.

(83) Bartmess, J. E. Thermodynamics of the Electron and the Proton [Erratum to document cited in CA121:19495]. *J. Phys. Chem.* **1995**, *99*, 6755–6755.

(84) Marenich, A. V.; Ho, J.; Coote, M. L.; Cramer, C. J.; Truhlar, D. G. Computational Electrochemistry: Prediction of Liquid-Phase Reduction Potentials. *Phys. Chem. Chem. Phys.* **2014**, *16*, 15068–15106.

(85) Blumberger, J.; Tavernelli, I.; Klein, M. L.; Sprik, M. Diabatic free energy curves and coordination fluctuations for the aqueous Ag+/Ag2+ redox couple: A biased Born-Oppenheimer molecular dynamics investigation. *J. Chem. Phys.* **2006**, *124*, 064507.

(86) Perdew, J. P.; Burke, K.; Ernzerhof, M. Generalized Gradient Approximation Made Simple. *Phys. Rev. Lett.* **1996**, *77*, 3865–3868.

(87) Perdew, J. P.; Burke, K.; Ernzerhof, M. Generalized Gradient Approximation Made Simple [Phys. Rev. Lett. 77, 3865 (1996)]. *Phys. Rev. Lett.* **1997**, *78*, 1396–1396.

(88) Tsuneda, T.; Suzumura, T.; Hirao, K. A new one-parameter progressive Colle–Salvetti-type correlation functional. *J. Chem. Phys.* **1999**, *110*, 10664–10678.

(89) Klamt, A.; Schüürmann, G. COSMO: a new approach to dielectric screening in solvents with explicit expressions for the screening energy and its gradient. *J. Chem. Soc., Perkin Trans. 2* **1993**, 799–805.

(90) Seminario, J. M. Calculation of intramolecular force fields from second-derivative tensors. *Int. J. Quantum Chem.* **1996**, *60*, 1271–1277.

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

(92) Martínez, L.; Andrade, R.; Birgin, E. G.; Martínez, J. M. PACKMOL: A package for building initial configurations for molecular dynamics simulations. *J. Comput. Chem.* **2009**, *30*, 2157–2164.

(93) Meza, J. C. Steepest descent. *Wiley Interdiscip. Rev. Comput. Stat.* **2010**, *2*, 719–722.

(94) Galántai, A. The theory of Newton's method. *J. Comput. Appl. Math.* **2000**, *124*, 25–44, Numerical Analysis 2000. Vol. IV: Optimization and Nonlinear Equations.

(95) Schneider, T.; Stoll, E. Molecular-dynamics study of a three-dimensional one-component model for distortive phase transitions. *Phys. Rev. B* **1978**, *17*, 1302.

(96) Berendsen, H. J. C.; Postma, J. P. M.; van Gunsteren, W. F.; DiNicola, A.; Haak, J. R. Molecular dynamics with coupling to an external bath. *J. Chem. Phys.* **1984**, *81*, 3684.

(97) Darden, T.; York, D.; Pedersen, L. Particle Mesh Ewald: An N·log(N) Method for Ewald Sums in Large Systems. *J. Chem. Phys.* **1993**, *98*, 10089–10092.

(98) Miyamoto, S.; Kollman, P. Settle: An analytical version of the SHAKE and RATTLE algorithm for rigid water models. *J. Comput. Chem.* **1992**, *13*, 952–962, cited By 4794.

(99) Seidel, C. A. M.; Schulz, A.; Sauer, M. H. M. Nucleobase-Specific Quenching of Fluorescent Dyes. 1. Nucleobase One-Electron Redox Potentials and Their Correlation with Static and Dynamic Quenching Efficiencies. *J. Phys. Chem.* **1996**, *100*, 5541–5553.

(100) Crespo-Hernández, C. E.; Close, D. M.; Gorb, L.; Leszczynski, J. Determination of Redox Potentials for the Watson-Crick Base Pairs, DNA Nucleosides, and Relevant Nucleoside Analogues. *J. Phys. Chem. B* **2007**, *111*, 5386–5395.

