Ab initio study of the optical properties of green fluorescent protein

Zaccheddu, M.

Ab initio study of the optical properties of Green Fluorescent Protein

Maurizio Zaccheddu

In the front cover:
Green Fluorescent Protein with the chromophore shown inside the protein cavity.

Ab initio study of the optical properties of Green Fluorescent Protein

PROEFSCHRIFT

ter verkrijging van
de graad van Doctor aan de Universiteit Leiden,
op gezag van Rector Magnificus Prof. Mr. P.F. van der Heijden,
volgens besluit van het College voor Promoties
te verdedigen op donderdag 24 April 2008
klokke 16.15 uur

DOOR

Maurizio Zaccheddu

geboren te Cagliari (Italië) in 1978

Promotor:	Prof. dr. C. Filippi
Referent:	Dr. F. Buda
Overige leden:	Prof. dr. J.M. van Ruitenbeek
	Prof. dr. J.M.J. van Leeuwen
	Prof. dr. P. Bolhuis
	Prof. dr. E.J. Baerends

To my parents

Contents

Chapter 1

Introduction

1.1 Photosensing in biological systems

The absorption of visible light and its conversion to other forms of energy is at the core of some of the most fundamental processes in biology. Indeed, life on earth owes its existence to photosynthesis, a process through which sunlight is harvested and converted into chemical energy by plants, algae and photosynthetic bacteria. Another familiar example of light absorption initiating a biological response over several temporal and length scales is vision: light stimulates a conformational change of the photosensitive component in the retina, which is followed by a cascade of chemical reactions ultimately culminating in the stimulation of the optical nerve. In general, photosensing in a biological system occurs through a photoreceptor protein that hosts a chromophore (i.e. the molecule bound to the protein proper and responsible for light absorption and emission) which undergoes a photochemical reaction, such as photoisomerization, excitation transfer, electron or proton transfer upon photoexcitation. Deepening our physical understanding of the primary excitation processes and of the subsequent energy transfer in these photobiological systems is important both from a fundamental point of view and because of existing and potential applications in biology, biotechnology and artificial photosynthetic devices.

An important example of photosensitive biosystems is the family of autofluorescent proteins, a class of biological labels that has revolutionized cellular biology in the last decades. These molecules absorb light at one wavelength and emit (i.e. fluoresce) at a specific and longer wavelength. Since they can often be coexpressed with non-fluorescent proteins without affecting the latter's functions, autofluorescent proteins have been used in a multitude of applications, for example as fluorescent labels to visualize and track

proteins in living cells, to monitor protein-protein interactions, and as indicators of pH and calcium concentration in vivo. For certain applications, it is however desirable to modify, enhance or suppress the molecular mechanisms that modulate the response of the chromophore to external inputs. Understanding the relationship between the microscopic structure and the spectral properties of these biosystems permits the rational design of new photoactive systems with novel functions through selective mutations of existing autofluorescent molecules. Examples include shifting their excitation and emission spectra, or altering the sensitivity to external factors such as pH or past exposure to light.

Theoretical calculations of the optical properties of photoactive systems complement experimental spectroscopic data by providing an atomistic description of the dynamical response of the protein upon light activation. In order to attack these challenging problems, the computational approach must however meet several difficult requirements. First, it should provide an accurate quantum-mechanical description of the ground state and of the electronic excitations of the photoactive site. It should then include a dynamical description of ground state fluctuations and possibly of photo-induced dynamical effects. Finally, the calculations must be able to treat a realistically large model of the biosystem in order to understand how modifications of the protein environment affect the optical properties of the chromophore. It is far from trivial to satisfy all the above requirements. In most cases, ground state properties of large systems can be reliably and efficiently computed from first principles, in particular through density functional theory (DFT) approaches, and sufficient knowledge has also been accumulated to establish the reliability of a given calculation. However, the computation of excitation energies is proving to be more complicated, and there are serious problems with the approaches employed in the study of large photoactive biomolecules. In surveying the vast theoretical literature on photosensitive systems, one finds that the large spread of semi-empirical and first-principle approaches used for a particular system yields an equally large spread of results and predictions.

The most appealing approach for the computation of excitations in large molecular systems is certainly time-dependent density functional theory (TDDFT) given its favorable scaling with system size. While generally reasonably accurate, conventional adiabatic TDDFT often fails to describe charge transfer excitations in extended conjugated systems and excitations characterized by two- and higher-electron excitations. As we will show in this thesis, these and other shortcomings may result in the poor description of the excitations of photoactive chromophores which usually are conjugate π-systems with electronic states often displaying multi-configurational character. The unre-

liability of density-functional-based approaches to accurately describe photoexcitations of biomolecules implies that the main researchers aggressively working in this field are employing conventional highly-correlated quantum chemical approaches as multi-reference configuration interaction (MRCI) and complete active space second-order perturbation theory (CASPT2). These approaches rely on expanding the wave function in Slater determinants and, as the system size increases and the energies of the single-particle orbitals become closely spaced, the space of orbitals which must be included to recover a significant fraction of electronic correlation grows enormously. Therefore, when these approaches are applied to large biomolecules, compromises must be taken as in the use of a small atomic basis or a reduced space of active orbitals. Consequently, while highly-correlated quantum chemical approaches are accurate for small systems where these techniques can be pushed to their limits, the same level of accuracy cannot in general be guaranteed when going to a large biosystem.

In this thesis, we employ a hierarchy of state-of-the-art computational methods to deal with the problem at different levels of accuracy. We believe that, for the description of the ground state properties of these photoactive biomolecules, conventional techniques are sufficiently accurate while, for excited states, we want to explore the performance of a different approach as new theoretical ways to handle excited states are needed. More specifically, ground state properties can be described using density functional theory in combination with ab-initio molecular dynamics to equilibrate the structures and study the thermal fluctuations of the chromophore and its immediate surroundings. The long-range protein-chromophore interactions can be included via hybrid quantum-classical simulation schemes, where the photoactive site is described quantum mechanically and the interaction with the rest of the macromolecule is treated using an atomic force field.

For the computation of excited states, on the other hand, we will use a different theoretical framework based on many-body quantum Monte Carlo techniques that has been developed over the last few years by Filippi and coworkers and has already yielded accurate excitations of a variety of small photoactive molecules. Moreover, to describe the long-range protein-chromophore interactions, we will combine for the first time quantum Monte Carlo with a molecular mechanics approach where the chromophore is treated quantum mechanically and the rest of the protein classically. The advantage of quantum Monte Carlo methods compared to highly correlated quantum chemical approaches is that they scale far more favorably with system size. While we already know that quantum Monte Carlo is competitive with highly-correlated quantum chemical approaches for small molecules, this study represents the first application of quantum Monte Carlo techniques to

the description of the excitations of a realistic complex biomolecular system.

Using this hierarchical combination of computational approaches, we study here the rich photophysical behavior of Green Fluorescent Protein (GFP), the prototype of the class of autofluorescent proteins and one of the most widely used fluorescent labels in cellular biology. In particular, we investigate the interplay between the spectral properties and the microscopic structural features of the chromophore-protein complex in its different forms. Beyond being extremely relevant in biotechnology, GFP represents a perfect playground for our theoretical investigation of photoactive biomolecules due to several reasons. First, this protein is experimentally very well characterized, serving as a stringent test for any approach aiming at describing excitation processes in biosystems. Then, GFP has already been the subject of a large number of theoretical semi-empirical and first-principle studies, none of which fully conclusive. Finally, despite the substantial body of literature, several issues which we will not touch in this thesis are still open and not convincingly addressed by theoretical calculations. These include the conformational changes in the chromophore and their relation to the so-called dark states which are reversibly accessible after photoexcitation during blinking and switching. For all these reasons, GFP is the ideal arena where to validate and possibly sharpen our proposed methodology while addressing the theoretical challenge to understand the nature of the excitations in this relevant autofluorescent protein.

1.2 The Green Fluorescent Protein

Green fluorescent protein (GFP) is the prototype of the class of autofluorescent proteins [1–3]. GFP is an intrinsically fluorescent protein and was first extracted [4] from the bioluminescent jellyfish *Aequorea victoria* of the Pacific Ocean, shown in Fig. 1.1. The Aequorea jellyfish bioluminesces (i.e. emits light as a result of a chemical reaction) at the rim of its bell and two proteins are involved in its bioluminescence, aequorin and the Green Fluorescent Protein. By a quick release of calcium ions, the jellyfish can induce the photoprotein aequorin to emit blue light, which is then transduced to green via radiationless energy transfer to a coupled Green Fluorescent Protein. The biological function of GFP in the jellyfish Aequorea is therefore to convert the blue emission of aequorin to green emission. Interestingly, it still remains unclear how and why these organisms use their bioluminescent capabilities as yellyfish do not flash at each other in the dark, nor glow continuously [5]. Moreover, it is not understood why these jellyfishes would synthesize a separate protein rather then mutate the chemiluminescent protein to shift its

wavelength, and why green emission should be ecologically superior to blue.

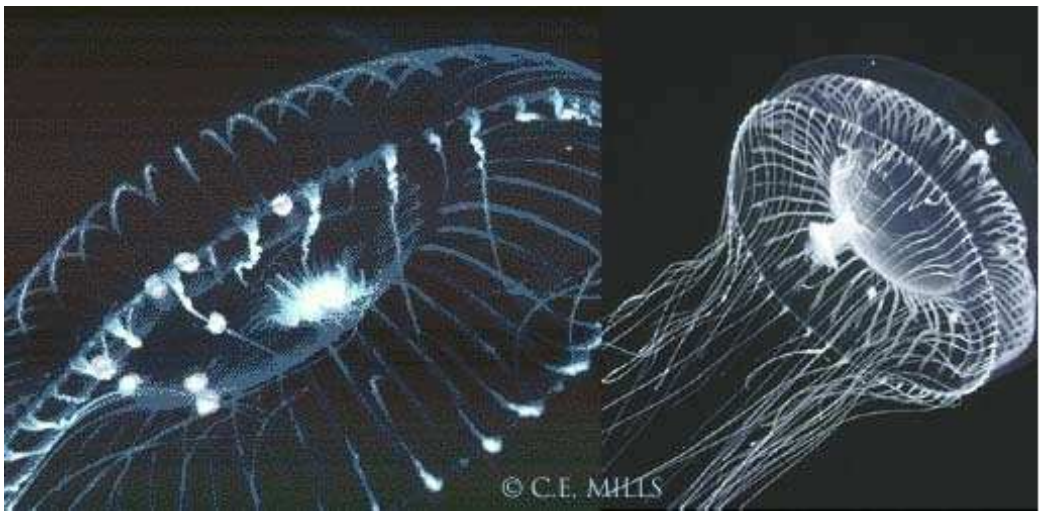

Figure 1.1: Two views of the hydromedusa Aequorea victoria from Friday Harbor, Washington [5].

Independently of the reason, evolution and natural selection has generated a very efficient optical devise and this optimization through evolution is probably a reason for the success of GFP in biotechnology. Over the last decades, GFP has in fact become one of the most widely used markers in cellular biology. The most successful and numerous applications of GFP are as a genetic fusion partner to host proteins, which maintain their normal functions but are now fluorescent and can be dynamically visualized in living cells and organisms. This property is dramatically illustrated in Fig. 1.2, where the genetic code of a mouse has been modified to express Green Fluorescent Protein. Moreover, significant experimental efforts have gone in engineering mutants of the original Aequorea victoria GFP with different colors, enhanced fluorescence and photostability or specific sensitivity to external factors such as temperature or pH. These mutagenesis studies have resulted in new fluorescent probes that range in color from blue to yellow. The search for mutants with longer-wavelength emission has been motivated by the difficulty to distinguish GFP emission from the background cellular fluorescence, as well as the desire to develop fluorescent resonance energy transfer (FRET) partners with the required overlap between absorption and emission spectra, to tag different proteins and study protein-protein interactions in vivo.

Because the construction of red-shifted mutants from the Aequorea victoria jellyfish GFP beyond the yellow spectral region has proven largely unsuccessful, longer-wavelength fluorescent proteins emitting in the orange and red

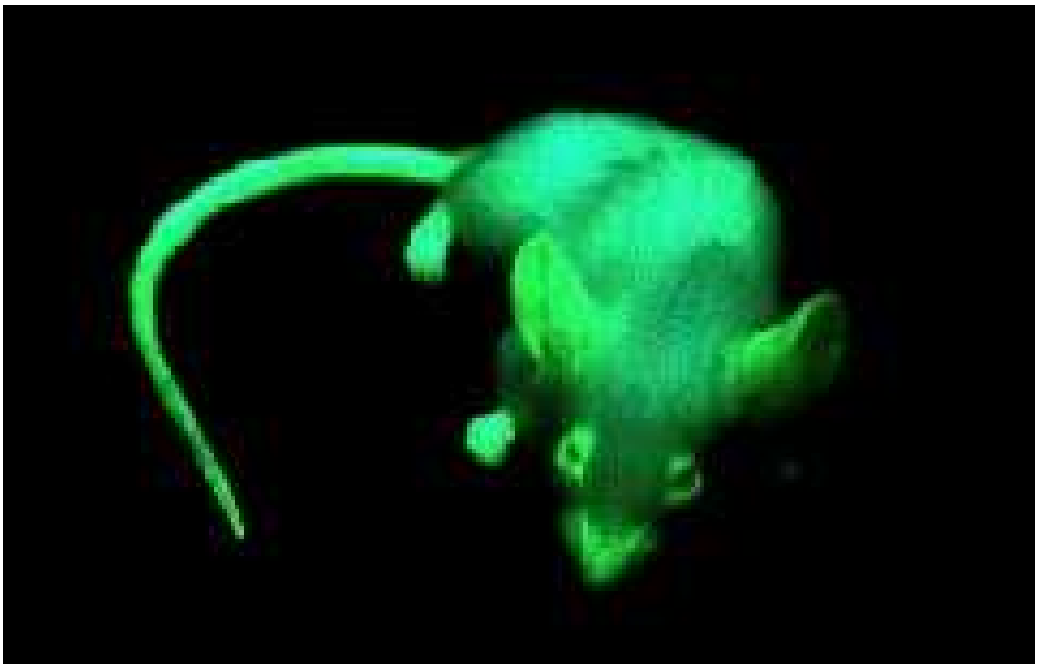

Figure 1.2: Mouse expressing Green Fluorescent Protein, illuminated under blue light [6].

spectral regions, have been extracted from other sea organisms as the marine anemone, *Discosoma striata*, and reef corals belonging to the class *Anthozoa*. Still other species have been mined to produce similar proteins having cyan, green, yellow, orange, and deep red fluorescence emission. Consequently, a broad range of fluorescent protein genetic variants is now available that feature fluorescence emission spanning almost the entire visible light spectrum. In the following, we will restrict our discussion to wild-type GFP, that is the original protein of Aequorea victoria.

The tertiary structure of wild-type Green Fluorescent Protein is shown in Figure 1.3. The fold comprises 11 β-sheets arranged in a barrel-like structure with a diameter of about 24 Å and a height of 42 Å. This structure forms the so-called β-can which is capped by short helical segments. The chromophore is well protected in the center of the barrel and is linked to the α-helical stretch which runs close to the central part of the barrel. This fold motif with minor variations is common to all proteins of the GFP family, including the fluorescent proteins extracted from other sea organisms. The correct folding of GFP in the β-can structure and the configuration of the residues around the chromophore are crucial to the formation and the fluorescence of the chromophore which is rigidly kept inside this chemically protective structure,

displaying high stability and quantum yield of fluorescence. In fact, the isolated chromophore is not fluorescent in aqueous solution, and denaturation yields a loss of fluorescence which is regained when the β-can structure is correctly reformed. The isolated chromophore is also shown in Fig. 1.3 and is a p-hydroxybenzylideneimidazolinone molecule formed autocatalytically by an intramolecular post-translational cyclization of three consequtive amino acids (Ser-65, Tyr-66, and Gly-67).

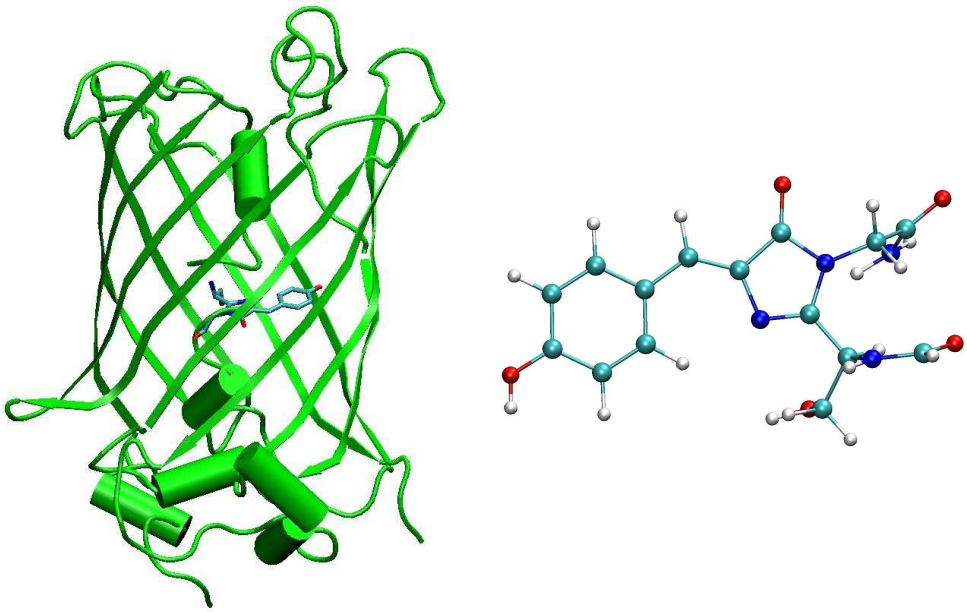

Figure 1.3: Tertiary structure of Green Fluorescent Protein represented in strand style (left). The β-can structure has a diameter of about 24 Å and a height of 42 Å. The chromophore is highlightened in the center of the protein cavity and is also shown isolated in vacuum (right).

The fluorescent mechanisms of wild-type GFP is prototypical of the GFP family. At thermal equilibrium, the absorption spectrum of wild-type GFP has two peaks at 398 nm (3.12 eV) and 478 nm (2.59 eV). Excitation at 398 nm results in an emission maximum in the region of 506 nm (2.45 eV) while irradiation at 478 nm yields emission with a maximum at 482 nm (2.57 eV) [7]. The absorption spectrum of wild-type GFP at room temperature is shown in Fig. 1.4. The two absorption bands at 398 and 478 nm were attributed early on to two interconvertible states of the protein with the chromophore in a neutral (protonated) A form and an anionic (deprotonated) B form, respectively. Upon photoexcitation of the neutral A form, the excited chromophore

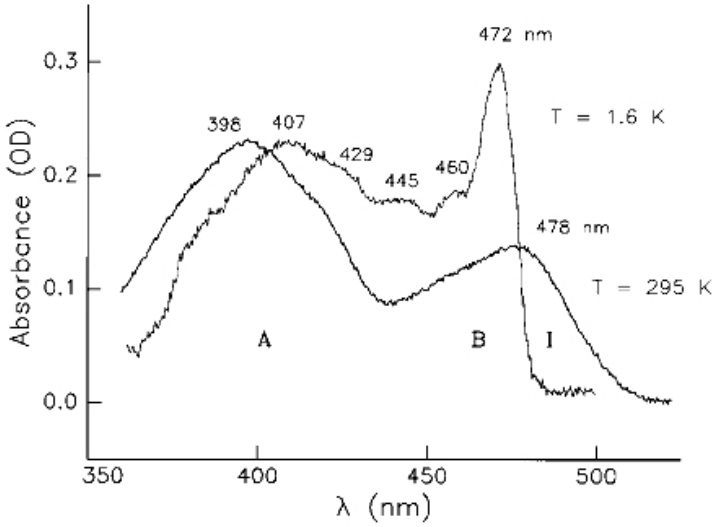

Figure 1.4: Absorption spectra of wild-type GFP at room temperature (T=295 K) and low temperature (T=1.6 K) from Ref. [7].

transfers a proton through a complex hydrogen-bond network to the residue Glu-222 forming a transient intermediate anionic state (I^*) which emits in the region of 506 nm (2.45 eV). After decay to the ground state (I), the system usually returns to state A through a ground state inverse proton transfer process. The green fluorescence at 482 nm (2.57 eV) following excitation of the B state stems from direct decay of the excited B^* state. Therefore, both the I and the B states are characterized by an anionic (deprotonated) chromophore but the I form has a protein environment similar to the neutral A form while the environment of the B form is structurally different from the A and I forms with the Thr-203 residue being rotated and forming a hydrogen bond with the phenolic oxygen. The fluorescence mechanisms of wild-type GFP is summarized in Fig. 1.5 where a schematic representation of the neutral and anionic chromophores and the corresponding protein binding sites is also shown. This model for the photocycle of wild-type GFP was originally proposed after ultrafast excited-state dynamics measurements and rationalized on the basis of the resolved x-ray structures of the neutral A form and of the B form as stabilized in GFP mutants. We will return to a detailed analysis of the three forms of wild-type GFP and their protein environments in Chapter 4.

Finally, we report here few additional experimental observations which are relevant for our theoretical calculations. In particular, the absorption

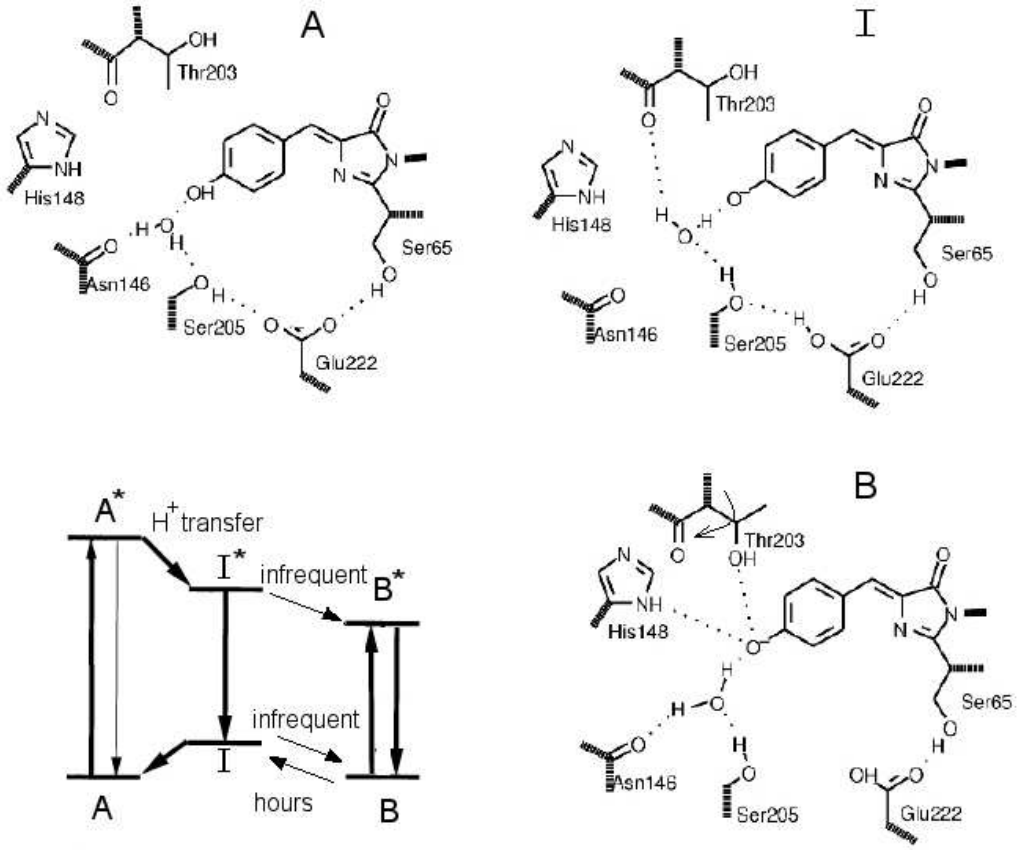

Figure 1.5: Scheme of the fluorescence mechanisms of wild-type Green Fluorescent Protein. The hydrogen bond network from the chromophore through the residues involved in the proton transfer is shown for the neutral A and the anionic I and B forms. Note the change in conformation of residue Thr-203 in going from the I to the B form. The figure is adapted from Ref. [3].

spectrum of wild-type GFP at 1.6 K is also shown in Fig. 1.4. At low temperature, the two maxima shift at 407 nm (3.05 eV) and 472 nm (2.63 eV), and the ratio of the absorbances of the A and B forms inverts with respect to room temperature indicating that the B form has a slightly lower ground state than the A form. The broad wing at the red side of the 472 maximum disappears and is attributed to the I form which is not populated at this low temperature. Finally, spectral hole-burning experiments have located the 0-0 transitions of the three forms and shown that the ground state of the I form is higher than the ground states of the A and B forms, and separated from them by energy barriers of several hundred wavenumbers. Moreover, the excited-state barrier between A^* and I^* is low while the barrier between

I^* and B^* is about 2000 cm^{-1} (0.25 eV), so the only possible interconversion is between the excited states of the A and I forms [7].

1.3 Previous theoretical work

The structural and optical properties of wild-type Green Fluorescent Protein have already been the subject of several theoretical investigations. We will not review the early semi-empirical and quantum chemical studies [8–10] since they were not able to unambiguously assign the charge states to the experimental absorption bands. Initially, a cationic and a zwitterionic form of the protein were even proposed as the protonated and the deprotonated state of the chromophore. Moreover, some early calculations yielded excitation energies for a particular charge state of the chromophore varying by more than 1 eV when slightly different quantum chemical approaches were employed [11]. We will focus instead on the most recent first-principle calculations of the excitations of wild-type GFP.

Particularly relevant is a recent first-principle study of the neutral and anionic forms of GFP by Marques *et al.* [12] who report a remarkably good agreement of the time-dependent density functional theory (TDDFT) spectra in the local density approximation (LDA) with experiments. These theoretical results are summarized in Fig. 1.6 where the TDDFT/LDA absorption peaks of 3.01 eV and 2.67 eV are compared with the experimental low-temperature maxima of 3.05 and 2.63 eV for the A and the B form, respectively. Few anomalous features characterize however these calculations and raise doubts about the definite and conclusive nature of this study. While the chromophore-protein structures are optimized in the presence of the protein environment using a DFT/LDA quantum mechanics in molecular mechanics (QM/MM) approach, the TDDFT excitation energies are then computed on the isolated chromophores without the sourrounding protein environment. Therefore, possible polarization effects of the protein are not included in the calculation of the excitations of the chromophore. Moreover, the authors model the anionic I form and not the B form by deprotonation of the neutral A form, but erroneously state to be simulating the B form.

Highly-correlated quantum chemical calculations have been recently published for the I and B forms by Sinicropi *et al.* [13] using complete active second-order perturbation theory (CASPT2) for a large model chromophore of GFP in the presence of a classical protein environment. With respect to the original x-ray structure of the neutral A form, only the coordinates of the chromophore and three water molecules are relaxed within the complete active space self consistent field (CASSCF) approach. For the construction

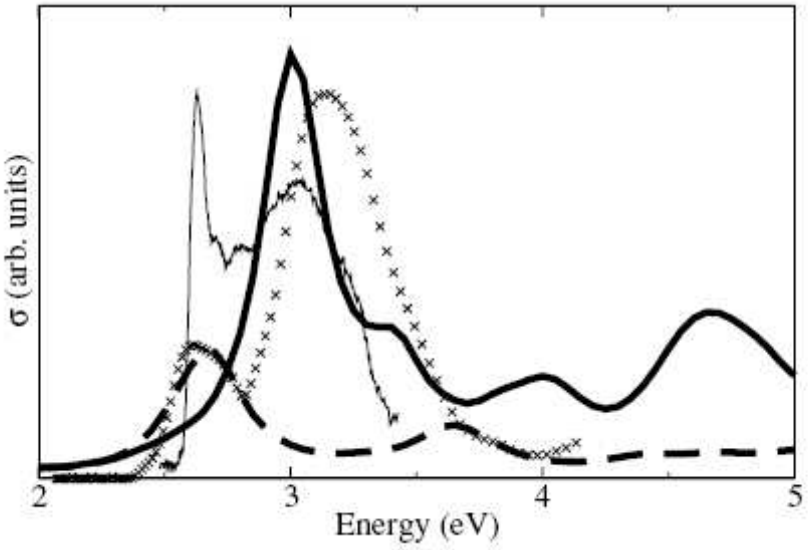

Figure 1.6: TDDFT/LDA spectra of the neutral (think solid line) and anionic (thick solid line) chromophores of wild-type GFP as computed by Marques *et al.* [12]. The experimental low-temperature (thin line) and room-temperature (crosses) spectra are also shown. The TDDFT calculations are performed for the isolated chromophores whose structures were optimized in ground state DFT/LDA QM/MM calculations. We note that the spectrum for the computed anionic I form is erroneously attributed to the B form. The figure is adapted from Ref. [12].

of the I form, the neutral chromophore is deprotonated and and some relevant residues are manually reoriented while, for the B form, residue Thr-203 is partially relaxed in its proper conformation. The CASSCF QM/MM embedding scheme is therefore very simple and lacks a complete relaxation of the chromophore-protein structure. Nevertheless, the CASPT2 absorption maximum of 2.81 eV for the B form appears to be reasonably close to the experimental value of 2.63 eV, while a better agreement with experiments is obtained for the emission maxima of both the I and B forms. Unfortunately, the authors do not report the excitation for the neutral A form of the protein, so it is not possible to access whether this approach is actually capable to correctly describe how the spectrum shifts with the protonation state of the chromophore.

1.4 This thesis

The main focus of this thesis is the computational study of the absorption properties of wild-type Green Fluorescent Protein in the neutral A and anionic I and B forms. We not only construct a series of model chromophores in the gas phase but also investigate how the spectral properties of the chromophore are modified by the protein environment using hybrid molecular mechanics in quantum mechanics approaches to account for the long-range chromophore-protein interactions. To compute the excitations of GFP, we employ both conventional time-dependent density functional theory as well as quantum Monte Carlo techniques. Since this thesis is the first application of mixed classical/quantum Monte Carlo methods to the computation of the excited states of a large biomolecule, it serves the dual purpose of both understanding the spectral tuning of the excitations of GFP by the protein proper as well as assessing the performance of quantum Monte Carlo to describe the excited states of a complex biosystem. This thesis is organized as follows.

In Chapter 2, we describe the computational methods we use in the thesis. We review highly-correlated quantum chemical approaches as well as density functional theory also in its time-dependent formulation. We discuss in depth quantum Monte Carlo methods, in particular the functional form of the trial wave function and the optimization scheme used to obtain the optimal parameters in the excited-state wave functions. We briefly describe molecular mechanics techniques and the hybrid quantum mechanics in molecular mechanics (QM/MM) scheme used for the study of Green Fluorescent Protein. The computational details conclude this Chapter.

In Chapter 3, we construct a set of models of the neutral and anionic chromophores of GFP in the gas phase to begin exploring the performance of adiabatic time-dependent density functional theory and quantum Monte Carlo approaches. The results are puzzling. TDDFT appears to be overestimating the excitations of a small anionic model chromophore as compared to photodistruction spectroscopy experiments and highly-correlated CASPT2 calculations while the experimental absorption maximum obtained with the same technique for a cationic model is reasonably well reproduced. If signatures of possible problems in the use of TDDFT can be found for the larger models that we have constructed, we are not able to rationalize the reasons for its apparent failure in the description of the smaller anionic model chromophore. Moreover, using quantum Monte Carlo techniques and sophisticated wave functions, we obtain excitations for the small anionic model in reasonable agreement with TDDFT. A significant difference with TDDFT is instead that QMC yields a large shift in the excitation when going from the

neutral to the anionic model of the GFP chromophore in the gas phase.

In Chapter 4, we construct the protein models of the neutral A and the two anionic I and B forms of wild-type GFP using a density functional theory QM/MM approach. The outcome of this ground state modeling is already surprising and shows how difficult it is to correctly describe a complex biosystem and how easy to be mislead in believing the correctness of a given model when comparing to relatively few experimental numbers. We carefully analyze the structures of our protein as well as of other models available in the literature and conclude that the DFT QM/MM calculations by Marques *et al.* [12] are incorrect due to what we believe is a wrong description of the binding site of the chromophore. Naturally, the incorrect description of the residues surrounding the chromophore affects its response to light and the perfect agreement of the TDDFT spectra for the corresponding isolated chromophore with experiments shown in Fig. 1.6 is in fact purely coincidental and due to the use of incorrect chromophore structures. Our TDDFT/MM calculations of our chromophore models in the presence of a classical protein environment yield an absorption maximum in agreement with experiments for the neutral A but not for the anionic I and B forms of GFP. The red-shift in excitation due to deprotonation of the chromophore is very badly underestimated by adiabatic TDDFT which sees almost no difference between the neutral and anionic excitations. We then explore for the first time the use of QMC in describing the excitations of a chromophore in its protein environment and perform QMC/MM calculations of the excitation energies of the three forms of wild-type GFP, using for the moment only a simple wave function. We find that the experimental shift between the different charge states of the chromophore-protein complex is well reproduced by QMC but the absolute excitation energies are overestimated as compared to experiments. We show some first steps to investigate the possible reasons for this error such as shortcomings in the QM/MM description of the chromophore-protein interaction, which, we believe, will resolve the issue in combination with the use of more sophisticated wave functions.

Chapter 5 is self-standing and outside the main thread of the thesis, and focuses on the cooperative effects of π-π and π-anion interactions, a relevant theme within supramolecular chemistry for the design of receptors of anionic species. In particular, we investigate the geometrical and energetic effects induced by π-π stacking on the anion-π system of the unusual triazine-triazine-nitrate complex recently observed experimentally, using semi-empirical dispersion corrected density functional theory and QMC methods. We reproduce and rationalize the highly asymmetrical features of the experimental structure, which are not imposed by the coordination of the anion-π-π subunit within the particular synthesized compound. We quantify the energetic

stabilization induced by π-π stacking and discuss ways to further enhance this cooperative effect in the design of anion-host architectures.

Chapter 2

Computational methods

2.1 Introduction

To investigate the photophysics of Green Fluorescent Protein, we will employ a variety of computational methods as there is not a single theoretical approach to date, which is capable to cover the different spatial and temporal scales which characterize this complex problem. Starting from the x-ray structure of the protein, we will build a realistic model of the protein environment surrounding the chromophore, that is, the optically active component of the protein. As the protein exhibits multiple forms corresponding to different protonations of the chromophore, the system must be properly relaxed to describe the different conformations. The electronic properties of the chromophore within its protein environment are then investigated quantum mechanically in the ground and excited states. These steps translate in a series of computations involving a hierarchy of theoretical approaches, ranging from classical molecular dynamics to correlated many-body techniques. In particular, we will present the results of the following calculations:

- Classical molecular mechanics (MM) calculations to equilibrate the protein in water solution at room temperature.
- Hybrid quantum mechanics in molecular mechanics (QM/MM) calculations based on density functional theory (DFT) to obtain an accurate description of the ground state geometry of the optically active chromophore.
- Time-dependent density functional theory (TDDFT) calculations to compute the excitation spectrum and access how the protein environment modulates the response of the chromophore to light.
- Correlated post-Hartree-Fock quantum chemical approaches to investi-

gate the role of correlation and possible shortcomings in the description of excited states within density functional theory. These calculations are also a prerequisite for the construction of the many-body wave function needed in the next step.

- Quantum Monte Carlo (QMC) calculations of the excitation spectrum. As the application of quantum Monte Carlo to excited states is rather new, further methodological developments were needed.

In this chapter, we give a short description of the theoretical methods we employ and of the relevant technical details. We begin with the quantum mechanical methods, in particular the multi-configuration self-consistent (MCSCF) approach and its state average (SA) version for the computation of excited states, the variational (VMC) and diffusion (DMC) Monte Carlo methods, and time-dependent density functional theory (TDDFT). We then describe how a quantum mechanical approach can be combined with classical molecular mechanics (QM/MM) for the hybrid treatment of a quantum site embedded in a larger classical system. Novel methodological developments will be presented in the following chapters to more clearly illustrate the context in which they are needed.

2.2 Quantum mechanical calculations

The many-electron Schrödinger equation gives an accurate description of materials at the quantum mechanical level but is an intractable $3N + 1$ dimensional partial differential equation, where the number of electrons N may be very large. In this thesis, we will consider molecular systems with typically 100-500 electrons for which we want to investigate the electronic properties of the ground and lowest excited states.

Most computational quantum mechanical studies of such large electronic systems circumvent the problem of the high dimensionality by employing simpler one-electron theories such as Kohn-Sham density-functional theory (DFT), which replaces the electron-electron interactions by an effective potential, thereby reducing the problem to a set of one-electron equations. Despite the successes of DFT in describing the electronic structure of complex molecular systems, the treatment of electronic correlation within DFT is only approximate, sometimes leading to incorrect results as we will see in the case of the excitation spectrum of the Green Fluorescent Protein. Therefore, one needs to resort to alternative approaches as the more costly wave function based methods. Here, we will not employ traditional quantum chemistry wave function methods as for instance the complete active space second or-

der perturbation theory (CASPT2) technique which is often used to treat excitations of organic molecules, but we will focus on quantum Monte Carlo (QMC) techniques, which, for ground state problems, have yielded in the past very accurate description of correlated properties also of large systems, where conventional quantum chemistry methods are extremely difficult to apply. We review here only those aspects of traditional quantum chemistry approaches which are need to understand how the wave function is constructed for quantum Monte Carlo calculations.

Let us define the notation we adopt in this thesis. As we neglect relativistic effects and work in the Born-Oppenheimer approximation, we will assume that we have a non-relativistic system of N interacting electrons described by the Hamiltonian:

$$\mathcal{H} = -\frac{1}{2}\sum_{i=1}^{N}\nabla_i^2 + \sum_{i=1}^{N} v_{\text{ext}}(\mathbf{r_i}) + \sum_{i<j}^{N}\frac{1}{|\mathbf{r}_i - \mathbf{r}_j|}\,, \tag{2.1}$$

where we used atomic units ($\hbar = m = e = 1$). The external potential $v_{\text{ext}}(\mathbf{r})$ is given either by the bare electron-ion Coulomb potential $-Z/r$ where Z is the charge of the ion, or by a pseudopotential describing the ion plus the core electrons which have been eliminated from the calculation. We denote with $\mathbf{R}$ the $3N$ particle coordinates, and with $\mathbf{x} = (\mathbf{r}, \sigma)$ the 3 spatial and 1 spin coordinates of one electron where $\sigma = \pm 1$.

2.2.1 Traditional quantum chemistry methods

The simplest approach for the description of a system of N interacting electrons is the the Hartree-Fock (HF) method, where the ground state many-body wave function is approximated as the optimal non-interacting solution, that is a Slater determinant of single-particle spin-orbitals $\{\Phi_i\}$:

$$\Psi_{\text{HF}}(\mathbf{x}_1, \ldots, \mathbf{x}_N) = \frac{1}{\sqrt{N!}}\begin{vmatrix} \Phi_1(\mathbf{x}_1) & \Phi_1(\mathbf{x}_2) & \cdots & \Phi_1(\mathbf{x}_N) \\ \Phi_2(\mathbf{x}_1) & \Phi_2(\mathbf{x}_2) & \cdots & \Phi_2(\mathbf{x}_N) \\ \vdots & \vdots & \vdots & \vdots \\ \Phi_N(\mathbf{x}_1) & \Phi_N(\mathbf{x}_2) & \cdots & \Phi_2(\mathbf{x}_N) \end{vmatrix}.$$

The optimal single-particle orbitals are determined by minimizing the expectation value E_{HF} of the interacting Hamiltonian $\mathcal{H}$ on the wave function Ψ_{HF}. If the spin-orbitals are written as the product of a spatial and a spin components, $\Phi_i(\mathbf{x}) = \phi_i(\mathbf{r})\chi_{s_i}(\sigma)$, one obtains that the spatial orbitals must

satisfy the self-consistent HF equations:

$$\left[-\frac{1}{2}\nabla^2 + v_{\text{ext}}(\mathbf{r}) + \sum_{j=1}^{N}\int \mathrm{d}\mathbf{r}'\,\frac{|\phi_j(\mathbf{r}')|^2}{|\mathbf{r}-\mathbf{r}'|}\right]\phi_i(\mathbf{r}) - \sum_{j=1}^{N}\delta_{s_i,s_j}\int \mathrm{d}\mathbf{r}'\,\frac{\phi_j^*(\mathbf{r}')\phi_i(\mathbf{r}')}{|\mathbf{r}-\mathbf{r}'|}\phi_j(\mathbf{r}) = \epsilon_i\phi_i(\mathbf{r})\,, \tag{2.2}$$

where the Lagrange multipliers ϵ_i arise from the orthonormality constraints between the orbitals. Each orbital sees the external potential, the Hartree electrostatic component, and the non-local Hartree-Fock exchange potential. The HF potential cancels the interaction of the electron with itself, that is the self-interaction contribution coming from the the Hartree potential, and keeps the electrons of the same spin apart so that each electron has a hole around it, known as the exchange hole, containing unit positive charge.

For atoms, the HF equations can be solved directly on a grid but, for molecular systems, the orbitals are expanded as a linear combination of atomic orbitals (LCAO) centered on the nuclear positions:

$$\phi_i(\mathbf{r}) = \sum_{\mu}^{\text{nuclei}}\sum_{j} a_{ji}^{\mu}\,\eta_{j\mu}(\mathbf{r}-\mathbf{r}_\mu)\,, \tag{2.3}$$

where $\mathbf{r}_\mu$ denotes the position of a nucleus. The LCAO coefficients, a_{ji}^{μ}, are optimized to yield the lowest variational energy. In general, most quantum chemistry codes work with a Gaussian atomic basis:

$$\eta(\mathbf{r}) = x^m y^n z^k \exp\left(-\alpha r^2\right), \tag{2.4}$$

as this choice allows all integrals to be computed analytically.

The difference between the exact energy E and the HF energy is called the correlation energy, $E_{\text{corr}} = E - E_{\text{HF}}$.

Post Hartree-Fock methods

Quantum chemical post-HF approaches express the many-body wave function $\Psi(\mathbf{x}_1, \ldots, \mathbf{x}_N)$ in terms of a non-interacting basis as they rely implicitly or explicitly in writing the wave function as an expansion in determinants of single-particle orbitals. With such an expansion, the matrix elements of the Hamiltonian on the basis and the overlap of the basis functions can be readily expressed and even computed analytically if a Gaussian basis set is employed to express the single-particle orbitals.

Conceptually, we can imagine to start from the solutions of the HF equations which give us a complete set of orthonormal orbitals, comprising the N occupied orbitals and $M - N$ virtual orbitals, where M is the size of the atomic basis set. We can then proceed as in the configuration interaction (CI) approach and construct a correlated wave function as

$$\Psi_{\mathrm{CI}} = c_0 D_{\mathrm{HF}} + \sum_{ab} c_{a\to b} D^{a\to b} + \sum_{abcd} c_{ab\to cd} D^{ab\to cd} + \dots \,, \tag{2.5}$$

where $D^{a\to b}$ denotes a single excitation from the HF determinant where the occupied orbital a is substituted with the virtual orbital b. Similarly, $D^{ab\to cd}$ indicates a double excitation with the orbitals a and b substituted with the virtual orbitals c and d. A full CI expansion is obtained if one includes up to N-body excitations to all virtual orbitals, and the result should then be extrapolated to the infinite basis limit by considering larger basis sets. We can rewrite a CI expansion in more compact form as

$$\Psi_{\mathrm{CI}} = \sum_{i=1}^{K} c_i C_i \,, \tag{2.6}$$

where C_i are spin and space-adapted configuration state functions (CSF), that is, fixed linear combination of determinants with proper spin and space symmetry. By applying the variational principle, one obtains the secular equations for the coefficients c_i:

$$\sum_{j=1}^{K} \langle C_i|\mathcal{H}|C_j\rangle c_j^{(k)} = E_{\mathrm{CI}}^{(k)} \sum_{j=1}^{K} \langle C_i|C_j\rangle c_j^{(k)} \,, \tag{2.7}$$

where $\langle C_i|C_j\rangle = \delta_{ij}$ as the orbitals are orthonormal.

An advantage of the CI approach is that one obtains not only an approximation to the ground state wave function but also to the higher excited states via the coefficients $c_i^{(k)}$. In fact, a generalized variational principle applies, known as the the McDonald's theorem, which states that the approximate solutions with energies $E_{\mathrm{CI}}^{(0)} \leq E_{\mathrm{CI}}^{(1)} \leq \ldots \leq E_{\mathrm{CI}}^{(K)}$ satisfy

$$E_i \leq E_{\mathrm{CI}}^{(i)} \,, \tag{2.8}$$

where E_i are the exact energies of the eigenstates of the Hamiltonian $\mathcal{H}$. A disadvantage of a CI expansion is that a great number of determinants must be included due to the lack of explicit dependence of the wave function from inter-electron coordinates which makes difficult the description of the cusp

occurring at the electron coalescence points. Moreover, the number of determinants increases very fast with the system size, in particular exponentially with the number of electrons N. A way to limit the number of determinants is to include the most important excitations, for instance single and double (CISD), which yields a computational cost of N^6 with consequent loss of size consistency.