(101) Alsenoy, N. R.-B. I. P.-J. C. V.; Mandado, M. Theoretical study of the adsorption of aromatic units on carbon allotropes including explicit (empirical) DFT dispersion

corrections and implicitly dispersion-corrected functionals: the pyridine case. *Phys. Chem. Chem. Phys.* **2015**, *17*, 575–587.

(102) Mandado, M.; Hermida-Ramón, J. M. Electron Density Based Partitioning Scheme of Interaction Energies. *J. Chem. Theory Comput.* **2011**, *7*, 633–641.

(103) Ropo, M.; Schneider, M.; Baldauf, C.; Blum, V. First-principles data set of 45,892 isolated and cation-coordinated conformers of 20 proteinogenic amino acids. *Sci. Data* **2016**, *3*.

(104) Jorgensen, W.; Chandrasekhar, J.; Madura, J.; Impey, R.; Klein, M. Comparison of simple potential functions for simulating liquid water. *J. Chem. Phys.* **1983**, *79*, 926–935.

(105) Zhao, Y.; Truhlar, D. G. The M06 suite of density functionals for main group thermochemistry, thermochemical kinetics, noncovalent interactions, excited states, and transition elements: two new functionals and systematic testing of four M06-class functionals and 12 other functionals. *Theor. Chem. Acc.* **2008**, *120*, 215–241.

(106) Dunning, T. H. Gaussian basis sets for use in correlated molecular calculations. I. The atoms boron through neon and hydrogen. *J. Chem. Phys.* **1989**, *90*, 1007–1023.

(107) Massey, F. The Kolmogorov-Smirnov Test for Goodness of Fit. *J. Am. Stat. Assoc.* **1951**, *46*, 68–78.

(108) Andersson, K.; Roos, B. O. *Modern Electronic Structure Theory*; pp 55–109.

(109) Andersson, K.; Malmqvist, P. A.; Roos, B. O.; Sadlej, A. J.; Wolinski, K. Second-order perturbation theory with a CASSCF reference function. *J. Phys. Chem.* **1990**, *94*, 5483–5488.

(110) Andersson, K.; Malmqvist, P.; Roos, B. O. Second-order perturbation theory with a

complete active space self-consistent field reference function. *J. Chem. Phys.* **1992**, *96*, 1218–1226.

(111) Palmisano, V. F.; Gómez-Rodellar, C.; Pollak, H.; Cárdenas, G.; Corry, B.; Faraji, S.; Nogueira, J. J. Binding of azobenzene and p-diaminoazobenzene to the human voltage-gated sodium channel Nav1.4. *Phys. Chem. Chem. Phys.* **2021**, *23*, 3552–3564.

(112) Humphrey, W.; Dalke, A.; Schulten, K. VMD: visual molecular dynamics. *J. Mol. Graph.* **1996**, *14*, 33–38.

(113) Pan, X.; Li, Z.; Zhou, Q.; Shen, H.; Wu, K.; Huang, X.; Chen, J.; Zhang, J.; Zhu, X.; Lei, J. Structure of the human voltage-gated sodium channel Nav1. 4 in complex with β1. *Science* **2018**, *362*, eaau2486.

(114) Lomize, M. A.; Pogozheva, I. D.; Joo, H.; Mosberg, H. I.; Lomize, A. L. OPM database and PPM web server: resources for positioning of proteins in membranes. *Nucleic Acids Res.* **2012**, *40*, D370–D376.

(115) Jo, S.; Kim, T.; Iyer, V. G.; Im, W. CHARMM-GUI: a web-based graphical user interface for CHARMM. *J. Comput. Chem.* **2008**, *29*, 1859–1865.

(116) Huang, J.; Rauscher, S.; Nawrocki, G.; Ran, T.; Feig, M.; De Groot, B. L.; Grubmüller, H.; MacKerell, A. D. CHARMM36m: an improved force field for folded and intrinsically disordered proteins. *Nat. Methods* **2017**, *14*, 71–73.

(117) Krylov, A. I.; Gill, P. M. Q-Chem: an engine for innovation. *Wiley Interdiscip. Rev. Comput. Mol. Sci.* **2013**, *3*, 317–326.