In the multi-configuration self consistent field (MCSF) approach, one optimizes not only the linear coefficients c_i but also the LCAO coefficients a_{ji} to minimize the total energy. A particular type of MCSCF calculation is the complete active space self-consistent (CASCF) approach, where a set of active orbitals is selected, whose occupancy is allowed to vary, while all other orbitals are fixed as either doubly occupied or unoccupied. In a CASSCF(n,m) calculation, n electrons are distributed among an active space of m orbitals and all possible resulting space- and spin-symmetry-adapted CSFs are constructed. The final CASSCF(n,m) wave function consists of a linear combination of these CSFs, like in a full CI calculation for n electrons in m orbitals, except that also the orbitals are now optimized to minimize the total energy.

When several states of the same symmetry are requested, there is a danger in optimizing the higher states that their energy is lowered enough to approach and mix with lower states, thus giving an unbalanced description of excitation energies. A well-established solution to this problem is the use of a state averaged (SA) CASSCF approach where the weighted average of the energies of the states under consideration is optimized

$$E_{\mathrm{SA}} = \sum_I w_I \frac{\langle \Psi_I | \mathcal{H} | \Psi_I \rangle}{\langle \Psi_I | \Psi_I \rangle}\,, \tag{2.9}$$

where $\sum_I w_I = 1$ and the states are kept orthogonal. The wave functions of the different states depend on their individual sets of CI coefficients using a common set of orbitals. Orthogonality is ensured via the CI coefficients and a generalized variational theorem applies. Obviously, the SA-CASSCF energy of the lowest state will be higher than the CASSCF energy obtained without SA. The most important step for a MCSCF/CASSCF calculation is the choice of the active space and, unfortunately, there is not a simple rule to select the proper orbitals. Usually, a great number of trial calculations are necessary to find out which orbitals must be included in the active space.

2.2.2 Density functional theory

When compared to conventional quantum chemistry methods, density functional theory (DFT) is particularly appealing since it does not rely on the

knowledge of the complete N-electron wave function but only of the electronic density. Density functional theory provides an expression for the ground state energy of a system of interacting electrons in an external potential as a functional of the ground state electronic density [14]. Let us assume for simplicity that the spin polarization of the system of interest is identically zero. In the Kohn-Sham formulation of density functional theory [15], the ground state density is written in terms of single-particle orbitals obeying the equations:

$$\left[-\frac{1}{2}\nabla^2 + v_{\text{eff}}\left([n]\,;\mathbf{r}\right)\right]\phi_i = \epsilon_i\phi_i, \tag{2.10}$$

where the electronic density is constructed by summing over the N lowest energy orbitals where N is the number of electrons:

$$n(\mathbf{r}) = \sum_{i=1}^{N} \left|\phi_i(\mathbf{r})\right|^2. \tag{2.11}$$

The effective Kohn-Sham potential is given by

$$v_{\text{eff}}\left([n]\,;\mathbf{r}\right) = v_{\text{ext}}(\mathbf{r}) + \int \frac{n(\mathbf{r}')}{|\mathbf{r}-\mathbf{r}'|}\mathrm{d}\mathbf{r}' + v_{\text{xc}}\left([n]\,;\mathbf{r}\right) \tag{2.12}$$

$v_{\text{ext}}(\mathbf{r})$ is the external potential. The exchange-correlation potential $v_{\text{xc}}\left([n]\,;\mathbf{r}\right)$ is the functional derivative of the exchange-correlation energy $E_{\text{xc}}\left[n\right]$ that enters in the expression for the total energy of the system:

$$\begin{aligned} E &= -\frac{1}{2}\sum_{i=1}^{N}\int \phi_i\nabla^2\phi_i\,\mathrm{d}\mathbf{r} + \int n\left(\mathbf{r}\right)v_{\text{ext}}\left(\mathbf{r}\right)\,\mathrm{d}\mathbf{r} \\ &+ \frac{1}{2}\iint \frac{n(\mathbf{r})n(\mathbf{r}')}{|\mathbf{r}-\mathbf{r}'|}\mathrm{d}\mathbf{r}\,\mathrm{d}\mathbf{r}' + E_{\text{xc}}\left[n\right]. \end{aligned} \tag{2.13}$$

Unfortunately, although the theory unlike HF is in principle exact, the energy functional contains an unknown quantity, called the exchange-correlation energy, $E_{\text{xc}}\left[n\right]$, that must be approximated in any practical implementation of the method. If the functional form of $E_{\text{xc}}\left[n\right]$, and consequently the exchange-correlation potential, were available, we could solve the N-electron problem by finding the self-consistent solution of a set of single-particle equations.

Approximate exchange-correlation functionals

Several approximate exchange-correlation functionals have been proposed in the literature, the most commonly used ones being the local density approximation (LDA), the generalized gradient approximation (GGA) and, more

recently, the hybrid functionals. The local density approximation [15] is the simplest functional:

$$E_{\mathrm{xc}}^{\mathrm{LDA}}[n] = \int \mathrm{d}\mathbf{r}\, \epsilon_{\mathrm{xc}}^{\mathrm{hom}}(n(\mathbf{r}))n(\mathbf{r}) \tag{2.14}$$

where $\epsilon_{\mathrm{xc}}^{\mathrm{hom}}(n)$ is the exchange correlation energy per electron of a uniform electron gas of density n. This functional is by construction exact for a homogeneous electron gas but has been shown to work surprisingly well also when the distribution of electrons is strongly inhomogeneous.

However, LDA does not always provide sufficiently accurate results. For example, it always overestimates the binding energy and the bond length of weak bonded molecules and solids. Therefore, a dependence of the exchange-correlation energy on the derivatives of the electronic density has been introduced in the so-called generalized gradient approximations (GGA), whose generic functional form (here restricted to second-order derivative) is

$$E_{\mathrm{xc}}^{\mathrm{GGA}}[n] = \int n(\mathbf{r})\, \epsilon_{\mathrm{xc}}^{\mathrm{GGA}}(n(\mathbf{r}), |\nabla n(\mathbf{r})|, \nabla^2 n(\mathbf{r}))\, \mathrm{d}\mathbf{r}. \tag{2.15}$$

Many different GGA's are available in the literature and, in this thesis, we will make use of the Becke-Lee-Yang-Parr (BLYP) [16] and the Perdew-Burke-Ehrennshof (PBE) [17] GGA functionals.

In recent years, hybrid functionals have become very popular in particular for chemical applications. These functionals introduce a dependence on the Kohn-Sham orbitals, and mix a portion of exact exchange from Hartree-Fock theory with the exchange and correlation GGA functional:

$$E_{\mathrm{xc}}^{\mathrm{hybrid}}[n] = E_{\mathrm{xc}}^{\mathrm{GGA}}[n] + c_{\mathrm{x}}(E_{\mathrm{x}}^{\mathrm{HF}}[n] - E_{\mathrm{x}}^{\mathrm{GGA}}[n])\,, \tag{2.16}$$

where $E_x^{\mathrm{GGA}}[n]$ and $E_{xc}^{\mathrm{GGA}}[n]$ are GGA exchange (x) and exchange-correlation (xc) energies, and $E_x^{\mathrm{HF}}[n]$ is the exact exchange which has the same form as the HF exchange energy:

$$E_{\mathrm{x}}[n] = -\frac{1}{2}\sum_{i=1}^{N}\sum_{j=1}^{N}\delta_{s_i,s_j}\iint \frac{\phi_i^*(\mathbf{r})\phi_j^*(\mathbf{r}')\phi_j(\mathbf{r})\phi_i(\mathbf{r}')}{|\mathbf{r}-\mathbf{r}'|}\,\mathrm{d}\mathbf{r}\,\mathrm{d}\mathbf{r}'\,. \tag{2.17}$$

The coefficient c_x controls the amount of Hartree-Fock exchange: It is unity for Hartree-Fock, zero for pure DFT, and fractional (typically around 0.25 [18]) for hybrid functionals. This parameter is usually fitted to reproduce a set of properties as for instance atomization energies of first and second-row molecules. A widely used hybrid functional available in most DFT codes is

the three parameter B3LYP functional [19] which combines LDA and the BLYP GGA with exact exchange:

$$\begin{aligned} E_{\mathrm{xc}}^{\mathrm{B3LYP}} &= E_{\mathrm{xc}}^{\mathrm{LDA}} + a_0(E_x^{\mathrm{HF}} - E_x^{\mathrm{LDA}}) \\ &+ a_x(E_x^{\mathrm{GGA}} - E_x^{\mathrm{LDA}}) + a_c(E_c^{\mathrm{GGA}} - E_c^{\mathrm{LDA}}) \end{aligned} \tag{2.18}$$

where $a_0 = 0.20$, $a_x = 0.72$, and $a_c = 0.81$. In this thesis, we will use the hybrid functional B3LYP or PBE0. For more information about DFT, we refer the reader to Refs. [20–22].

Time-dependent density functional theory

Time-dependent density-functional theory (TDDFT) represents a rigorous formalism for the calculations of excitation energies. Similarly to ground state density functional theory, TDDFT is formally exact but relies in practice on the use of approximate exchange-correlation functionals.

The central theorem of TDDFT is the Runge-Gross theorem [23] which generalizes the Hohenberg-Kohn theorem to a time-dependent Hamiltonian, and proves the one-to-one correspondence between the external time-dependent potential $\mathrm{v_{ext}}(\mathbf{r},\mathrm{t})$ and the time-dependent electronic density, $\mathrm{n}(\mathbf{r},\mathrm{t})$. This theorem leads to construct a time-dependent Kohn-Sham scheme for a system of non-interacting electrons in an effective external time-dependent potential:

$$\left[-\frac{1}{2}\nabla^2 + v_{\mathrm{eff}}\left([n]\,;\mathbf{r},t\right)\right]\phi_i(\mathbf{r},t) = i\frac{\partial}{\partial t}\phi_i(\mathbf{r},t), \tag{2.19}$$

which yields the exact electronic density constructed from the Kohn-Sham orbitals as

$$n(\mathbf{r},t) = \sum_{i=1}^{N} \left|\phi_i(\mathbf{r},t)\right|^2 . \tag{2.20}$$

The Kohn-Sham effective potential is given by

$$v_{\mathrm{eff}}\left([n]\,;\mathbf{r},t\right) = v_{\mathrm{ext}}(\mathbf{r},t) + \int \frac{n(\mathbf{r}',t)}{|\mathbf{r}-\mathbf{r}'|}\mathrm{d}\mathbf{r}' + v_{\mathrm{xc}}\left([n]\,;\mathbf{r},t\right) , \tag{2.21}$$

where the first term in the external potential, the second term takes in account the electrostatic interaction between the electrons, and the last term is the exchange-correlation potential. It is important to stress that the time-dependent Kohn-Sham potential is not the same functional of the density

as the ground-state Kohn-Sham potential (Eq. 2.12) but equals the functional derivative of the exchange-correlation component of the action functional [23, 24].

Like in the ground-state DFT approach, the only fundamental approximation in TDDFT is the time-dependent exchange-correlation potential and the quality of the results crucially depends on the quality of this approximation. The simplest approximation is the so-called adiabatic approximation:

$$v_{\mathrm{xc}}^{\mathrm{adiab}}([n];\mathbf{r},t) = v_{\mathrm{xc}}^{\mathrm{gs}}([n];\mathbf{r})|_{n=n(\mathbf{r},t)} \tag{2.22}$$

where $v_{\mathrm{xc}}^{\mathrm{gs}}$ is some given ground-state exchange-correlation potential. The adiabatic approximation therefore assumes that the self-consistent potential is local in time and responds instantaneously and without memory to any temporal change in the charge density. As the $v_{\mathrm{xc}}^{\mathrm{gs}}$ is a ground-state property, we expect that this approximation works best for time-dependent systems whose density does not change too much from the ground-state one. By inserting the LDA or the BLYP potential (or whatever functional one prefers), we obtain what we denote as the approximate adiabatic TDDFT/LDA or TDDFT/BLYP approach.

The excitation energies can be readily obtained from a TDDFT calculation by knowing how the system responds to a small time-dependent perturbation. The key quantity is the linear density response function χ which measures the change in the density of the system due to a small perturbation in the external potential:

$$\delta n_\sigma(\mathbf{r},\omega) = \int \mathrm{d}\mathbf{r}'\chi(\mathbf{r},\mathbf{r}',\omega)\,\delta v_{\mathrm{ext}}(\mathbf{r}',\omega) \tag{2.23}$$

and which allows one to compute the dynamic polarizability and therefore access the photoabsorption cross section. Through the time-dependent Kohn-Sham scheme (Eqs. 2.19–2.21), we can rewrite the same change in the density as

$$\delta n_\sigma(\mathbf{r},\omega) = \int \mathrm{d}\mathbf{r}'\chi_{\mathrm{KS}}(\mathbf{r},\mathbf{r}',\omega)\,\delta v_{\mathrm{eff}}(\mathbf{r}',\omega)\,, \tag{2.24}$$

where χ_{KS} is the density response function of the non-interacting Kohn-Sham electrons which can be written in terms of the unperturbed time-independent Kohn-Sham orbitals. Then, using the definition of the exchange-correlation potential (Eq. 2.20), we can obtain the linear change in the potential as

$$\delta v_{\mathrm{eff}}(\mathbf{r},\omega) = \delta v_{\mathrm{ext}}(\mathbf{r},\omega) + \int \mathrm{d}\mathbf{r}' \left[\frac{1}{|\mathbf{r}-\mathbf{r}'|} + f_{\mathrm{xc}}(\mathbf{r},\mathbf{r}',\omega)\right] \delta n(\mathbf{r}',\omega)\,, \tag{2.25}$$

where $f_{\rm xc}([n];\mathbf{r},\mathbf{r}',\omega)$ is the Fourier transform of the exchange-correlation kernel:

$$f_{\rm xc}([n];\mathbf{r},\mathbf{r}',t-t') = \frac{\delta v_{\rm xc}([n];\mathbf{r},t)}{\delta n(\mathbf{r}',t')}\,. \tag{2.26}$$

Combining Eqs. 2.23–2.25, we derive a Dyson-like equation for the response function

$$\begin{aligned}\chi(\mathbf{r},\mathbf{r}',\omega) &= \chi_{\rm KS}(\mathbf{r},\mathbf{r}',\omega) \\ &+ \int\mathrm{d}\mathbf{x}\int\mathrm{d}\mathbf{x}'\,\chi(\mathbf{r},\mathbf{x},\omega)\left[\frac{1}{|\mathbf{x}-\mathbf{x}'|}+f_{\rm xc}(\mathbf{x},\mathbf{x}',\omega)\right]\chi_{\rm KS}(\mathbf{x}',\mathbf{r}',\omega)\,,\end{aligned} \tag{2.27}$$

which yields the response χ of the interacting system via a self-consistent solution if the exact exchange-correlation kernel is known. Since a full solution of this equation is numerically quite difficult, one obtains the excitation energies by knowing that the density response function, χ, has poles at the frequencies which correspond to the excitation energies of the interacting system. Similarly, the poles of the Kohn-Sham response function, $\chi_{\rm KS}$, correspond to the non-interacting excitation energies given by the difference of Kohn-Sham eigenvalues.

Through a series of algebraic manipulations, it is possible to reformulate linear-response TDDFT in terms of the so-called Casida's equations [25] where the poles of the response functions, $\Omega = E_m - E_0$, are determined as solutions of the non-Hermitian eigenvalue problem:

$$\begin{bmatrix}\mathbf{A} & \mathbf{B}\\ \mathbf{B} & \mathbf{A}\end{bmatrix}\begin{pmatrix}\vec{X}\\ \vec{Y}\end{pmatrix} = \Omega\begin{bmatrix}\mathbf{-1} & \mathbf{0}\\ \mathbf{0} & \mathbf{1}\end{bmatrix}\begin{pmatrix}\vec{X}\\ \vec{Y}\end{pmatrix}, \tag{2.28}$$

where the matrices $\mathbf{A}$ and $\mathbf{B}$ are defined as

$$\begin{aligned}A_{ia,i'a'} &= \delta_{ii'}\delta_{aa'}(\epsilon_a-\epsilon_i)+K_{ia,i'a'}\,,\\ B_{ia,i'a'} &= K_{ia,a'i'} = (ia|\frac{1}{|\mathbf{r}-\mathbf{r}'|}|a'i') + (ia|f_{\rm xc}|a'i')\,.\end{aligned} \tag{2.29}$$

The eigenvalues of these equations give the excitation energies and the eigenvectors can be used to compute the oscillator strengths.

It is important to note that TDDFT adiabatic approximation includes only dressed one-electron excitations [26]. This can be seen in the context of the Tamm-Dancoff approximation (TDA), which consists of neglecting the $\mathbf{B}$ matrices to obtain $\mathbf{A}\vec{X} = \Omega\vec{X}$. The number of possible solutions to this equation is the dimensionality of $\mathbf{A}$ which is the number of single excitations. In fact, the linear-response time-dependent Hartree-Fock TDA (exchange-only density functional theory) is simply the well-known configuration interaction

singles (CIS) method. The computational cost of TDDFT scales approximately like $O(N^3)$, so TDDFT represents a very appealing theoretical approach to compute the excitation energies of large molecular systems. While often reasonably accurate, the main difficulties encountered in conventional TDDFT include the underestimation of the ionization threshold [27], the underestimation of charge transfer excitations [28–30], and the lack of explicit two- and higher-electron excitations [26, 31]. These shortcomings may be fatal in describing the excitations of biomolecular systems as these systems are often characterized by charge transfer in the excited states and may display multi-configurational character.

2.2.3 Quantum Monte Carlo methods

The variational (VMC) and diffusion (DMC) Monte Carlo methods we present in this Section share with conventional quantum chemistry methods that they are wave function based approaches. However, differently from quantum chemistry approaches, they attempt to solve the Schrödinger equation stochastically, and have consequently significantly more freedom in the choice of the functional form of the many-body correlated wave function. Moreover, both approaches have a more favorable scaling with the number of electrons, that is, N^4 when compared to N^6 of CISD or N^7 of the coupled cluster single and double with perturbative triples approach. Therefore, even though they are more costly than DFT which only scales as N^3, they are significantly faster than conventional highly-correlated quantum chemistry methods, and can be applied to larger systems and to solids to provide accurate answers in situations when DFT is shown to be inadequate.

Variational Monte Carlo

The variational Monte Carlo method is the simplest QMC approach and uses Monte Carlo techniques to evaluate the expectation value of an operator for a given wave function. Let us assume we are given the many-body trial wave function Ψ_{T} and that we are interested in computing the expectation value of the Hamiltonian $\mathcal{H}$:

$$E_V = \frac{\int \Psi_{\mathrm{T}}^*(\mathbf{R})\mathcal{H}\Psi_{\mathrm{T}}(\mathbf{R})d\mathbf{R}}{\int \Psi_{\mathrm{T}}^*(\mathbf{R})\Psi_{\mathrm{T}}(\mathbf{R})\mathrm{d}\mathbf{R}} \tag{2.30}$$

This expectation value can be rewritten as

$$E_V = \frac{\int |\Psi_{\mathrm{T}}(\mathbf{R})|^2[\Psi_{\mathrm{T}}(\mathbf{R})^{-1}\mathcal{H}\Psi_{\mathrm{T}}(\mathbf{R})]d\mathbf{R}}{\int |\Psi_{\mathrm{T}}(\mathbf{R})|^2 d\mathbf{R}} = \int \rho(\mathbf{R})E_L(\mathbf{R})\mathrm{d}\mathbf{R}\,, \tag{2.31}$$

where

$$\rho(\mathbf{R}) = \frac{|\Psi_{\mathrm{T}}(\mathbf{R})|^2}{\int |\Psi_{\mathrm{T}}(\mathbf{R})|^2 \mathrm{d}\mathbf{R}}\,, \tag{2.32}$$

and the local energy is defined as

$$E_L(\mathbf{R}) = \Psi_{\mathrm{T}}(\mathbf{R})^{-1}\mathcal{H}\Psi_{\mathrm{T}}(\mathbf{R})\,, \tag{2.33}$$

Since $\rho(\mathbf{R})$ is a positive quantity and integrates to 1, we can interpret it as a probability distribution and use Monte Carlo techniques to sample a set of configurations $\{\mathbf{R}_m\}$ distributed according to $\rho(\mathbf{R})$. The expectation value can then be estimated as an average of the local energy $E_L(\mathbf{R})$ evaluated on these configurations:

$$E_V \approx \frac{1}{M}\sum_{m=1}^{M} E_L(\mathbf{R}_m) \tag{2.34}$$

Note that in this derivation, we can substitute the Hamiltonian $\mathcal{H}$ with any operator $\mathcal{O}$ diagonal in space representation.

For a realistic system of electrons, the square of the many-body wave function is a complicated probability distribution in a high-dimensional space, of which we do not usually know how to compute the normalization. Therefore, we cannot use direct sampling techniques but we employ the Metropolis algorithm to generate a sequence of configurations $\{\mathbf{R}_m\}$ distributed according to $\rho(\mathbf{R})$. The Metropolis algorithm is a general method to sample an arbitrary probability distribution without knowing its normalization, and is an application of a Markov chain. In a Markov chain, one changes the state of the system randomly from an initial state $\mathbf{R}_{\mathrm{i}}$ to a final state $\mathbf{R}_{\mathrm{f}}$ according to the stochastic transition matrix $M(\mathbf{R}_{\mathrm{f}}|\mathbf{R}_{\mathrm{i}})$ which satisfies

$$M(\mathbf{R}_{\mathrm{f}}|\mathbf{R}_{\mathrm{i}}) \geq 0 \quad \text{and} \quad \sum_{\mathrm{f}} M(\mathbf{R}_{\mathrm{f}}|\mathbf{R}_{\mathrm{i}}) = 1\,. \tag{2.35}$$

To sample the desired distribution $\rho(\mathbf{R})$, one evolves the the system by repeated application of a Markov matrix M which satisfies the *stationarity condition*

$$\sum_{i} M(\mathbf{R}_{\mathrm{f}}|\mathbf{R}_{\mathrm{i}})\ \rho(\mathbf{R}_i) = \rho(\mathbf{R}_f)\,,$$

for any state $\mathbf{R}_{\mathrm{f}}$. The stationarity condition condition tells us that if we start from the desired distribution ρ, we will continue to sample ρ. Moreover, if the stochastic matrix M is ergodic, this condition ensures that any initial distribution will evolve to ρ under repeated applications of M. Therefore, ρ

is the right eigenvector of M with eigenvalue 1 and it is also the dominant eigenvector.

In practice, one imposes the more stringent *detailed balance* condition

$$M(\mathbf{R}_\mathrm{f}|\mathbf{R}_\mathrm{i})\ \rho(\mathbf{R}_i) = M(\mathbf{R}_\mathrm{i}|\mathbf{R}_\mathrm{f})\ \rho(\mathbf{R}_f) \tag{2.36}$$

which is a sufficient but not necessary condition to satisfy the stationarity condition as can be easily seen by summing both sides of the equation over $\mathbf{R}_\mathrm{i}$ and using Eq. 2.35. The transition M is then rewritten as the product of a proposal matrix T and the acceptance A:

$$M(\mathbf{R}_\mathrm{f}|\mathbf{R}_\mathrm{i}) = A(\mathbf{R}_\mathrm{f}|\mathbf{R}_\mathrm{i})\ T(\mathbf{R}_\mathrm{f}|\mathbf{R}_\mathrm{i})\,, \tag{2.37}$$

where M and T are stochastic matrices but A is not. The detailed balance condition finally becomes

$$\frac{A(\mathbf{R}_\mathrm{f}|\mathbf{R}_\mathrm{i})}{A(\mathbf{R}_\mathrm{i}|\mathbf{R}_\mathrm{f})} = \frac{T(\mathbf{R}_\mathrm{i}|\mathbf{R}_\mathrm{f})\ \rho(\mathbf{R}_f)}{T(\mathbf{R}_\mathrm{f}|\mathbf{R}_\mathrm{i})\ \rho(\mathbf{R}_i)}. \tag{2.38}$$

For a given T, the choice originally made by Metropolis *et al.* [32] for the acceptance is

$$A(\mathbf{R}_\mathrm{f}|\mathbf{R}_\mathrm{i}) = \min\left\{1, \frac{T(\mathbf{R}_\mathrm{i}|\mathbf{R}_\mathrm{f})\ \rho(\mathbf{R}_f)}{T(\mathbf{R}_\mathrm{f}|\mathbf{R}_\mathrm{i})\ \rho(\mathbf{R}_i)}\right\}\,, \tag{2.39}$$

and is the one which maximizes the acceptance. In choosing the proposal matrix T, we observe that the Metropolis algorithm generates points which are sequentially correlated so that the effective number of independent observations in a Monte Carlo run of M steps is M/T_corr, where T_corr is the autocorrelation time of the observable of interest. Therefore, to achieve a fast evolution and reduce T_corr, the optimal T should yield a high acceptance and at the same time allow large proposed moves. The choice of T will of course be limited by the fact that we need to be able to sample T directly. We use here the algorithm described in Ref. [33], which uses a non-symmetrical T and which we properly modified to deal with pseudopotentials.

In short, the generalized Metropolis algorithm will consist of the following steps:

1. Choose the distribution $\rho(\mathbf{R})$ and the transition probability $T(\mathbf{R}_\mathrm{f}|\mathbf{R}_\mathrm{i})$.
2. Initialize the configuration $\mathbf{R}_\mathrm{i}$.
3. Advance the configuration from $\mathbf{R}_\mathrm{i}$ to $\mathbf{R}'$:
   a) Sample $\mathbf{R}'$ from $T(\mathbf{R}'|\mathbf{R}_\mathrm{i})$.

b) Calculate the ratio

$$q = \frac{T(\mathbf{R}_\mathrm{i}|\mathbf{R}')}{T(\mathbf{R}'|\mathbf{R}_\mathrm{i})} \frac{\rho(\mathbf{R}')}{\rho(\mathbf{R}_\mathrm{i})} \,. \tag{2.40}$$

c) Accept or reject: If $q > 1$ or $q > r$ where r is a uniformly distributed random number in (0,1), set the new configuration $\mathbf{R}_\mathrm{f} = \mathbf{R}'$. Otherwise, set $\mathbf{R}_\mathrm{f} = \mathbf{R}_\mathrm{i}$.

4. Throw away the first κ configurations corresponding to the equilibration time.
5. Collect the averages and block them to obtain the error bars.

Two final comments on the Metropolis algorithm. First, the distribution $\rho(\mathbf{R})$ does not have to be normalized since only ratios enter in the acceptance. Therefore, it is possible to sample the square of complex wave functions (Eq. 2.32) whose normalization we do not know. Second, if $M_1, M_2, \ldots, M_n$ are matrices which satisfy the stationarity condition, the matrix $M = \prod_{i=1}^{n} M_i$ also satisfies the stationarity condition. Consequently, particles can be moved one at the time, a necessary feature as the system size grows since the size of the move would need to be decreased to have a reasonable acceptance of a move of all particles.

Many-body wave functions used in quantum Monte Carlo

The use of VMC to compute the expectation values of quantum mechanical operators allows great freedom in the choice of the trial wave function which on the other hand determines the accuracy as well as the efficiency of the calculation. Therefore, the form of wave function should yield accurate results while being compact and easy to evaluate.

The ingredients entering in the wave function most commonly used in quantum Monte Carlo can be understood by inspecting the advantages and limitations of traditional quantum chemistry approaches. Methods such as configuration interaction (CI) expand the many body wave function in a linear combination of Slater determinants of single-particle spin-orbitals. This form allows the evaluation of the high-dimensional integrals in all expectation values but the convergence of the expansion is very slow, in part because of the difficulty in describing the cusps which occur as two electrons approach each other. Quantum Monte Carlo uses a much more compact representation of the wave function which is usually given by a sum of few determinants (tens and not millions like in a CI calculation) multiplied by a component which can exactly impose the cusps at the inter-particle coalescence points.

Slater-Jastrow wave function

The trial wave functions used in our quantum Monte Carlo calculations are of the Jastrow-Slater form, thus they are a product between a sum of determinants of single particle orbitals, and a Jastrow correlation factor:

$$\Psi(\mathbf{r}_1,\ldots,\mathbf{r}_N) = \mathcal{J}(\mathbf{r}_1,\ldots,\mathbf{r}_N)\sum_k d_k D_k^{\uparrow}(\mathbf{r}_1,\ldots,\mathbf{r}_{N_\uparrow}) D_k^{\downarrow}(\mathbf{r}_{N_\uparrow+1},\ldots,\mathbf{r}_N)\,, \tag{2.41}$$

where $D_k^{\uparrow}$ and $D_k^{\downarrow}$ are Slater determinants of single particle orbitals for the up and down spin electrons, respectively. The orbitals are a linear combination of Slater functions centered on the atoms for all-electron calculations while are expanded on a Gaussian basis when pseudopotentials are employed. The Jastrow correlation function is a positive function of the interparticle distances and explicitly depends on the electron-electron separations.

The wave function is here written in spin-assigned form where the dependence on the spin variables $\{\sigma_i\}$ has disappeared and the wave function appears to no longer be fully antisymmetric. To obtain such an expression, we start from a full wave function $\Psi(\mathbf{x}_1,\ldots,\mathbf{x}_N)$ depending on both spatial and spin coordinates, and expand it on its spin components. For a a system of N electrons with $N = N_\uparrow + N_\downarrow$ and $S_z = (N_\uparrow - N_\downarrow)/2$, we introduce a spin function ζ_1

$$\zeta_1(\sigma_1,\ldots,\sigma_N) = \chi_\uparrow(\sigma_1)\ldots\chi_\uparrow(\sigma_{N_\uparrow})\chi_\downarrow(\sigma_{N_\uparrow+1})\ldots\chi_\downarrow(\sigma_N)\,. \tag{2.42}$$

and construct a set of $K = N!/(N_\uparrow!N_\downarrow!)$ distinct spin functions ζ_i by permuting the indices in ζ_1. Since the spin functions ζ_i form a complete orthonormal set in spin space, we can decompose the wave function Ψ in terms of its spin components as

$$\Psi(\mathbf{x}_1,\ldots,\mathbf{x}_N) = \sum_{i=1}^{K} F_i(\mathbf{r}_1,\ldots,\mathbf{r}_N)\zeta_i(\sigma_1,\ldots,\sigma_N)\,. \tag{2.43}$$

As Ψ is antisymmetric under the interchange of particle indices, each function F_i is antisymmetric under the interchange of like-spin electrons and all F_i are the same except for a relabelling of the particle indices and a change in sign for odd permutations. Therefore, the wave function can be rewritten as:

$$\Psi(\mathbf{x}_1,\ldots,\mathbf{x}_N) = \mathcal{A}\left\{F_1(\mathbf{r}_1,\ldots,\mathbf{r}_N)\zeta_1(\sigma_1,\ldots,\sigma_N)\right\} \tag{2.44}$$

It is easy to show using orthonormality of the functions ζ_i that the expectation value of an operator $\mathcal{O}$ which is spin-independent is the same if we use the fully antisymmetric wave function Ψ or just one spatial function, say F_1:

$$\langle\Psi|\mathcal{O}|\Psi\rangle = \langle F_1|\mathcal{O}|F_1\rangle\,. \tag{2.45}$$

Since it is more convenient to use the function F_1 than the full wave function Ψ, in quantum Monte Carlo, we always work with spin-assigned wave functions. To obtain F_1, we simply assign the spin-variables of the particles as

$$\begin{array}{cccccccc} \text{Particle} & 1 & 2 & \dots & N_\uparrow & N_{\uparrow+1} & \dots & N \\ \sigma & 1 & 1 & \dots & 1 & -1 & \dots & -1 \end{array}$$

so that $F_1(\mathbf{r}_1, \dots, \mathbf{r}_N) = \Psi(\mathbf{r}_1, 1, \dots, \mathbf{r}_{N_\uparrow}, 1, \mathbf{r}_{N_\uparrow+1}, -1, \dots, \mathbf{r}_N, -1)$.

Finally, the Jastrow-Slater spin-assigned wave function obtained by imposing $\sigma = +1$ for first $N_\uparrow$ particles and $\sigma = -1$ for the others is given by

$$\begin{aligned} \Psi(\mathbf{r}_1, \dots, \mathbf{r}_N) &= F_1(\mathbf{r}_1, \dots, \mathbf{r}_N) \\ &= \mathcal{J} \sum_k d_k \, D_k^\uparrow(\mathbf{r}_1, \dots, \mathbf{r}_{N_\uparrow}) D_k^\downarrow(\mathbf{r}_{N_\uparrow+1}, \dots, \mathbf{r}_N) \end{aligned} \quad (2.46)$$

where $\mathcal{J} = \mathcal{J}(\mathbf{r}_1, \dots, \mathbf{r}_N)$ is the Jastrow factor.

The Jastrow factor is generally chosen to be a positive function of the interparticle distances and therefore does not affect the sign of the wave function which is solely determined by the determinantal component. At a large interparticle distances, it plays no role since it becomes constant, which we achieve by using scaled variables as shown below. The Jastrow factor is of fundamental importance in describing correlation at short and intermediate distances. In particular, the electron-electron cusp conditions are imposed through the Jastrow factor: At the electron-electron coalescence points, the potential energy diverges at infinity, so the kinetic energy must have an opposite divergence to the potential to keep the local energy finite. It is possible to ensure this cancellation if the trial wave function satisfies a set of cusp conditions and displays a proper discontinuity of the derivatives at the coalescence points.

The form of the Jastrow factor which we use depends on electron-electron and the electron-nucleus distances and describes electron-electron, electron-nucleus and electron-electron-nucleus correlations:

$$\begin{aligned} \mathcal{J}(\mathbf{r}_1, \dots, \mathbf{r}_N) = \prod_{\alpha,i} \exp\left\{A(r_{i\alpha})\right\} \times \prod_{i<j} \exp\left\{B(r_{ij})\right\} \times \\ \times \prod_{\alpha, i<j} \exp\left\{C(r_{i\alpha}, r_{j\alpha}, r_{ij})\right\} . \end{aligned} \quad (2.47)$$

The electron-nucleus terms A should be included if the determinantal part is constructed using orbitals obtained from a DFT or a HF calculation and not reoptimized after the inclusion of the electron-electron Jastrow factor.

As the electron-electron term alters the single-particle density by reducing/increasing it in high/low density regions, the resulting density will in general be worse that the original DFT or HF density which can be repristinated by the inclusion of the electron-nucleus terms. The electron-electron term B is introduced to impose the electron-electron cusp conditions and to keep the electrons apart as the electron-electron interaction is repulsive. Finally, the electron-electron-nucleus terms C can in principle exactly describe a two-electron atom or ion in an S state. Higher body correlations are clearly less important as, due to the exclusion principle, it is rare for three or more electrons to be close since at least two electrons must necessarily have the same spin.

To keep the Jastrow factor finite at large distances, we use scaled variables $\bar{r} = (1 - e^{-\kappa r})/\kappa$ for the A and B terms, and $\bar{r} = e^{-\kappa r}$ for the C terms. The particular form we employ in this work is:

$$\begin{aligned} A(r_{i\alpha}) &= \frac{a_1 \bar{r}_{i\alpha}}{1 + a_2 \bar{r}_{i\alpha}} + \sum_{p=2}^{N^a_{\rm ord}} a_{p+1} \bar{r}^p_{i\alpha} \\ B(r_{ij}) &= \frac{b_1 \bar{r}_{ij}}{1 + b_2 \bar{r}_{ij}} + \sum_{p=2}^{N^b_{\rm ord}} b_{p+1} \bar{r}^p_{ij} \\ C(r_{i\alpha}, r_{j\alpha}, r_{ij}) &= \sum_{p=2}^{N^c_{\rm ord}} \sum_{0}^{k=p-1} \sum_{0}^{l=l_{\rm max}} c_{mkl} \bar{r}^k_{ij} (\bar{r}^l_{i\alpha} + \bar{r}^l_{j\alpha}) (\bar{r}_{i\alpha} \bar{r}_{j\alpha})^m, \end{aligned} \quad (2.48)$$

where $m = (p - k - l)/2$, and $l_{\rm max}$ is $p-k$ if $k \neq 0$ and $p-k-2$ if $k = 0$. Only terms for which $m = (p - k - l)/2$ is an integer are included. The a and c coefficients are different for different atom types. The only spin dependence is in b_1 which is is used to satisfy the electron-electron cusp conditions: $b_1 = 1/2$ for antiparallel spin, and $b_1 = 1/4$ for parallel electrons.

Diffusion Monte Carlo

VMC is a very useful tool which allows us to explore which type of correlation is relevant in the system under study with a relatively small amount of computer time. For instance, we can investigate how the complexity of the trial wave function influences the description of the state of interest. However, its major drawback is that the result will uniquely depend on the the quality of the trial wave function which cannot be constructed in an automatic way and whose functional form must be chosen for each particular problem. Therefore, we would like to have a way to remove (at least in part) the bias introduced by the wave function.

Projector Monte Carlo is a more powerful method than VMC and removes (at least in part) the bias of the trial wave function from the results. It is a stochastic implementation of the power method for finding the dominant eigenstate of a matrix or integral kernel. In a projector Monte Carlo method, one uses an operator that inverts the spectrum of $\mathcal{H}$ to project out the ground state of $\mathcal{H}$ from a given trial state. Different operators have been used as projectors but, here, for simplicity, we only discuss diffusion Monte Carlo (DMC) which we use in our calculations.

In DMC, we use as projection operator $\exp[-\tau(\mathcal{H} - E_{\mathrm{T}})]$, and given an initial trial wave function $\Psi^{(0)}$, we repeatedly apply the projection operator to obtain the sequence of wave functions:

$$\Psi^{(n)} = e^{-\tau(\mathcal{H}-E_{\mathrm{T}})}\Psi^{(n-1)} \,. \tag{2.49}$$

If we expand the initial wave function $\Psi^{(0)}$ on the eigenstates Ψ_i with energies E_i of $\mathcal{H}$, we obtain for $\Psi^{(n)}$:

$$\Psi^{(n)} = \sum_i \Psi_i \, \langle \Psi^{(0)} | \Psi_i \rangle e^{-n\tau(E_i - E_{\mathrm{T}})} \,, \tag{2.50}$$

where $\langle \Psi^{(0)} | \Psi_i \rangle$ is the overlap between $\Psi^{(0)}$ and the eigenstate Ψ_i. Since the coefficients of the excited states die off exponentially fast relative to the coefficient of the ground state, we obtain

$$\lim_{n\to\infty} \Psi^{(n)} = \Psi_0 \langle \Psi^{(0)} | \Psi_0 \rangle e^{-n\tau(E_0 - E_{\mathrm{T}})} \,. \tag{2.51}$$

Therefore, if we choose the trial energy $E_{\mathrm{T}} \approx E_0$ to keep the over all normalization of $\Psi^{(n)}$ fixed, the projection yields the ground state Ψ_0 of the Hamiltonian. Note that the starting wave function must have a non-zero overlap with the ground state one.

To see how to perform the projection, let us first rewrite Eq. 2.49 in integral from and obtain

$$\Psi^{(n)}(\mathbf{R}', t+\tau) = \int \mathrm{d}\mathbf{R} \, G(\mathbf{R}', \mathbf{R}, \tau) \Psi^{(n-1)}(\mathbf{R}, t) \,, \tag{2.52}$$

where the coordinate Green's function is defined as

$$G(\mathbf{R}', \mathbf{R}, \tau) = \langle \mathbf{R}' | e^{-\tau(\mathcal{H}-E_{\mathrm{T}})} | \mathbf{R} \rangle \,. \tag{2.53}$$

If we can sample the trial wave function and the Green's function in Eq. 2.52, we can perform this high-dimensional integral by Monte Carlo integration. For fermions, since the wave function must be antisymmetric, it cannot be

interpreted as a probability distribution. Therefore, for the moment, we will assume that we are dealing with bosons which are characterized by a positive ground state wave function.

Using the Trotter-Suzuki formula it is possible to show that the approximate Green's function for small τ is given by:

$$\langle \mathbf{R}'|e^{-\mathcal{H}\tau}|\mathbf{R}\rangle \approx \frac{1}{(2\pi\tau)^{3N/2}} \exp\left[-\frac{(\mathbf{R}'-\mathbf{R})^2}{2\tau}\right] \exp\left[-\tau\,\mathcal{V}(\mathbf{R})\right] . \quad (2.54)$$

Therefore, the iteration in Eq. 2.52 can be interpreted as a Markov process with the difference that the Green's function is not normalized and we obtain a branching random walk: the first factor in the short-time Green's function is the Green's function for diffusion while the second term multiplies the distribution by a positive scalar. Since the short-time expression of the Green's function is only valid in the limit of τ approaching zero, in practice, DMC calculations must be performed for different values of τ and the result extrapolated for τ which goes to zero.

The use of this Green's function would however yield a highly inefficient and unstable algorithm since the potential can vary significantly from configuration to configuration or also be unbounded like the Coulomb potential. For example, the electron-nucleus potential diverges to minus infinity as the two particles approach each other, and the branching factor will give raise to an unlimited number of walkers. Even if the potential is bounded, the approach becomes inefficient with increasing size of the system since the branching factor also grows with the number of particles.

These difficulties can be overcame by using *importance sampling* which was originally proposed by Kalos [34] for Green's function Monte Carlo and extended by Ceperley and Alder [35] to DMC. We start from Eq. 2.52, multiply each side by a trial wave function Ψ and define the probability distribution $f^{(n)}(\mathbf{R}) = \Psi(\mathbf{R})\Psi^{(n)}(\mathbf{R})$ which satisfies

$$f^{(n)}(\mathbf{R}', t+\tau) = \int \mathrm{d}\mathbf{R}\, \tilde{G}(\mathbf{R}', \mathbf{R}, \tau) f^{(n-1)}(\mathbf{R}, t)\,, \quad (2.55)$$

where the importance sampled Green's function is given by

$$\tilde{G}(\mathbf{R}', \mathbf{R}, \tau) = \Psi(\mathbf{R}')\langle \mathbf{R}'|e^{-\tau(\mathcal{H}-E_\mathrm{T})}|\mathbf{R}\rangle/\Psi(\mathbf{R})\,. \quad (2.56)$$

It is possible to show that resulting drift-diffusion-branching short-time Green's function is given by

$$\tilde{G}(\mathbf{R}', \mathbf{R}, \tau) = (2\pi\tau)^{3N/2} \exp\left[-\frac{(\mathbf{R}'-\mathbf{R}-\tau\mathbf{V}(\mathbf{R}))^2}{2\tau}\right] \times \\ \times \exp\left\{-\tau\left[(E_\mathrm{L}(\mathbf{R}) + E_\mathrm{L}(\mathbf{R}'))/2 - E_\mathrm{T}\right]\right\} + O(\tau^2)\,. \quad (2.57)$$

where the quantum velocity is defined as

$$\mathbf{V}(\mathbf{R}) = \frac{\nabla\Psi(\mathbf{R})}{\Psi(\mathbf{R})}\,. \tag{2.58}$$

There are two important new features of $\tilde{G}(\mathbf{R}', \mathbf{R}, \tau)$. First, the quantum velocity $\mathbf{V}(\mathbf{R})$ pushes the walkers to regions where $\Psi(\mathbf{R})$ is large. In addition, the local energy $E_{\mathrm{L}}(\mathbf{R})$ instead of the potential $\mathcal{V}(\mathbf{R})$ appears in the branching factor. Since the local energy becomes constant and equal to the eigenvalue as the trial wave function approaches the exact eigenstate, we expect that, for a good trial wave function, the fluctuations in the branching factor will be significantly smaller. In particular, imposing the cusp conditions on the wave function will remove the instabilities coming from the singular Coulomb potential.

The DMC algorithm will now be:

1. A set of M_0 configurations is sampled from $|\Psi(\mathbf{R})|^2$ using the Metropolis algorithm. This is the zero-th *generation* and the number of configurations is the *population* of the zero-th generation.
2. The walkers are advanced as $\mathbf{R}' = \mathbf{R} + \xi + \tau\mathbf{V}(\mathbf{R})$ where ξ is a normally distributed $3N$ dimensional random vector, and the last term is the drift.
3. For each walker, compute the factor

$$p = \exp\left\{-\tau[(E_{\mathrm{L}}(\mathbf{R}) + E_{\mathrm{L}}(\mathbf{R}'))/2 - E_{\mathrm{T}}]\right\}\,. \tag{2.59}$$

   Branch the walker by treating p as the probability to survive at the next step: if $p < 1$, the walker survives with probability p while, if $p > 1$, the walker continues and new walkers with the same coordinates are created with probability $p - 1$. This is achieved by creating a number of copies of the current walker equal to the integer part of $p + \eta$ where η is a random number between (0,1).
4. The trial energy E_{T} is adjusted to keep the population stable around the target population M_0.