(118) McCullagh, M.; Franco, I.; Ratner, M. A.; Schatz, G. C. DNA-based optomechanical molecular motor. *J. Am. Chem. Soc.* **2011**, *133*, 3452–3459.

(119) Vanommeslaeghe, K.; Hatcher, E.; Acharya, C.; Kundu, S.; Zhong, S.; Shim, J.; Darian, E.; Guvench, O.; Lopes, P.; Vorobyov, I., et al. CHARMM general force field: A

force field for drug-like molecules compatible with the CHARMM all-atom additive biological force fields. *J. Comput. Chem.* **2010**, *31*, 671–690.

(120) Aquilante, F.; Lindh, R.; Bondo Pedersen, T. Unbiased auxiliary basis sets for accurate two-electron integral approximations. *J. Chem. Phys.* **2007**, *127*, 114107.

(121) Forsberg, N.; Malmqvist, P.-A. Multiconfiguration perturbation theory with imaginary level shift. *Chem. Phys. Lett.* **1997**, *274*, 196–204.

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

(123) Schrödinger, L.; DeLano, W. PyMOL. `http://www.pymol.org/pymol`.

(124) Sundlov, J. A.; Fontaine, D. M.; Southworth, T. L.; Branchini, B. R.; Gulick, A. M. Crystal structure of firefly luciferase in a second catalytic conformation supports a domain alternation mechanism. *Biochemistry* **2012**, *51*, 6493–6495.

(125) Tian, C.; Kasavajhala, K.; Belfon, K. A. A.; Raguette, L.; Huang, H.; Migues, A. N.; Bickel, J.; Wang, Y.; Pincay, J.; Wu, Q.; Simmerling, C. ff19SB: Amino-Acid-Specific Protein Backbone Parameters Trained against Quantum Mechanics Energy Surfaces in Solution. *J. Chem. Theory Comput.* **2020**, *16*, 528–552.

(126) Becke, A. D. Becke's three parameter hybrid method using the LYP correlation functional. *J. Chem. Phys* **1993**, *98*, 5648–5652.

(127) Stephens, P. J.; Devlin, F. J.; Chabalowski, C. F.; Frisch, M. J. Ab Initio calculation of vibrational absorption and circular dichroism spectra using density functional force fields. *J. Phys. Chem.* **1994**, *98*, 11623–11627.

(128) 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.

(129) Francl, M. M.; Pietro, W. J.; Hehre, W. J.; Binkley, J. S.; Gordon, M. S.; DeFrees, D. J.; Pople, J. A. Self-consistent molecular orbital methods. XXIII. A polarization-type basis set for second-row elements. *J. Chem. Phys.* **1982**, *77*, 3654–3665.

(130) Krishnan, R.; Binkley, J. S.; Seeger, R.; Pople, J. A. Self-consistent molecular orbital methods. XX. A basis set for correlated wave functions. *J. Chem. Phys.* **1980**, *72*, 650–654.

(131) McLean, A. D.; Chandler, G. S. Contracted Gaussian basis sets for molecular calculations. I. Second row atoms, Z=11-18. *J. Chem. Phys.* **1980**, *72*, 5639–5648.

(132) Mofford, D. M.; Reddy, G. R.; Miller, S. C. Aminoluciferins extend firefly luciferase bioluminescence into the near-infrared and can be preferred substrates over D-luciferin. *J. Am. Chem. Soc.* **2014**, *136*, 13277–13282.

(133) Götz, A. W.; Williamson, M. J.; Xu, D.; Poole, D.; Le Grand, S.; Walker, R. C. Routine Microsecond Molecular Dynamics Simulations with AMBER on GPUs. 1. Generalized Born. *J. Chem. Theory Comput.* **2012**, *8*, 1542–1555.

(134) Salomon-Ferrer, R.; Götz, A. W.; Poole, D.; Le Grand, S.; Walker, R. C. Routine Microsecond Molecular Dynamics Simulations with AMBER on GPUs. 2. Explicit Solvent Particle Mesh Ewald. *J. Chem. Theory Comput.* **2013**, *9*, 3878–3888.