In Ref. [36], the reader can find a thourough description of several improvements one can bring to the simple algorithm outlined above.

So far, we have assumed that the wave function is positive everywhere and we have not yet addressed the problem posed by the fact that electrons are fermions and that the trial wave function must be antisymmetric. Unfortunately, straightforward generalizations of the DMC algorithm to handle

both signs of the wave functions, even if formally correct, lead to the fermion sign problem: The bosonic component grows at the expenses of the fermionic one and the antisymmetric signal is lost in the noise. To avoid this problem, we can simply forbid moves in which the sign of the trial wave function changes and the walker crosses the nodes which are defined as the set of points where the trial wave function is zero. This procedure is known as the *fixed-node approximation*. Forbidding node crossing is equivalent to finding the solution of the evolution equation with the boundary condition that it has the same nodes as the trial wave function. The Schrödinger equation is therefore solved exactly inside the nodal regions but not at the nodes where the solution will have a discontinuity of the derivatives. The fixed-node solution will be exact only if the nodes of the trial wave function are exact. In general, the fixed-node energy will be an upper bound to the exact energy, in particular the best upper bound consistent with the boundary conditions given by the nodes of the trial wave function.

Wave function optimization

The quality of the trial wave function controls the statistical efficiency of the VMC and DMC algorithms and determines the final accuracy of the results. The ability of optimizing the parameters of the trial wave function is crucial for the success of quantum Monte Carlo methods. For the optimization of the parameters in the trial wave function of a system in its ground state, we use the optimization method within energy minimization recently proposed by Umrigar, Filippi, and Sorella [37], which we briefly describe below.

The Jastrow-Slater wave function depends on a set of parameters p:

$$\Psi(p, \mathbf{R}) = \mathcal{J}(\alpha, \mathbf{R}) \sum_{i=1}^{N_{CSF}} c_i C_i(\eta, \mathbf{R}) \tag{2.60}$$

where the parameters p are given by parameters α in the Jastrow factors, the linear parameters c_i in from of the CSFs, and the LCAO coefficients η which enter in the single-particle orbitals. An optimal set of linear coefficients is readily obtained by solving the generalized eigenvalue problem

$$\sum_{j=1}^{N_{\mathrm{CSF}}} H_{ij} c_j = E_I \sum_{j=1}^{N_{\mathrm{CSF}}} S_{ij} c_j \,. \tag{2.61}$$

The Hamiltonian and overlap matrix elements are estimated by a finite-sample average in variational Monte Carlo as

$$H_{ij} = \left\langle \frac{\mathcal{J}C_i}{\Psi} \frac{\mathcal{H}\mathcal{J}C_j}{\Psi} \right\rangle_{\Psi^2} , \quad S_{ij} = \left\langle \frac{\mathcal{J}C_i}{\Psi} \frac{\mathcal{J}C_j}{\Psi} \right\rangle_{\Psi^2} \tag{2.62}$$

where the statistical average is over the Monte Carlo configurations sampled from Ψ^2. Importantly, the use of the non-symmetric estimator of the Hamiltonian matrix of Eq. (2.62) yields a strong zero-variance principle [38] and results in a particularly efficient approach.

To find the optimal linear (c) as well as non-linear (α and η) parameters, we linearize the wave function around the current set of parameters p, and consider the changes in Ψ given by $\Psi_k = (\partial\Psi/\partial p_k)$, which can be made orthogonal to Ψ as

$$\bar{\Psi}_k = \Psi_k - \left\langle \frac{\Psi_k}{\Psi} \right\rangle_{\Psi^2} \Psi \,. \tag{2.63}$$

We now work in the semi-orthogonal basis of the functions $\{\bar{\Psi}_0, \bar{\Psi}_k\}$ where $\bar{\Psi}_0 = \Psi$, and find the variations Δp_i in the parameters as the lowest eigenvalue solution of the generalized eigenvalue problem

$$H_{ij}\Delta p_j = E\, S_{ij}\Delta p_j \tag{2.64}$$

where $\Delta p_0 = 1$. When the parameter values are far from optimal, the new parameters $p_i + \Delta p_i$ can be worse than the old ones. Therefore, to ensure convergence, a positive constant a_{diag} is added to the diagonal of the Hamiltonian matrix (apart the first element):

$$H_{ij} = H_{ij} + a_{diag}\delta_{ij}(1 - \delta_{i0}) \tag{2.65}$$

which is adjusted at each optimization step.

Wave function optimization and excited states

As we are not only interested in ground state problems, we want to be able to optimize the parameters of the multiple (ground and excited) states described by the wave functions:

$$\Psi_I = \sum_{i=1}^{N_{\mathrm{CSF}}} c_i^I \mathcal{J} C_i \,. \tag{2.66}$$

which share the same Jastrow factor and orbitals but different linear coefficients. To this end, we follow a state-average (SA) approach to determine a set of orbitals and a Jastrow factor which give a comparably good description of the states under considerations while preserving orthogonality among the states. We alternate the optimization of the linear coefficients as outlined above to a micro-iteration in which the optimized quantity with respect to

the orbital and Jastrow variations is the weighted average of the energies of the states under consideration:

$$E_{\mathrm{SA}} = \sum_{I \in \mathrm{A}} w_I \frac{\langle \Psi_I | \mathcal{H} | \Psi_I \rangle}{\langle \Psi_I | \Psi_I \rangle}, \tag{2.67}$$

where the weights w_I are fixed and $\sum_I w_I = 1$. Therefore, at convergence, the averaged energy E_{SA} is stationary with respect to all parameter variations subject to the orthogonality constraint while the individual state energies E_i are stationary with respect to variations of the linear coefficients but not with respect to variations of the orbital or Jastrow parameters. In this approach, the wave functions are kept orthogonal and a generalized variational theorem applies.

To improve the orbital and Jastrow parameters at each SA micro-iteration step, we extend the linear optimization approach of Ref. [37] we described above to the SA optimization of multiple states. Under a common variation in an orbital or Jastrow parameter p_i, the changes in the states Ψ_I are given by $\Psi_k^I = (\partial \Psi^I / \partial p_k)$ and can be made orthogonal to the corresponding state as

$$\bar{\Psi}_k^I = \Psi_k^I - \left\langle \frac{\Psi^I}{\Psi_g} \frac{\Psi_k^I}{\Psi_g} \right\rangle_{\Psi_g^2} \Psi^I . \tag{2.68}$$

To linearize the minimization with respect to the non-linear parameters, we work in the semi-orthogonal basis of the functions $\{\bar{\Psi}_0^I, \bar{\Psi}_k^I\}$ where $\bar{\Psi}_0^I = \Psi^I$, and find the variations Δp_i in the parameters as the lowest eigenvalue solution of the generalized eigenvalue problem

$$H_{ij}^{\mathrm{SA}} \Delta p_j = E \, S_{ij}^{\mathrm{SA}} \Delta p_j \tag{2.69}$$

where $\Delta p_0 = 1$. The SA Hamiltonian matrix is computed as

$$H_{ij}^{\mathrm{SA}} = \sum_{I \in \mathrm{A}} w_I \left\langle \frac{\bar{\Psi}_i^I}{\Psi_g} \frac{H \bar{\Psi}_j^I}{\Psi_g} \right\rangle_{\Psi_g^2}, \tag{2.70}$$

and an analogous definition holds for the SA overlap matrix elements. The matrix elements for all states are computed in a single variational Monte Carlo run with guiding wave function Ψ_g. At convergence and for the optimal linear coefficients, the minimal energy E_{SA} (Eq. 2.67) is obtained: if the iterative scheme converges, the matrix elements H_{i0}^{SA} are zero and, consequently, the derivatives of E_{SA} with respect to the parameter p_i are zero as they equal H_{i0}^{SA}.

In summary, one iteration of excited state optimization consists of the following steps: *i*) Sample the quantities needed for the optimization of the linear coefficients with the appropriate guiding wave function; *ii*) diagonalize the matrix (Eq. 2.61) to obtain the optimal linear coefficients for the states under consideration; *iii*) sample for all states the quantities needed in the linear equations (Eq. 2.69) and obtain the parameters Δp_i; *iv*) construct a set of improved orbitals and Jastrow parameters as $p_i \rightarrow p_i + \Delta p_i$. As in the optimization of a ground state wave function, when the non-linear parameters are far from the optimal values, the optimization may need to be stabilized by shifting all diagonal elements except the first one as $H_{ij}^{\mathrm{SA}} \rightarrow H_{ij}^{\mathrm{SA}} + a_{\mathrm{adiag}}\delta_{ij}(1-\delta_{i0})$.

The use of pseudopotentials

While the QMC methods can be extended to large systems containing many electrons, the computational effort increases dramatically with the atomic number Z as the scaling is approximately Z^6, rendering all-electron calculations quickly intractable. The problem is caused by the core electrons which yield large energies and large fluctuations of the energy. The most common way to overcome this difficulty is to replace the core electrons by pseudopotentials, an approximation which is usually rather good as the core is chemically inert. An electron-nucleus pseudopotential is usually non-local and the most commonly used form is a potential which is local in the radial coordinate and non-local in the angular part as

$$\langle \mathbf{r}|v^{\mathrm{NL}}|\mathbf{r}'\rangle = \sum_{l=0}^{l_{max}} v^l(r)\,\delta(r-r') \sum_{m=-l}^{l} \mathrm{Y}_{lm}(\Omega)\mathrm{Y}_{lm}^*(\Omega')\,, \tag{2.71}$$

where l_{max} the maximum angular momentum considered, and the function v^l is radial and vanishes outside a core radius r_c. The non-local potential acting on the trial wave functions gives

$$\begin{aligned} &\langle \mathbf{R}|\mathcal{V}^{\mathrm{NL}}|\Psi\rangle \\ = &\sum_{i=1}^{N}\sum_{l=0}^{l_{max}} v^l(r_i) \sum_{m=-l}^{l} \mathrm{Y}_{lm}(\Omega_i) \int \mathrm{d}\Omega_i' \, \mathrm{Y}_{lm}^*(\Omega_i')\, \Psi(\mathbf{r}_1,\ldots,\mathbf{r}_i',\ldots,\mathbf{r}_N)\,, \end{aligned} \tag{2.72}$$

where the integral is over a sphere of radius $r_i' = r_i$ centered on the pseudoatom. This angular integration poses no particular problem in a VMC calculation and is done by a numerical quadrature on a regular polyhedron defined by a set of vertices whose number will depend on the value of l_{max} [39].

The use of nonlocal pseudopotential in DMC is more problematic since the Green's function (Eq. 2.53) is no longer positive. A possible way to circumvent this problem is to introduce the so-called *locality approximation* and define a new effective core potential by localizing the non-local potential on the trial wave function [40] as

$$\mathcal{V}_{\text{eff}}(\mathbf{R}) = \frac{1}{\Psi(\mathbf{R})}\langle \mathbf{R}|\mathcal{V}^{\text{NL}}|\Psi\rangle \,. \tag{2.73}$$

This new effective potential is explicitly many-body but is local and can be easily incorporated in a DMC algorithm. However, the potential depends now on the trial wave function, and the DMC energy computed with the mixed estimator is no longer necessarily variational and depends on the quality of the trial wave function. As the trial wave function approaches the fixed-node solution obtained without the locality approximation, the DMC energy converges to the correct fixed-node energy quadratically fast in the error on the trial wave function. A different approach to handle non-local pseudopotential which is variational and improves the accuracy upon the DMC approach with the locality approximation was recently proposed in Ref. [41].

Quantum Monte Carlo and excited states

Even though the DMC method was designed to project the ground state of a given Hamiltonian, it can also be used to study excited states. A straightforward way of obtain information on excited states is to construct a trial wave function which is a good approximation of the wave function of the state of interest. Naturally, this wave function can be used within VMC and, if the state is the lowest state of a one-dimensional representation of the point group of the molecules, DMC will yield a solution which is variational.

If the excited state is not the lowest state in its symmetry, we can still use DMC in the fixed-node approximation as the nodal constraints will prevent the collapse to the ground state solution. However, we may not be variational. It is only guaranteed that, if the nodal surface of the trial wave function is the same as the one of an exact state, fixed-node DMC yields the exact solution. Therefore, the role of the trial wave function is even more important when studying excited states within DMC since it not only imposes fermionic antisymmetry but also selects the state of interest.

2.3 Quantum mechanics in molecular mechanics techniques

Photochemical processes like absorption or fluorescence in biological systems are quite difficult to study because light-induced transitions require a quantum mechanical treatment of the system. However, as a biological system is very large and comprises thousands of atoms, the whole system cannot be treated at quantum level with the theoretical approaches available today. Fortunately, the primary process of photoabsorption often involves only few residues of the protein and is localized in a small spatial region, while the rest of the protein acts in this phase as spectator. Therefore, one can study the system by partitioning it into a quantum subsystem which is chemically active and small enough to be computationally treatable, and a larger environment which is assumed to be chemically inert and is simulated by less expensive classical molecular mechanics methods. These hybrid approaches are called quantum mechanics in molecular mechanics (QM/MM) approaches. In Fig. 2.1, we show how the QM/MM partitioning is applied in the simulation of the Green Fluorescent Protein.

The classical molecular mechanics calculations rely on empirical force fields to approximately describe the interactions of the classical particles. The most general form of these inter-particle potentials takes into account inter-molecular dispersion interactions between neutral atoms and electrostatic interactions between charged atoms, and intra-molecular interactions to describe the rotational, vibrational and torsional degrees of freedom of the molecule. To define a force field, one needs to specify the form of the potential and the values of the parameters whose number is very large as the same atom type may have different properties depending on which atom it is bonded to. As bonds cannot be created or broken in such a model, the bonding structure must be established at the beginning of the simulation.

Different QM/MM techniques are available in the literature, which differ for instance in how the interface between the MM and QM parts is treated and or in how the protein environment is simulated. In our work, we adopt the approach by Röthlisberger and coworkers [42] as implemented in the code CPMD [46]. The non-bonded interactions between the MM and the QM parts are modelled as

$$H_{\mathrm{NB}} = \sum_{I \in MM} q_I \int \mathrm{d}\mathbf{r}\, \frac{\rho(\mathbf{r})}{|\mathbf{r} - \mathbf{r}_I|} + \sum_{I \in MM, J \in QM} v_{\mathrm{vdW}}(r_{IJ}) \tag{2.74}$$

where $\rho(\mathbf{r})$ is the density of the electrons and the nuclei of the QM system, q_I are the MM partial charges at positions $\mathbf{r}_I$, and the van der Waals interactions

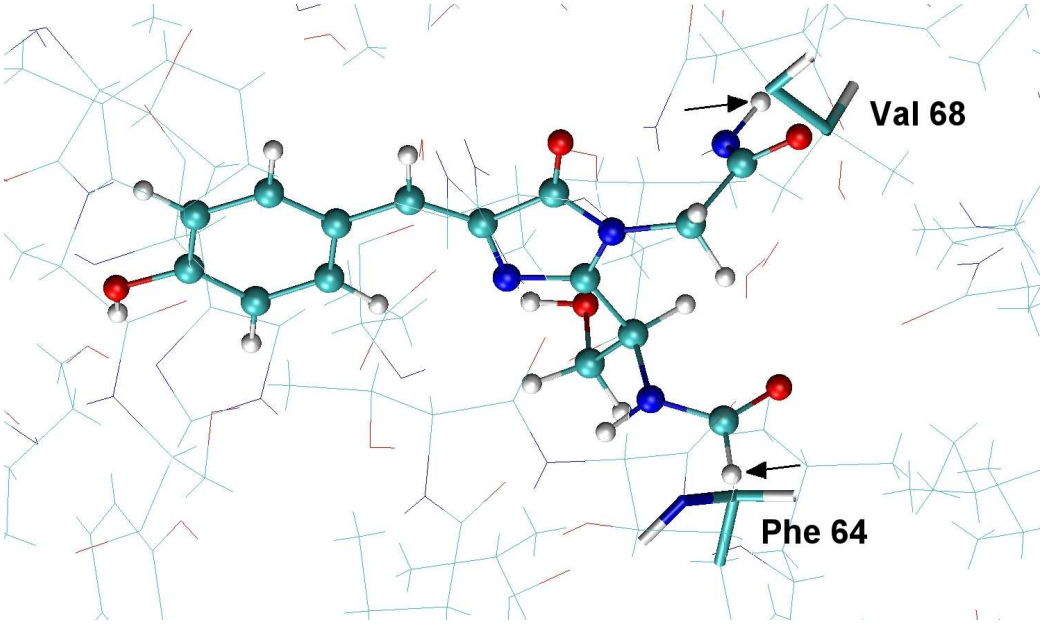

Figure 2.1: QM/MM partitioning in the simulation of Green Fluorescent Protein. The optically active part of the system is treated with a QM method while the rest of the protein is the MM part. The chromophore model is showed with a ball and sticks representation, while the MM part with solid lines. The two arrows indicate the two hydrogen-link atom between the QM and the MM. The MM part closest to the QM atoms is represented with cylinders.

between QM and MM atoms are described by the classical force field v_{vdW}.

The use of this expression is however problematic if the QM system is close to some positively charged MM atoms since the electrons will be attracted by the MM charges and the density overpolarized. This so-called spill-out effect is non physical and is particularly severe when a plane-wave basis is used as in the CPMD code. To avoid this problem in the CPMD implementation of QM/MM, a screening term is introduced for the point charges which are in proximity of the QM system. The electrostatic interaction of electrons with the close MM atoms (NN) in the non-bonded Hamiltonian is rewritten as

$$H_{\mathrm{NB}}^{\mathrm{el}} = \sum_{I \in \mathrm{NN}} q_I \int \mathrm{d}\mathbf{r}\, \rho^{\mathrm{el}}(\mathbf{r}) v_I(|\mathbf{r} - \mathbf{r}_I|) \tag{2.75}$$

where

$$v_I(r) = \frac{r_{cI}^n - r^n}{r_{cI}^{n+1} - r^{n+1}} \tag{2.76}$$

with r_{cI} the covalent radius of the atom type I and n=4. We refer the reader to Ref. [42] for a thorough discussion of the technical tricks used to reduce the large computational effort involved in computing the first term of Eq. 2.74 within plane-waves.

Finally, it is important to specify how to treat bonded interactions between the QM and MM parts, which arise when the QM/MM boundary cuts through a chemical bond of a molecule. In this case, we adopt the solution to "cap" the broken bond with a link atom, in particular a hydrogen atom. The QM system sees the hydrogen atom while the MM atoms do not.

2.4 Computational details

We employ different codes for the various steps of the calculations. For completeness, we conclude this chapter with a brief overview of the programs used, specifying in which step they were employed and briefly describing the technical details.

The **Amber Molecular Dynamics Package** [43] is a set of computational tools widely used to simulate bio-molecular systems. In particular, the module *Xleap* is used to add the missing hydrogens to the starting X-ray structure, to center the protein in a box of water solution (option *solvatebox*), to add the counterions to the water (option *addIons*) and finally to construct starting from the coordinates of the system the *Amber force field* which is then used in the Amber module *Sander* to perform classic molecular dynamics simulations. the Amber package to equilibrate the protein before starting the QM/MM calculations. The Amber force field 2003 is used to parameterize the MM part [44, 45]. In all MM simulations, a cutoff of 8 Å is used for the non-bonded van der Walls interactions.

The **CPMD** [46] code is used to perform the QM/MM calculation within density functional theory to describe the QM system while the Gromos [47] force field is used for the MM part. The QM/MM interface was developed by Röthlisberger and coworkers [48, 49], and, since it makes use of the MM Gromos libraries [47], it is necessary to convert the force field from the Amber to the Gromos format. The CPMD code is a plane-wave/pseudopotential code particularly designed for ab-initio molecular dynamics. The finite QM part is treated within a supercell approach using a sufficiently large periodic cell to avoid interactions between neighboring images. Moreover, we employ the isolated system module in CPMD which allows studying an isolated molecule or complex within periodic boundary conditions. We perform all calculations using a 70 Ry plane-wave cutoff and the Troullier-Martins pseudopotentials [50]. The Poisson equations are solved with the Tucker-

man's method [51]. All ground state calculations are carried out using the PBE [52] functional.

The **Gaussian03** [53] code is a quantum-chemistry code we use to perform ground state DFT and linear-response TDDFT calculations. It uses Gaussian basis sets, and a wide range of exchange-correlation functionals is available. We employ this code to perform geometry optimization of several molecular models of the chromophore in vacuum and to compute their TDDFT spectra. We also calculate the TDDFT spectra in the presence of the protein environment using the geometries obtained in the QM/MM approach with the CPMD code. The electrostatic effect of the protein environment is taken into account by using point charges with the coordinates obtained in the QM/MM calculation and the values of the charges as in the Amber force field. It is not possible to screen the charges with Gaussian03 but, as the basis is Gaussian, we do not expect significant spill-out problems as for a plane-wave basis.

The Amsterdam Density Functional **ADF** [54] code is a software package for first-principles electronic structure DFT calculations, using Slater functions for the construction of the orbitals. We employ this code to perform some TDDFT calculations with the SAOP [55] and LB94 [56] exchange-correlation potentials.

The **GAMESS** [57] code is a general ab-initio quantum chemistry package which we mostly use to generate the starting QMC wave functions either through a DFT or a SA-CASSCF calculation. This code employs Gaussian basis sets. The electrostatic effect of the protein environment can be introduced through the use of screened point charges described by the potential

$$v(r) = \frac{q}{r}\left(1 - A\, e^{-Br}\right) \tag{2.77}$$

where q is the value of the charge, and A and B are free parameters. We choose $A = 1$ to remove the divergence at the origin, and adjust the value of the B to reproduce the CPMD potential outside the maximum. We find that the choice $B = 1/(1.0828 \times r_c)^2$ where r_c is the CPMD core radius gives similar potentials in the valence region.

Finally, the code **CHAMP** is used for all the quantum Monte Carlo calculations. It can perform VMC and DMC calculations, and optimize the wave function parameters by energy minimization.

Chapter 3

Chromophore in vacuum

3.1 Chromophore models of GFP

Before studying the neutral (protonated) and anionic (deprotonated) forms of GFP within the protein environment, it is important to understand the performance of our theoretical techniques on simpler models. We therefore begin our investigation of GFP with the computation of the electronic excitations of model chromophores in the gas phase since the lower level of complexity of these systems allow us to push the limits of our theoretical tools and better understand their limitations. Moreover, absorption experiments and calculations using highly-correlated techniques are available for several chromophore models in the gas phase [58–60].

The chromophore models studied in this Chapter are depicted in Fig. 3.1, and can be divided in three groups: The anionic (deprotonated) chromophores (A, B, C), the neutral (protonated) chromophores (D, E, F), and a positively charged chromophore (G). For both the anionic and the neutral case, three models of increasing size are constructed. The geometries of all models are optimized within all-electron DFT with a cc-pVTZ basis and two different functionals, BLYP and B3LYP, using the Gaussian03 code [53]. We always refer to Fig. 3.1 and the its labels when describing the models below.

Anionic models

The anionic *minimal* model (A) is the smallest possible representation of the GFP chromophore with the additional nice feature to posses C_s symmetry, which significantly accelerates most quantum chemical computations. Given its favorable size and symmetry, it is a simple starting point to investigate the performance of quantum Monte Carlo in describing GFP. Even though no experimental characterization has been done on this system, correlated

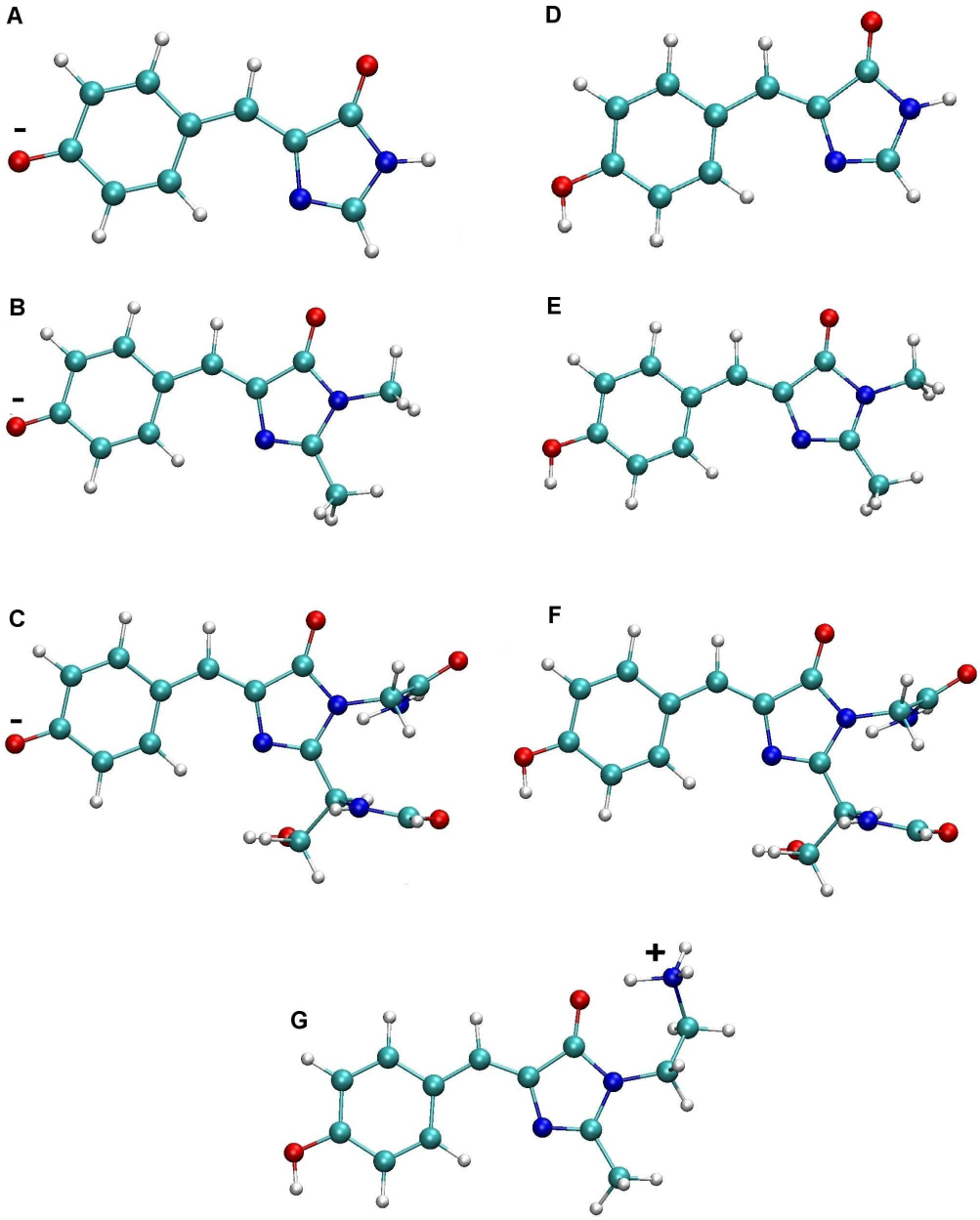

Figure 3.1: Gas-phase chromophore models: The three anionic chromophores in the minimal (A), methyl-terminated (B), and protein-cut (C) models; the three neutral chromophore in the minimal (D), methyl-terminated (E), and protein-cut (F) models; the positively charged (neutral$^+$) model (G).

CASPT2 calculations [60] are available to which we can compare.

The anionic methyl-terminated model (B) is only slightly larger than the minimal model (A). Since it only differs in the termination with the two hydrogens substituted with methyl groups, we expect that the electronic properties of models (A) and (B) are rather similar. Even though model (B) has the disadvantage that C_s symmetry is lost and the system has now no symmetry, we construct this chromophore since it was synthesized and investigated in the gas-phase spectroscopic experiments described above. Therefore, for model (B), we can directly compare the calculated excitation energies with experiments, which place its absorption maximum at 2.59 eV [58].

The largest anionic chromophore (C) is the *protein-cut* model with a total of 39 (24 C, N and O) atoms. It corresponds to the chromophore we employ as the QM part of the QM/MM simulations when the protein environment is included for a realistic study of wild-type GFP. The geometry of model (C) is optimized in vacuum and a comparison of the structural and electronic properties of this model with the one in the protein will allow us to understand the geometrical and electrostatic changes induced by the protein environment.

Neutral and neutral$^+$ models

For the neutral chromophore, we also construct three models of increasing size, which are analogous to the anionic case, that is the minimal (D), methyl-terminated (E) and protein-cut (F) models. To date, no highly-correlated quantum chemical calculations nor experiments are available for these neutral chromophores in the gas phase. Nevertheless, the study of the neutral chromophores allow us to analyze the changes in electronic properties with respect to the anionic gas-phase chromophores as well as as a function of system size. Moreover, we can investigate the geometrical and electronic impact of the protein environment on the large model (F).

The experimental technique employed for the study of the anionic gas-phase chromophore (B) makes use of an electrostatic ion storage ring and can therefore be applied only to negatively or positively charged molecules. Therefore, not being able to study neutral chromophores, the same experimental group synthesized a positively charged chromophore (G) by attaching to the methyl-terminated model (E) a positively charged NH_3^+ group [59]. We refer to model (G) as the neutral$^+$ chromophore. The relevance of this chromophore lies in the claim by the same experimental group that the positively charged group is just a spectator, and that the absorption spectrum of chromophore (G) should therefore be very close to the one of the neutral chromophore. In the photodestruction spectroscopy experiment, the absorp-

tion maximum of model (G) is identified at 2.99 eV and an excitation of 3.11 eV is attributed to the neutral chromophore by correcting the experimental value by the TDDFT/B3LYP difference between the vertical excitation of the (G) and the (E) models. We note that, in the neutral$^+$ chromophore, the positively charged group is hydrogen bonded to the oxygen of the imidazole ring.

3.2 Structural analysis of the models

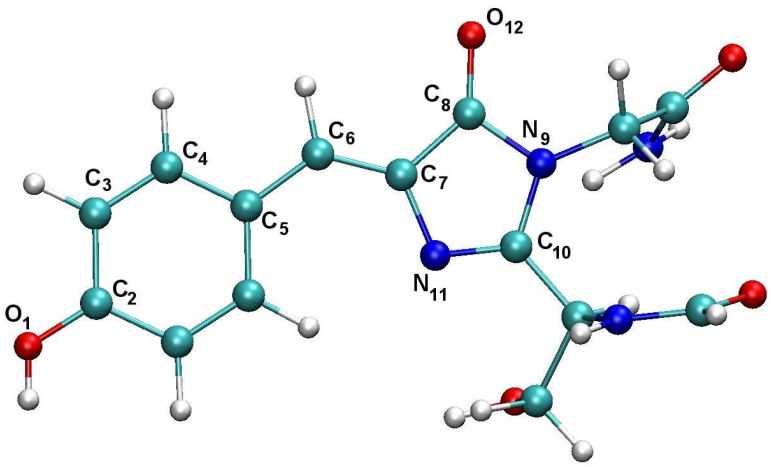

Figure 3.2: Atom numbering used for the chromophore of GFP.The benzene ring is essentially symmetric respect to the C_2-C_5 axis.

The structural features of the chromophore play a fundamental role in determining its excited state properties, and it is therefore important to accurately describe the geometry and how this is affected by the charge state of the chromophore and further modified when the chromophore is embedded in the protein environment. In particular, the degree of bond-length alternation in the conjugate chain running through the chromophore is correlated to the size of the bright $\pi \rightarrow \pi^*$ transition, with a stronger bond-length alternation yielding a larger excitation energy. For π-conjugate linear chains, this correspondence between gap and bond-length alternation is well known as, in going from shorter to longer chains, the energy gap decreases while the double bonds lengthen and the single bonds shorten. The GFP chromophore

is also a π-conjugated system even though the presence of the rings makes it a more complicated system than a linear oligomer.

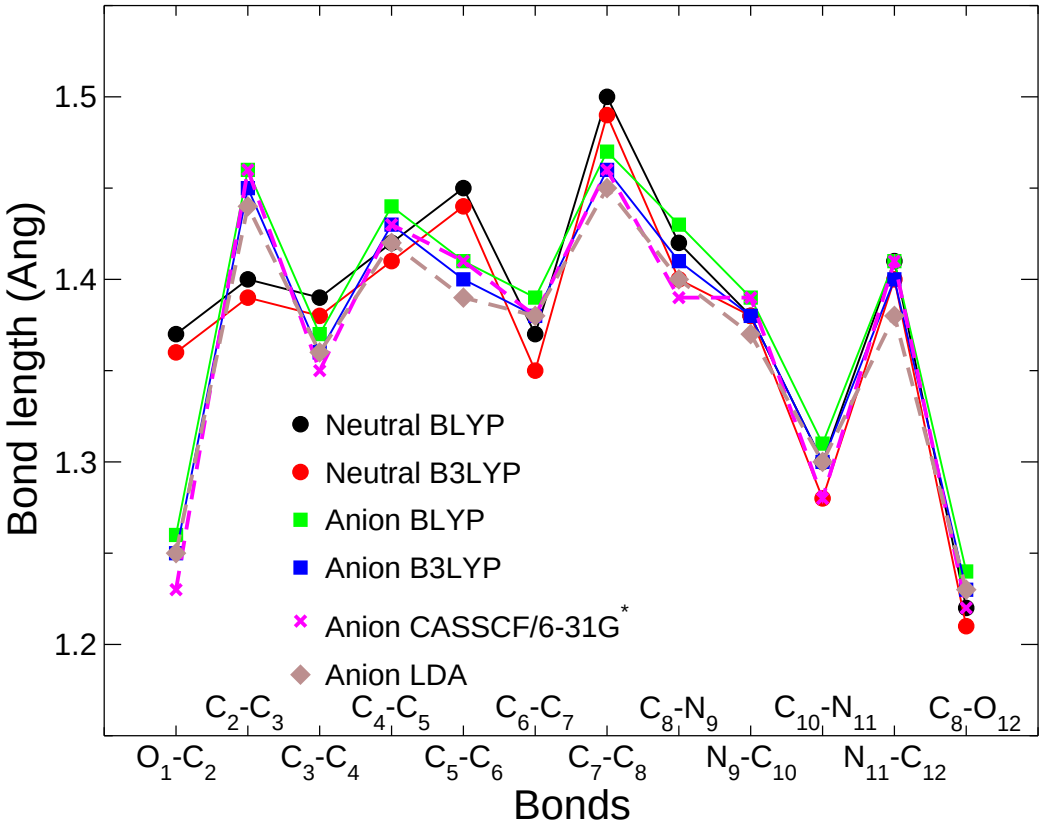

Figure 3.3: Bond lengths of the minimal anionic (A) and neutral (D) models of the GFP chromophore as obtained in a DFT/cc-pVTZ geometry optimization with the BLYP, B3LYP and LDA functionals. The results from a CASSCF/6-31G* calculation [60] are also shown. The bonds of the central carbon bridge are C_5—C_6 and C_6—C_7.

The structures of the chromophores are constructed by relaxing them in the ground state using DFT and various exchange-correlation functionals. Even though all models have been generated with at least the BLYP and B3LYP functionals, we only discuss in detail how the functional influences the geometry of the chromophore for the minimal anionic (A) and neutral (D) models as the conclusions apply equally well to all other models. The bond lengths along the chromophore of the minimal models computed with various functionals are shown in Fig. 3.3. For the atom labelling, we refer the reader to Fig. 3.2, where the heavy atoms of the chromophore are numbered starting from the phenolic oxygen along the top ridge of the phenol, through

the bridge and around the imidazole ring.

For both the anionic (A) and the neutral (D) minimal model, the geometries obtained with the BLYP and the B3LYP functional are very similar, with the use of B3LYP shortening all bonds by a very small amount (about 0.01 Å). The structural parameters obtained using the PBE functional which we employ to generate the QM/MM protein models of GFP are not shown as they are practically indistinguishable from the BLYP bond lengths with an average agreement of 0.001 Å. For the anionic model (A), we also compute the bond lengths with the LDA functional as the local density instead of a generalized gradient approximation functional is used in the QM/MM calculations by Marques *et al.* [12] which we extensively discuss in Chapter 4. As expected, we find that LDA overbinds with respect to B3LYP but the difference is not very large with an average deviation of 0.01 Å.

In summary, all exchange-correlation DFT functionals yield equivalent geometries for both the minimal anionic (A) and neutral (D) chromophores, and a similar finding holds for all other models depicted in Fig. 3.1. No reference structure exists for the neutral chromophores in the gas phase but, for the minimal anionic (A) model, we can compare our DFT geometries to the correlated CASSCF/6-31G* calculations by Martin *et al.* [60]. As shown in Fig. 3.3, we find good agreement between the CASSCF and the DFT structures and, in particular, the B3LYP bond lengths are quite close to the CASSCF values with a maximum deviation of only 0.02 Å. Since there is no evidence that anionic molecules are better described by DFT than neutral ones, we can safely assume that the accuracy of DFT for all models of GFP is well within the spread of the different functionals we tested. Therefore, when we compare excitation spectra for gas-phase models, the eventual differences one observes should be attributed to the theoretical approaches employed to compute the excitation energy rather then to the approach followed to optimize the geometry.

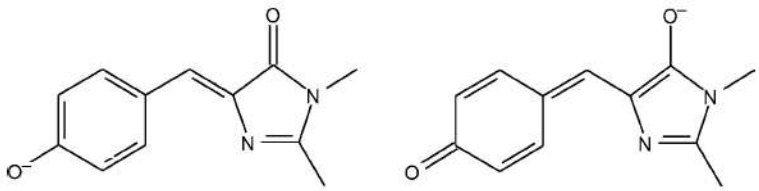

Figure 3.4: Scheme of the two resonant forms of the anionic chromophore: Benzenoid (left) and quinonoid (right).

While the main structural features are largely independent on the functional, the most evident difference is between the bond lengths of the neutral and the anionic model. To better understand the geometrical changes with the charge state of the chromophore, we show the two resonant forms of the anionic chromophore in Fig. 3.4. In the benzenoid form, the negative charge is localized on the phenolic oxygen and this bond structure is therefore also characteristic of the neutral chromophore. Upon deprotonation, the quinonoid form is also accessible where the negative charge has migrated to the imidazole oxygen. The change in bond length alternation and its reduction in the anionic chromophore is a measure of the mixing between the two forms. As we will see in Chapter 4, the protein environment can further tune the mixing by driving the resonance towards the benzenoid or the quinonoid form and therefore changing the bond structure of the chromophore.

The structural changes of Fig. 3.3 between the minimal neutral (D) to the anionic (A) model can now be more easily understood in terms of the two resonant forms. The neutral model is characterized by the aromaticity of the phenolic ring (with rather similar bond lengths between all carbon atoms of the ring) and by a double-single bond-length alternation at the carbon bridge given by the bonds C_5—C_6 and C_6═C_7. In the anionic model, the hydroxyl group is deprotonated and the oxygen-carbon bond, O_1—C_2, shortens by about 0.1 Å as compared to the neutral model, loosing its single-bond character. As a consequence, the aromaticity of the phenolic ring is reduced, yielding a quinoid structure of the ring (double bonds between C_2—C_3 and the opposite carbon bond). The degree of bond alternation in the central carbon bridge is also decreased in the anionic chromophore, yielding a significantly stronger bond C_5—C_6 than in the neutral model: The two central bonds, C_5—C_6 and C_6═C_7 differ by about 0.08 Å in the neutral model and only 0.02 Å in the anionic chromophore. Beyond the central bridge, the deprotonation of the phenolic oxygen does not have as large an effect, and the two most significant changes are a reduction by 0.03 Å of the single bond C_7—C_8 and a lengthening by 0.02 Å of the C_8—O_{12} bond.

Finally, we note that the main structural changes between the anionic and neutral chromophores we describe for the minimal (A) and (D) models are also present when the larger models of the chromophore are considered. In Table 3.1, we list the main bond lengths optimized within DFT/BLYP for the anionic and neutral chromophores of the minimal (A), methyl-terminated (B) and protein-cut (C) models and the corresponding (D, E, F) neutral models. Both in the anionic and the neutral case, the bond lengths of the three models are practically identical, so the addition of the longer tails in the protein-cut chromophore has no effect on the relevant geometrical parameters of the chromophore.

	Anionic model			Neutral model		
	A	B	C	D	E	F
	Bond length (Å)					
O_1—C_2	1.26	1.26	1.26	1.37	1.38	1.37
C_2—C_3	1.46	1.46	1.47	1.40	1.40	1.41
C_3—C_4	1.37	1.37	1.37	1.39	1.39	1.38
C_4—C_5	1.44	1.44	1.44	1.42	1.42	1.43
C_5—C_6	1.41	1.41	1.41	1.45	1.45	1.45
C_6—C_7	1.39	1.39	1.40	1.37	1.37	1.38
C_7—C_8	1.47	1.46	1.46	1.50	1.49	1.50
C_8—N_9	1.43	1.43	1.44	1.42	1.42	1.43
N_9—C_{10}	1.39	1.40	1.40	1.38	1.40	1.39
C_{10}—N_{11}	1.31	1.31	1.32	1.30	1.31	1.33
N_{11}—C_7	1.41	1.41	1.41	1.41	1.41	1.41
C_8—O_{12}	1.24	1.24	1.24	1.22	1.23	1.23
	Dihedral angle (°)					
D($C_4C_5C_6C_7$)	180.0	180.0	178.3	180.0	180.0	179.5

Table 3.1: DFT/BLYP structural parameters of the minimal (A), methyl-terminated (B), and protein-cut (C) anionic models and the corresponding (D, E, F) neutral models. We list the most representative bond lengths and one dihedral angle. See Fig. 3.2 for the labeling of the atoms.

3.3 TDDFT excited states

We compute the low-lying TDDFT vertical excitations of the various model chromophores in the gas phase and locate the excitation with the largest oscillator strength which corresponds to the maximum light absorption. We compare these bright excitations to the available reference data, that is, to photodistruction spectroscopy experiments for the anionic methyl-terminated (B) [58] and the neutral$^+$ (G) model [59], and to theoretical CASPT2 calculations for the anionic minimal (A) model [60]. These gas-phase calculations will already give us a feeling of which performance we may expect from TDDFT when the chromophores are embedded in the protein environment.

The all-electron linear-response TDDFT calculations are performed with the BLYP and B3LYP functionals at the corresponding BLYP and B3LYP ground state structures. We use the Gaussian03 code [53] and a cc-pVTZ basis as for the ground state calculations. For all model chromophores, we report the lowest two singlet excitations with their oscillator strength and

character which allow us to distinguish bright from dark states, and identify the electronic transition involved in the excitation. We also give two quantities which we use as indicators of the reliability of the TDDFT excitations, that is, minus the Kohn-Sham energy of the highest occupied molecular orbitals (HOMO) which indicates the start of the DFT ionization continuum, and the Kohn-Sham gap between the lowest unoccupied (LUMO) and the highest occupied molecular orbitals.

Before presenting the TDDFT results, we briefly discuss what we should expect for the excitations of the model chromophores in the gas phase. First, the excitation energy should decrease when progressing to larger chromophore sizes in the series (A, B, C) and (D, E, F) for the anionic and neutral models, respectively. This can be easily understood based on the simple argument of how the level spacing for a particle in a box behaves as the box is made larger. Moreover, we would expect the excitations of the anionic chromophores to be lower than the ones of the corresponding neutral models due to a combination of geometrical and electronic considerations. As discussed in the previous Section, the degree of bond alternation in proximity of the central carbon bridge is reduced in the anionic as compared to the neutral models, so the excitation of the bright $\pi \rightarrow \pi^*$ will be reduced. Moreover, as the anionic state can be described as a mixing of the benzenoid or the quinonoid forms, we expect also the excitation to be more delocalized and therefore lower than the neutral case. We will now see whether our picture of the behavior of the excitations is met in practice in the TDDFT calculations.

Anionic models

The TDDFT results for anionic minimal (A), methyl-terminated (B), and protein-cut (C) models are summarized in Table 3.2. For all three models and both exchange-correlation functionals, the excitation energy with the largest oscillator strength has a dominant HOMO $\rightarrow$ LUMO character. As the highest occupied (HOMO) and lowest unoccupied (LUMO) molecular orbitals are rather similar for the three models and the two functionals, we only plot them for the BLYP functional and the protein-cut model in Fig. 3.5. The highest occupied orbital has π-bonding character on the two carbon bonds of the central bridge while the lowest unoccupied orbital is a π-antibonding orbital on both bonds, and some degree of charge transfer can be observed across the bridge from the phenolic to the imidazole ring.

As expected, the electronic properties of the minimal (A) and the methyl-terminated (B) models are rather similar. The addition of the methyl groups only lowers the BLYP and B3LYP excitation energies by 0.08 and 0.07 eV, respectively, while preserving the same character of the excitation and the

oscillator strength. In the protein-cut (C) model, the BLYP and B3LYP functionals behave instead rather differently as the BLYP excitation with the largest oscillator strength is no longer the lowest state. However, if we consider the bright excitation for both the BLYP and B3LYP functionals, we find that the absorption maximum is further lowered by 0.03 and 0.05 eV with respect to the smaller methyl-terminated (B) model. Therefore, as expected, the excitation energy decreases with increasing size of the model.

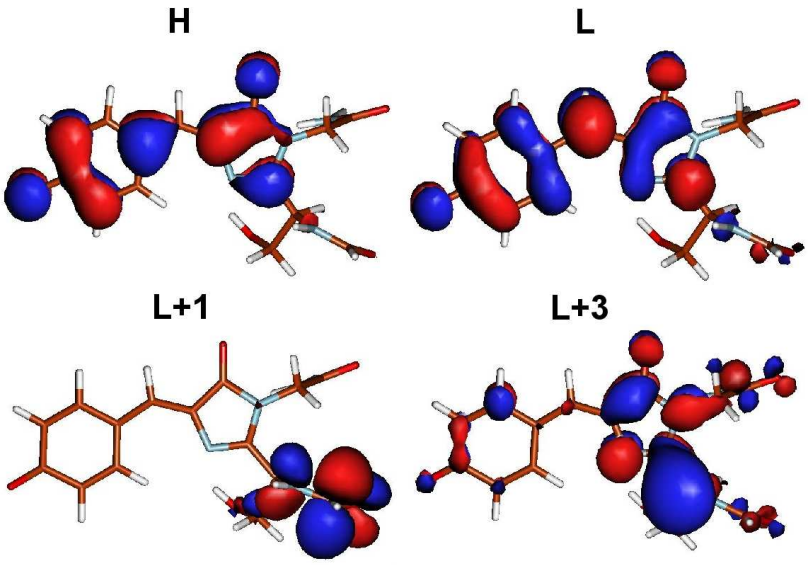

Figure 3.5: The DFT/BLYP orbitals for the anionic protein-cut (C) model chromophore in the gas phase. An isosurface of 0.025 is shown in red and an isosurface of -0.025 in blue. We only show the orbitals which are involved in the TDDFT/BLYP excitations.

We note that, as the model becomes larger, more orbitals are involved in the description of the BLYP excited states while the character of the lowest B3LYP excitation remains unchanged and predominantly HOMO $\rightarrow$ LUMO. In particular, for the protein-cut (C) model, the lowest BLYP excitation has HOMO $\rightarrow$ (LUMO+1) character while the excitation with the largest oscillator strength has non negligible contributions from transitions to the (LUMO+1) and (LUMO+3) orbitals, which are depicted in Fig. 3.5. While the (LUMO+3) orbital is largely localized on the imidazole ring, the (LUMO+1) orbital is confined in the far tail of the model. As we expect equivalent electronic properties for a protein-cut chromophore obtained by differently setting the QM/MM boundary to shorten this tail, it appears

rather unphysical that the (LUMO+1) orbital would play such a relevant role in the excited states of the chromophore.

We now compare the linear-response TDDFT/BLYP and B3LYP excitation energies with the available reference data. For the methyl-terminated (B) anionic model, experiments [58] place the absorption maximum at 2.59 eV, significantly lower than the excitations of 2.89 and 3.09 eV obtained with the BLYP and B3LYP functionals, which therefore overestimate the experiment by as much as 0.30 and 0.50 eV, respectively. Correlated CASPT2 calculations by Martin *et al.* [60] for the anionic minimal (A) model give a vertical excitation energy of 2.67 eV which is in very good agreement with experiment as we consider that the electronic properties of models (A) and (B) must be rather similar. For the minimal (A) model, BLYP and B3LYP yield instead a significantly higher excitation energy of 2.97 and 3.16 eV, respectively. We discuss below the possible reasons of the apparent failure of TDDFT in describing the excitation energies of the anionic chromophore in the gas phase.

Neutral and neutral$^+$ models

The TDDFT results for neutral minimal (D), methyl-terminated (E), and protein-cut (F) models are summarized in Table 3.3. Similarly to the anionic case, the B3LYP functional gives the lowest excitation as the one with the largest oscillator strength and HOMO $\rightarrow$ LUMO character for all three models. When the BLYP functional is employed, this is still true for the smaller models but not for the protein-cut (F) chromophore where the excitation with the largest oscillator strength is the second one, and has a large HOMO $\rightarrow$ LUMO component but a dominant (HOMO-2) $\rightarrow$ LUMO transition.

We plot the orbitals which are involved in the BLYP excitations for the protein-cut (F) model in Fig. 3.6. The highest occupied (HOMO) and lowest unoccupied (LUMO) molecular orbitals are rather similar for the three models and the two functionals. The HOMO orbital has π-anti-bonding/bonding character on the first/second bond of the central carbon bridge while the situtation is reversed in the LUMO where the sequence is instead π-bonding/anti-bonding. The (HOMO-2) orbital which has dominant weight in the two lowest BLYP excitations is localized in the upper tail of the chromophore. Similarly to case of the BLYP excitations of the anionic protein-cut model, it is rather unphysical that the (HOMO-2) $\rightarrow$ LUMO excitation is so relevant as we could generate a different chromophore with equivalent electronic properties but a shorter upper tail by placing a different QM/MM boundary. This is likely an indication that the adiabatic TDDFT/BLYP description of the largest model is encountering some problems.

If we focus on the excitations with the largest oscillator strength for both functionals, we see that the B3LYP excitation energies are always larger than the BLYP ones by 0.3 eV, and that the excitation correctly decreases as the size of the model increases as it was the case for the anionic chromophores. We also note that the excitations of the neutral chromophores are always larger than the ones of the corresponding anionic models by roughly 0.2-0.3 and 0.4 eV for the BLYP and the B3LYP functional, respectively.

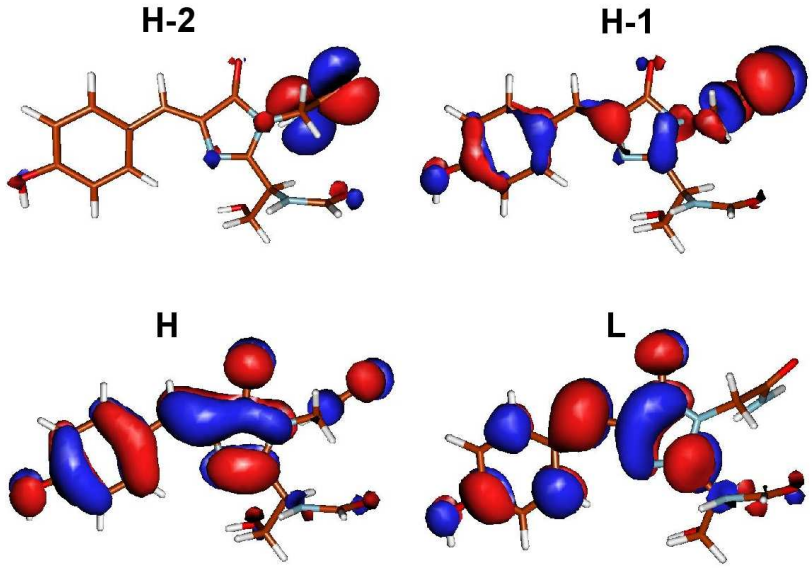

Figure 3.6: The DFT/BLYP orbitals for the neutral protein-cut (C) model chromophore in the gas phase. An isosurface of 0.025 is shown in red and an isosurface of -0.025 in blue. We only show the orbitals which are involved in the TDDFT/BLYP excitations.

In Table 3.4, we report the TDDFT excitations for the neutral$^+$ (G) model, which must be compared to the absorption maximum of 2.99 eV measured in photodistruction spectroscopy experiments [59]. To estimate the excitation of the methyl-terminated neutral (E) model, the same experimental group corrects the absorption maximum of the neutral$^+$ (G) chromophore by the difference between the theoretical TDDFT/B3LYP excitations of the (G) and (E) models optimized at the MP2 level, and obtains an excitation of 3.11 eV for the (E) model. We prefer to avoid this procedure as we obtain for instance different TDDFT corrections of 0.08 and 0.25 eV, when using the BLYP and the B3LYP functional with the corresponding DFT geometries. We instead focus on the direct comparison with the photodistruction

spectroscopy experiments for the neutral$^+$ (G) model and the available experimental data obtained with the same spectroscopy technique for the anionic methyl-terminated (B) model [58]. For the neutral$^+$ (G) model, we find that TDDFT/BLYP gives an excitation of 3.02 eV in very close agreement with experiments while the B3LYP excitation of 3.21 eV overestimates experiments by 0.22 eV. The experimental shift of 0.4 eV between the absorption maximum of the anionic methyl-terminated (B) model (2.59 eV) and the neutral$^+$ (G) model (2.99 eV) is not reproduced by the TDDFT calculations. The BLYP and B3LYP shifts are equal to 0.13 and 0.12 eV, respectively, and therefore significantly smaller than the experimental value.

3.3.1 Assessing the performance of TDDFT

Before discussing the mixed performance of TDDFT when compared to the reference data for the anionic (A and B) and the neutral$^+$ (G) models, we analyze some general features observed in the excitations of the gas-phase chromophores which will also be present when the chromophore is embedded in the protein environment. For the excited states of the smaller models, pure (BLYP) and hybrid (B3LYP) functionals yield excited states characterized by the same type of electronic transitions. For the larger models, BLYP mix several transitions in contrast to the hybrid B3LYP functional which preserves a dominant single-excitation character. For all excited states, B3LYP yields higher excitation energies than the corresponding pure functional.

That B3LYP yields higher TDDFT excitation energies than BLYP can be in part understood from the fact that TDDFT applies a correction to the single-particle Kohn-Sham excitations given by eigenvalue differences. The Kohn-Sham eigenvalues behave rather differently if the functional contains a fraction of exchange as in B3LYP. The occupied orbitals will be lower as the hybrid functional is partially self-interaction corrected and the exchange-correlation potential no longer decays exponentially. On the other hand, the unoccupied orbitals will be higher as the virtual orbitals see a different number of electrons than the occupied ones due to the presence of Hartree-Fock exchange in the functional. Consequently, the Kohn-Sham excitations will be higher for the B3LYP than for the BLYP functional, as can be seen by inspecting for instance the B3LYP and BLYP HOMO-LUMO gaps in Tables 3.2, 3.3 and 3.4.

As TDDFT corrects the Kohn-Sham eigenvalue differences, it is important to try to establish, if possible, when these corrections are meaningful. Due to the spatial locality of the approximate adiabatic exchange-correlation kernel f_{xc}, the TDDFT correction to the Kohn-Sham excitations would vanish for a pure functional if the dominant transition is between non-overlapping

occupied and virtual orbitals. Therefore, if the TDDFT/BLYP excitation reduces to the Kohn-Sham eigenvalue difference we are in the presence of an excitation characterized by charge transfer which is poorly described by adiabatic TDDFT. When a hybrid functional is employed, the kernel is non-local in space and the TDDFT correction may not necessarily vanish. Therefore, as most excited states have a dominant HOMO $\rightarrow$ LUMO transition, we always report in the Tables the HOMO-LUMO gap since a comparison of the TDDFT/BLYP excitation with the corresponding gap reveals if the TDDFT excitation has charge transfer character and is consequently unreliable.

Another important indicator of the reliability of the TDDFT excitations is minus the eigenvalue of the HOMO orbital ($-\epsilon_{\mathrm{HOMO}}$) as it locates the start of the TDDFT continuum and represents the ionization threshold. This threshold is usually underestimated in DFT as the Kohn-Sham potential for pure functionals decays exponentially and therefore too quickly. The use of hybrid functionals corrects to some degree this problem as they are partially self-interaction corrected. If the TDDFT excitation lies above the DFT ionization threshold, its value cannot in general be trusted.

We can now discuss the TDDFT performance for the anionic models. In Table 3.2, we observe that, for all models, the TDDFT/BLYP excitation with the largest oscillator strengths has HOMO $\rightarrow$ LUMO character and its energy lies well above the HOMO-LUMO gap. Therefore, this indicates that, for all models, the TDDFT/BLYP corrections to the Kohn-Sham excitations are not negligible and the brightest excited states are not charge-transfer excitations. For the protein-cut anion (C) model, the lowest BLYP excitation is predominantly a HOMO $\rightarrow$ (LUMO+1) transition and its energy of 2.65 eV closely agrees with the HOMO-(LUMO+1) gap of 2.68 eV. We had already noted the unphysical relevance of such a charge-transfer excitation as the (LUMO+1) orbital is localized in the far tails of the chromophore.

While charge transfer does not appear to pose a severe problem for the anionic chromophores, we note that the excitation energies of all models and for both functionals are significantly above the DFT ionization threshold. The use of B3LYP raises the ionization threshold as compared to BLYP but not sufficiently to bring it above the lowest excitation energy. As the chromophores are anionic, one may wonder whether the lowest excited state may actually be a quasi-stable excited state in the continuum which autoionizes after absorption, and therefore really lies above the ionization threshold. To investigate this point, we perform TDDFT calculations for the minimal (A) and the protein-cut (C) anionic models with the LB94 potential and the statistical average of orbital potential (SAOP) approach, which both yield the correct Coulombic tail ($-1/r$) in the exchange-correlation potential.

The TDDFT/LB94 and SAOP results obtained with the ADF code and

a ETpVQZ basis are reported in Table 3.5 for the ground state DFT/BLYP geometries. The use of an asymptotically correct exchange-correlation potential raises the ionization threshold well above the relevant excitation energy. Therefore, the BLYP/B3LYP excited states are not quasi-bound states in the continuum but the BLYP/B3LYP ionization threshold is simply underestimated. The TDDFT/LB94 and SAOP excitation energies for the minimal (A) model differ only by -0.08 and 0.02 eV from the BLYP excitation, respectively. For the protein-cut (C) anionic chromophore, the TDDFT/SAOP excitations are very similar to the BYLP ones with the lowest excitation having HOMO → (LUMO+1) character and the (LUMO+1) orbital being localized in the tails of the chromophore. The second SAOP excitation with the largest oscillator strength is only 0.07 eV higher than the brightest BLYP excitation. Therefore, even though the use of an asymptotically corrected potential has removed the issue of the underestimation of the DFT ionization threshold, the picture remains practically unchanged with respect to the use of the BLYP functional.

For the excitations of the neutral chromophores of Table 3.3, the underestimation of the DFT ionization threshold does not pose a problem as, for all models and both functionals, the DFT ionization threshold is above the excitation with the largest oscillator strength. As far as the charge-transfer character of the excitations, the behavior of the neutral minimal (D) and methyl-terminated (E) models is very similar to their anionic counterparts: The TDDFT/BLYP excitations have predominantly HOMO → LUMO character and the energy is significantly higher than the HOMO-LUMO gap. On the other hand, for the protein-cut (F) model, the lowest excitation has (HOMO-2) → LUMO character and the BLYP excitation energy of 2.97 eV is very close to the (HOMO-2)-LUMO gap of 2.99 eV indicating the charge-transfer character of the excitation. For this model, the second excitation has the largest oscillator strength but its energy is rather close to the lowest state and the predominant transition is still (HOMO-2) → LUMO with the (HOMO-2) orbital being localized on one the tails of the chromophores. This may point at some difficulties of adiabatic TDDFT in describing the excitations of the larger neutral chromophore.

In summary, it is not clear why TDDFT/BLYP and B3LYP should perform poorly in the description of the smaller anionic minimal (A) and methyl-terminated (B) models as compared to experimental and CASPT2 reference data. The excitations of these smaller models do not appear to be characterized by charge transfer and curing the underestimation of the ionization threshold with the use of asymptotically corrected functionals leaves the excitation energy practically unvaried. Therefore, it is not evident why TDDFT should be superior in describing the excitations of the corresponding neutral

(D) and (F) models or the neutral$^+$ (F) chromophore. We note that, for the larger protein-cut models, the presence of significant contributions from charge-transfer transitions also in the excitation with the largest oscillator strength raises some doubts about the reliability of the TDDFT excitations, in particular for the neutral protein-cut (F) model.

Table 3.2: TDDFT/BLYP and B3LYP excitation energies (eV) and oscillator strengths (in parenthesis) for the minimal (A), methyl-terminated (B), and protein-cut (C) anionic models of the GFP chromophore, computed using a cc-pVTZ basis. DFT/BLYP and B3LYP geometries are consistently used. The dominant electronic transitions and their contributions in parenthesis (if > 0.1) are also listed. The lowest two singlet excitations are given together with minus the Kohn-Sham energy of the highest occupied molecular orbitals ($-\epsilon_{\mathrm{HOMO}}$) and the Kohn-Sham gap between the lowest unoccupied and the highest occupied molecular orbitals ($\Delta\epsilon_{\mathrm{HL}}$). The experimental absorption maximum for the methyl-terminated (B) chromophore is 2.59 eV [58] while CASPT2 calculations for the minimal (A) model give a vertical excitation of 2.67 eV [60].

Model	minimal	methyl-terminated	protein-cut
		BLYP functional	
$S_0 \rightarrow S_1$	2.97(0.75)	2.89(0.77)	2.65(0.17)
	H→L(0.54)	H→L(0.54)	H→L+1(0.62)
		H→L+2(0.13)	H→L(0.29)
		H→L+3(0.11)	
$S_0 \rightarrow S_2$	3.25(0.00)	3.60(<0.01)	2.86(0.62)
	H-2→L(0.70)	H-2→L(0.59)	H→L(0.45)
		H→L+5(0.18)	H→L+1(0.32)
		H→L+3(0.17)	H→L+3(0.18)
$\Delta\epsilon_{\mathrm{HL}}$	1.82	1.79	1.77
$-\epsilon_{\mathrm{HOMO}}$	0.59	0.53	1.29
		B3LYP functional	
$S_0 \rightarrow S_1$	3.16(0.88)	3.09(0.92)	3.04(0.96)
	H→L(0.57)	H→L(0.58)	H→L(0.58)
$S_0 \rightarrow S_2$	4.01(0.00)	4.24(<0.01)	4.02(0.02)
	H-3→L(0.70)	H-2→L(0.61)	H→L+2(0.70)
		H→L+6(0.14)	
		H→L+3(0.14)	
$\Delta\epsilon_{\mathrm{HL}}$	2.96	2.93	2.91
$-\epsilon_{\mathrm{HOMO}}$	1.29	1.22	1.98

Table 3.3: TDDFT/BLYP and B3LYP excitation energies (eV) and oscillator strengths (in parenthesis) for the minimal (D), methyl-terminated (E), and protein-cut (F) neutral models of the GFP chromophore, computed using a cc-pVTZ basis. DFT/BLYP and B3LYP geometries are consistently used. See caption in Table 3.2 for further explanations.

Model	minimal	methyl-terminated	protein-cut
	BLYP functional		
$S_0 \rightarrow S_1$	3.22(0.59)	3.10(0.52)	2.97(0.20)
	H→L(0.57)	H→L(0.57)	H-2→L(0.51)
	H-5→L(0.13)	H-5→L(0.14)	H→L(0.36)
		H-2→L(0.13)	H-1→L(0.22)
$S_0 \rightarrow S_2$	3.66(0.01)	3.56(0.13)	3.10(0.34)
	H-2→L(0.57)	H-2→L(0.63)	H-2→L(0.44)
	H-3→L(0.27)	H→L+2(0.19)	H→L(0.39)
	H→L+1(0.19)		H-4→L(0.22)
	H→L+2(0.18)		
$\Delta\epsilon_{\mathrm{HL}}$	2.23	2.19	2.13
$-\epsilon_{\mathrm{HOMO}}$	4.98	4.76	4.87
	B3LYP functional		
$S_0 \rightarrow S_1$	3.54(0.68)	3.46(0.66)	3.42(0.71)
	H→L(0.61)	H→L(0.60)	H→L(0.60)
$S_0 \rightarrow S_2$	3.56(<0.01)	3.66(<0.01)	3.58(<0.01)
	H-1→L(0.69)	H-1→L(0.69)	H-1→L(0.66)
			H-3→L(0.21)
$\Delta\epsilon_{\mathrm{HL}}$	3.51	3.55	3.49
$-\epsilon_{\mathrm{HOMO}}$	5.85	5.62	5.78

Table 3.4: TDDFT/BLYP and B3LYP excitation energies (eV) and oscillator strengths (in parenthesis) for the neutral$^+$ (G) model of the GFP chromophore, computed using a cc-pVTZ basis. DFT/B3LYP ground state geometries are always used. See caption in Table 3.2 for further explanations.

	BLYP	B3LYP
$S_0 \rightarrow S_1$	3.02(0.73)	3.21(0.86)
	H→L(0.56)	H→L(0.60)
	H-1→L(0.10)	
$S_0 \rightarrow S_2$	3.26(<0.01)	3.73(0.02)
	H-3→L(0.50)	H-1→L(0.67)
	H-1→L(0.46)	H→L+2(0.13)
	H-2→L(0.11)	
$\Delta\epsilon_{\mathrm{HL}}$	2.01	3.19
$-\epsilon_{\mathrm{HOMO}}$	7.80	8.63

Table 3.5: TDDFT/LB94 and SAOP excitation energies (eV) and oscillator strengths (in parenthesis) for the anionic minimal (A) and protein-cut (C) models computed using a ET-pVQZ basis and the BLYP ground state geometries. See caption in Table 3.2 for further explanations.

	minimal		protein-cut
	LB94	SAOP	SAOP
$S_0 \rightarrow S_1$	2.89(0.75)	2.99(0.79)	2.75(0.29)
	H→L(0.95)	H→L(0.95)	H→L+1(0.67)
			H→L(0.32)
$S_0 \rightarrow S_2$			2.93(0.62)
			H→L(0.60)
			H→L+1(0.32)
$\Delta\epsilon_{\mathrm{HL}}$	1.73	1.86	1.83
$-\epsilon_{\mathrm{HOMO}}$	6.31	5.10	5.85

3.4 QMC excitation energies

The use of QMC methods in combination with a careful construction of the many-body trial wave function has proven successful in describing the excitations of small prototypical photoactive molecules in the gas phase [61–64]. Here, we begin our exploration of the applicability of quantum Monte Carlo to the study of excitations in realistic biosystems, by computing the QMC vertical excitations of the model chromophores of GFP in the gas phase. For the smaller anionic minimal (A) model, we will also perform a thourough investigation of the possible limitations of QMC calculations for excited states. In particular, the dependence of the excitation energy on the trial wave function will be analyzed by optimizing several thousands parameters in excited-state trial wave functions of increasing complexity, using a robust and efficient optimization method we have recently developed [65].

For the QMC calculations in the gas phase, we focus on the chromophores which are relevant when comparing to reference experimental or CASPT2 data, that is, the anionic minimal (A) and methyl-terminated (B), the neutral minimal (D), and the neutral$^+$ (G) models. We employ the ground state geometries optimized within DFT/B3LYP for all models expect for the anionic minimal (A) model where, for historical reasons, we have used the DFT/BLYP ground state geometry. In quantum Monte Carlo, we compute the vertical $\pi \rightarrow \pi^*$ excitations as the difference between the total energies of the ground and excited states, which we obtain in a three step procedure. First, a conventional state-average (SA) complete active space (CAS) self-consistent field (SCF) calculation is performed. The resultant SA-CASSCF wave functions are then multiplied by a Jastrow factor to include dynamical correlation, and partially or fully reoptimized for the ground and excited states. Finally, the Jastrow-Slater wave functions are used in a variational Monte Carlo (VMC) calculation, and the VMC results are further improved via diffusion Monte Carlo (DMC).

The trial Jastrow-Slater wave functions

We briefly remind the reader about the many-body wave functions we use since they are the key ingredient which determines the quality of our QMC calculations, and are here chosen of the Jastrow-Slater type, with the particular form:

$$\Psi_I^{\mathrm{VMC}} = \Psi_I^{\mathrm{CAS}} \prod_{A,i,j} \mathcal{J}(r_{ij}, r_{iA}, r_{jA}) \,, \tag{3.1}$$

where I labels the state of interest, r_{ij} denotes the distance between electrons i and j, and r_{iA} the distance of electron i from nucleus A. For most calcula-

tions, we use a Jastrow factor $\mathcal{J}$ which correlates pairs of electrons and each electron separately with a nucleus, and employ different Jastrow factors to describe the correlation with different atom types. We also investigate the effect of including electron-electron-nucleus correlation terms in the Jastrow factor. We refer the reader to Section 2.2.3 for further information on the specific form of the Jastrow factor.

The determinantal component consists of a complete active space (CAS) expansion which includes all possible space- and spin-adapted configuration state functions (CSF) obtained by placing n electrons in the active space of m orbitals, which is known as a CASSCF(n,m) wave function. Since the relevant excited state of all our models has the same symmetry as the ground state, the ground and excited-state wave functions are obtained in a state average (SA) CASSCF calculation, so the orbitals minimize the weighted average of the energies of the states of interest and the expansion coefficients are determined to preserve their orthogonality as explained in Section 2.2.1. Therefore, the wave functions of the different states share the same Jastrow factor and the same orbitals, but have different linear expansion coefficients on the CSFs. We note that, for the largest CASSCF wave functions, we may only keep the CSFs whose coefficient is above a chosen threshold. In this case, the threshold is separately applied to the ground- and the excited-state determinantal expansion, and the union of the surviving CSFs is then kept in both wave functions.

For the minimal anionic (A) model, we optimize all parameters in the Jastrow and the determinantal component of the wave function by energy minimization. Since the optimal orbitals and expansion coefficients in Ψ_I^{CAS} may differ from the CASSCF values obtained in the absence of the Jastrow factor $\mathcal{J}$, it is important to investigate the impact on the excitation energies when we reoptimize the determinantal parameters in the presence of the Jastrow component. If the wave function were the lowest state of a given symmetry, we could simply follow the energy-minimization approach of Ref. [66]. However, since the excited state in our models is not the lowest in its symmetry, we obtain both the Jastrow and orbitals parameters which minimize the average energy over the state of interest and the lower states, while the linear coefficients in the CSF expansion ensure that orthogonality is preserved among the states [65] as described in Section 2.2.3. For the other chromophore models studied in QMC, we either optimize a subset of parameters as the Jastrow factor and the linear coefficients, or simply optimize the Jastrow parameters in energy minimization for the ground state, and use the same Jastrow factor for the excited state calculations with an unoptimized SA-CASSCF determinantal component.

The trial wave functions of both states are then used in two separate

diffusion Monte Carlo (DMC) calculations, which produce the best energy within the fixed-node approximation, that is, the lowest-energy state with the same zeros (nodes) as the trial wave function.

All QMC calculations are performed with scalar-relativistic energy-consistent Hartree-Fock pseudopotentials specifically constructed for use in QMC, and with the corresponding Gaussian basis sets [67]. For most calculations, we employ a cc-pVDZ basis but we explore the effect of augmenting the basis with diffuse functions [68]. For this purpose, we generate an augmented cc-pVDZ basis (aug-cc-pVDZ) by adding one s and one p function with exponents 0.0469 and 0.04041 for carbon, 0.07896 and 0.06856 for oxygen, and 0.06124 and 0.05611 for nitrogen, respectively, and one s function with exponent 0.02974 for hydrogen.

3.4.1 The anionic minimal model: A case study

The anionic minimal (A) model represents a perfect playground to understand what the correct ingredients for our QMC calculations are. This model has the smallest number of atoms, so the calculations are faster, and has C_s symmetry which reduces the number of parameters in the determinantal component by roughly a factor of two. This reduction is convenient if we want to reoptimize the parameters of the SA-CASSCF wave function after including the Jastrow factor, and also because it is easier to converge very large CASSCF calculation with the quantum chemistry code GAMESS. In addition to these computational advantages, it is important to understand the anionic minimal (A) model which appears to be incorrectly described by adiabatic linear-response TDDFT (see Tables 3.2 and 3.5).

The determinantal CASSCF component

The first step in the generation of the many-body trial wave function is the construction of the SA-CASSCF determinantal component. Therefore, we want to explore the dependence of the excitation energy on the dimensions of the active space of the CAS wave function, and construct CASSCF(n,n) expansions of n electrons distributed over n orbitals with increasing values of n. The active space is build over the π/π^* orbitals (A″ symmetry) as the excitation of interest has $\pi \rightarrow \pi^*$ character and these orbitals are expected to be most relevant in describing the excitation. As 8 π orbitals are doubly occupied at the Hartree-Fock level, the maximum number of electrons in the active space can be $n = 16$.

In Table 3.6, we show the SA-CASSCF(n,n) excitation energies as a function of the dimension n of the active space for the anionic minimal (A) model

Table 3.6: SA-CASSCF(n,n) lowest excitation energy in eV of the anionic minimal (A) model as a function of the dimension n of the active space. We employ three different Gaussian basis sets, that is, cc-pVDZ, cc-pVTZ and aug-cc-pVDZ.

n	cc-pVDZ	cc-pVTZ	aug-cc-pVDZ
2	4.10	4.05	3.92
4	3.57	3.53	3.38
6	3.47	3.44	3.21
8	3.17	3.12	3.02
10	3.25	3.21	3.20
12	3.21	–	–
14	3.11	–	–

in combination with different Gaussian basis sets. With the code GAMESS which we use to perform all CASSCF calculations, we are able to converge the SA-CASSCF(n,n) wave functions only up to $n = 14$ when the cc-pVDZ basis is employed, while we only reach $n = 10$ for the larger basis sets. From the calculations with the cc-pVDZ basis, we note that enlarging the active space significantly reduces the CASSCF excitation, which reaches however a roughly constant value beyond $n = 8$. As far as the dependence from the basis set, we observe that increasing the valence character of the basis from double (cc-pVDZ) to triple (cc-pVTZ) reduces the excitation energy by less than 0.05 eV. The augmentation of the cc-pVDZ basis set with diffuse functions has a slightly larger effect on the excitation energy which is lowered by almost 0.2 eV. However, this gain appears to be lost if the active space is enlarged to $n = 10$ where the difference between the excitation energies with and without augmentation is only 0.05 eV. This finding is in agreement with the all-electron CASSCF/6-31G* calculations by Martin *et al.* who observe a reduction of 0.06 eV in the CASSCF(12,11) excitation energy by augmenting the basis with diffuse functions. We note however that their CASSCF/6-31G* excitation of 3.68 eV is higher than our CASSCF(10,10) value, a difference which is possibly due to their calculation being all-electron.

In summary, from the CASSCF study, we can draw several conclusions which are useful in setting up the quantum Monte Carlo wave functions. First, we do not need to increase the valence nature of the basis and using a cc-pVDZ basis should be sufficient. It is however important to check the effect of augmentation on the cc-pVDZ basis, which may be visible if one makes use of small CAS expansions. As far as active space, we should investigate the effect of increasing the dimensions of the CAS at least up to $n = 8$.

The impact of the trial wave function on the QMC excitation

Having established the necessary ingredients in the determinantal component of the wave function, we now want to determine where DMC places the excitation of the anionic minimal (A) model, and perform several tests which are summarized in Table 3.7. For these calculations, we employ different determinantal components and an electron-nucleus and electron-electron Jastrow factor, and optimize the Jastrow and determinantal parameters within energy minimization in a state average approach. For some of the wave function forms, we also report the energy obtained without reoptimizing the CASSCF component, and using for both states the optimal Jastrow factor determined for the ground state by energy minimization.

Table 3.7: Variational (VMC) and diffusion Monte Carlo (DMC) excitation energies in eV of the minimal anionic (A) model of the GFP chromophore. The dimension of the CAS(n,n) determinantal component, the threshold on the CSFs and the number of CSFs included in the wave function are listed. Energies are given for the fully optimized (Optimized) wave function and for the wave function (Unoptimized) where only the Jastrow factor is optimized in correspondence of the ground state. The CASSCF excitation energy is given in the last column. The basis set used is specified.

	CSF		Unoptimized		Optimized		
	Thr	Number	VMC	DMC	VMC	DMC	CASSCF
			cc-pVDZ basis				
CAS(2,2)	0.00	3	3.38(4)	3.18(4)	3.42(4)	3.15(4)	4.10
CAS(4,4)	0.00	20	3.33(4)	3.29(5)	3.31(4)	3.25(4)	3.57
CAS(6,6)	0.10	7	–	–	3.18(4)	3.11(7)	3.47
CAS(8,8)	0.10	9	–	–	3.06(4)	3.03(5)	3.17
CAS(8,8)	0.07	15	–	–	3.24(4)	3.10(5)	3.17
CAS(8,8)	0.05	25	–	–	3.11(4)	3.04(5)	3.17
			aug-cc-pVDZ basis				
CAS(2,2)	0.00	3	3.24(4)	3.11(7)	–	–	3.92

We begin our analysis of the DMC results with the simplest CAS(2,2) wave function which is constructed from the HOMO and the LUMO orbitals as this ansatz will also be employed in the preliminary QMC study

of the chromophores embedded in the protein environment. Optimizing the orbitals and linear coefficients of the CAS(2,2) determinantal component in the presence of the Jastrow factor significantly lowers the absolute VMC and DMC energies of the two states (not shown). However, the VMC and DMC excitation energies for the optimized and unoptimized wave functions are equivalent within statistical error. When moving to a larger CAS(4,4) expansion, the situation is rather similar in the sense that optimization does not significantly affect the VMC and DMC excitation energies. In comparison to the CAS(2,2) results, the DMC excitations energies of the CAS(4,4) wave functions appear higher even though the difference with the CAS(2,2) values is still within two standard deviations.

When going to larger CAS expansions, we only keep the CSFs with coefficient above a chosen threshold as the wave function would otherwise not be tractable within quantum Monte Carlo. The number of CSFs grows indeed very rapidly with the size of the active space: For instance, the CAS(8,8) wave function contains 1764 CSFs which become 19404 in a CAS(10,10) expansion. Since each CSF consists of several determinants, the full determinantal wave function can then be formed by several thousand determinants, which would render the QMC calculation exceedingly slow. We note that the wave function obtained by applying a threshold to a given CAS expansion does not reduce to one of the smaller CAS wave functions as different excitations may now be important and survive the threshold. Using a CAS(6,6) wave function with a rather large threshold of 0.1, we obtain a VMC excitation energy which is 0.22(6) eV smaller than the CAS(2,2) value, and a DMC excitation of 3.11(7) eV which is still compatible with both the CAS(2,2) and CAS(4,4) results within statistical error. Finally, we optimize the Jastrow, orbital, and linear parameters of a CAS(8,8) with a threshold of 0.1, and subsequently, reduce the threshold to 0.07 and 0.05 reoptimizing only the linear expansion coefficients. The VMC excitation does not behave monotonically as, by lowering the threshold to the intermediate 0.05 value, CSFs more relevant for the ground than the excited state may have entered the wave function. However, within statistical error, the DMC excitation energies are always rather comparable with each other falling in the range 3.03(5)-3.10(5) eV, and are certainly lower by at least 0.1 eV than the DMC values obtained with the optimized CAS(2,2) and CAS(4,4) wave functions.

In summary, it appears that the effect of increasing the CAS expansion and therefore improving the many-body wave function is to reduce the difference between the VMC and the DMC excitation energies, with the VMC gap approaching the DMC value from above. The DMC energy is very robust and depends not too strongly on the size of the active space with the DMC gap being lowered by only roughly 0.1-0.2 eV when going from a CAS(2,2) to a

CAS(8,8) wave function. The effect on the DMC excitation of optimizing the determinantal component in the presence of the Jastrow factor appears to be rather small. Finally, we note that the use of the cc-pVDZ basis is sufficient as augmenting the basis does not affect the DMC excitation energy of an unoptimized CAS(2,2) wave function and we know from the CASSCF study that the impact of diffuse functions would anyhow be washed out in going to larger CAS expansions. Further tests on the effect of using electron-electron-nucleus terms in the Jastrow factor, of performing a time-step extrapolation of the DMC results, and of treating the non-local pseudopotentials beyond the locality approximation do not change the present picture.

While this analysis reassures us about the robustness of our DMC approach in computing the brightest excitation energy of the minimal anionic (A) model, we must note that the DMC excitation energy is in the range 3.0-3.2 eV and therefore in rather good agreement with the TDDFT values of 2.97 and 3.16 eV obtained with the BLYP and B3LYP functionals, respectively. Consequently, the DMC excitation energy is significantly higher than the CASPT2/6-31G* value of 2.67 eV for the minimal anionic model [60], and not in good agreement with the absorption maximum of 2.59 eV obtained in photodistruction spectroscopy experiments for the closely related anionic methyl-terminated (B) model [58].

3.4.2 The neutral and anionic models at comparison

To better understand the apparent overestimation by DMC of the excitation of the minimal anionic (A) model, we want to compare the DMC results obtained for different model chromophores, that is, the minimal neutral (D), the anionic methyl-terminated (B) and the neutral$^+$ (G) model. This will allow us to observe how the excitation correlates to the charge state of the chromophore, and to compare the QMC excitations to other available reference data. Since the excitation energy of the anionic minimal (A) model is not particularly affected by either the optimization of the determinantal component or the use of large CAS expansions, we perform all QMC calculations consistently using an unoptimized CAS(2,2) determinantal component in the wave function. We will take into account in our analysis that the excitations may be overestimated by roughly 0.1-0.2 eV due to the use of the small CAS expansion. The results are summarized in Table 3.8.

We first compare the results for the minimal anionic (A) and neutral (D) models and note that DMC yields an excitation for the neutral species 0.9(1) eV higher than the one for the anionic counterpart. This finding is in agreement with the difference of 0.8 eV found in semi-empirical configuration interactions (CI) calculations for larger anionic and neutral chromophore

Table 3.8: Variational (VMC) and diffusion Monte Carlo (DMC) energies in eV for the minimal anionic (A) and neutral (D), the anionic methyl-terminated (B), and the neutral$^+$ (G) model, computed using an unoptimized CAS(2,2) wave function. The TDDFT/BLYP and B3LYP energies are also listed together with the photodistruction spectroscopy experimental absorption maxima from Refs. [58]^a and [59]^b.

	VMC	DMC	BLYP	B3LYP	Expt.
Minimal anionic	3.38(4)	3.18(4)	2.97	3.16	–
Minimal neutral	4.27(7)	4.09(7)	3.22	3.54	–
Methyl-term. anionic	3.43(5)	3.20(5)	2.89	3.09	2.59^a
Neutral$^+$	3.67(8)	3.37(7)	3.02	3.21	2.99^b

models in the gas phase [69]. The semi-empirical CI excitations energies for the anionic and neutral models are equal to 3.52 and 2.70 eV, respectively, but the parameters of the semiempirical approach were tuned to yield an excitation for the anionic model reasonably close to the experimental gas-phase value [70]. On the other hand, TDDFT yields a significantly smaller difference between the excitations of the neutral and the anionic minimal models, that is, 0.25 and 0.38 eV with the BLYP and the B3LYP functional, respectively. Therefore, while the TDDFT excitation for the anionic chromophore is in reasonable agreement with the DMC value, the TDDFT excitation of the neutral is lower than the DMC energy by 0.87 and 0.54 eV for the BLYP and the B3LYP functional, respectively.

If we compare the DMC excitation energies of the anionic methyl-terminated (B) model and the neutral$^+$ (G) models with TDDFT and available experimental numbers, we should keep in mind that the DMC excitations are overestimated by 0.1-0.2 eV since, for the moment, we only employed simple CAS(2,2) wave functions. While a reduction of 0.1-0.2 eV brings the DMC energies of both models in close agreement with the TDDFT/B3LYP excitations, the resulting excitation for the neutral$^+$ (G) model remains higher by 0.1-0.2 eV than the absorption maximum located in photodistruction spectroscopy experiments. Moreover, as already discussed, DMC appears to be overestimating the excitation of the anionic form by roughly 0.4-0.5 eV.

3.5 Conclusions

Before studying the neutral and anionic forms of GFP within the protein environment, we constructed a series of model chromophores of GFP in the gas phase to begin exploring the performance of adiabatic TDDFT and quantum Monte Carlo approaches. We built three models of increasing complexity for both the anionic (deprotonated) and the neutral (protonated) chromophore of GFP, and also investigated a particular cationic model which was recently characterized in photodistruction spectroscopy experiments. Reference experimental data obtained with the same spectroscopy technique as well as CASPT2 theoretical results are available for two of the smaller anionic models we study.

The results are rather puzzling. Adiabatic TDDFT reproduces the experimental absorption maximum of the cationic model reasonably well if a pure functional is used, but appears to be overestimating the excitation energies of two small anionic chromophores by 0.3-0.5 eV depending on the functional employed. Moreover, the TDDFT excitations energies for the neutral models are not too dissimilar from the excitations of the corresponding anionic chromophores while one would expect a significant shift in the excitation upon deprotonation. We analysed various possible shortcomings of TDDFT in describing the anionic form of GFP but we were not able to identify evident problems of TDDFT, especially in the calculations for the smaller models of the GFP chromophore. The excitations of these models do not appear to be characterized by charge transfer and curing the underestimation of the DFT ionization threshold with the use of asymptotically corrected functionals does not change the excitation energy. Therefore, it is not evident why TDDFT should be superior in describing the excitations of the corresponding neutral models or of the cationic chromophore. On the other hand, for the larger model chromophores that we will employ within the protein, we found significant contributions from charge-transfer transitions in the TDDFT excitations, which are a signature of potential problems with the TDDFT description of these larger models, especially in the neutral form.

Using quantum Monte Carlo approaches, we performed a thourough study of the excitation energy of the smallest anionic model using sophisticated trial many-body wave functions and state-of-the-art optimization techniques. We find that the DMC excitation energy is very close to the TDDFT result, and therefore higher than the reference photodistruction spectroscopy experimental result by roughly 0.4-0.5 eV. We could attempt to use even more complex wave functions but, based on the tests done so far, the DMC result appears to be rather robust. Differently from TDDFT, when comparing the anionic

and neutral chromophores, we observed a sizable shift in the excitation as the DMC excited state of the neutral form is 0.9(1) eV higher than the energy of the corresponding anionic model. Finally, the shift between the anionic and cationic model chromophores observed in photodistruction spectroscopy experiments is not well reproduced within DMC but, for both models, the DMC excitations appear to be rather close to the TDDFT values obtained using hybrid exchange-correlation functionals.

Chapter 4

Treating the protein environment in QM/MM

4.1 Construction of the protein model

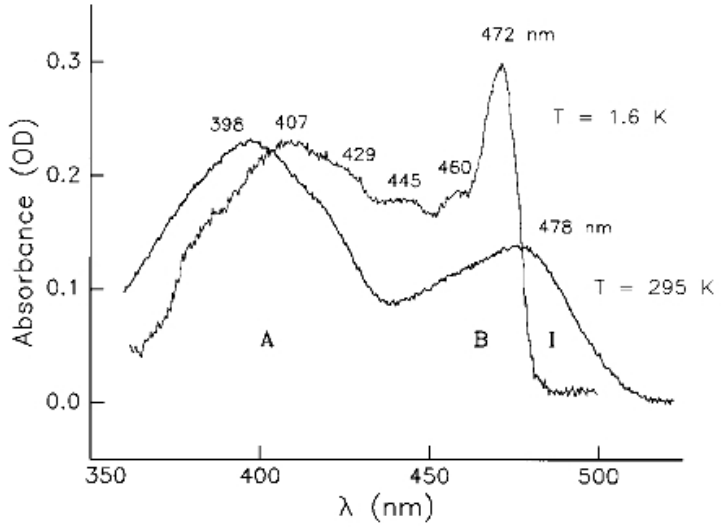

Figure 4.1: Absorption spectra of wild-type GFP at room temperature (T=295 K) and low temperature (T=1.6 K). Adapted from Ref. [7]

As the focus of this Chapter is the construction of a theoretical model to describe the photophysics of GFP, we briefly remind the reader some of the key spectral properties of this autofluorescent protein. As shown

in Fig. 4.1, the room temperature spectrum of wild-type GFP is characterized by two maxima which are attribuited to two interconvertible states (protonated/deprotonated) of the protein. The absorption band at 398 nm (3.12 eV) is from the neutral A form while the band at 478 nm (2.59 eV) from the anionic B form of the protein. Upon photoexcitation of A, the excited chromophore transfers a proton through a hydrogen-bond network to the surrounding environment forming a transient intermediate anionic state (I^*) which emits in the region of 506 nm (2.45 eV). After decay to the ground state (I), the system usually returns to state A through a ground state inverse proton transfer process. The green fluorescence at 482 nm (2.57 eV) following direct excitation of the B state stems from direct decay of the excited B^* state. In Fig. 4.1, the spectrum at 1.6 K is also shown where the ratio of the absorbances of the A and B forms inverts and the two maxima shift at 407 nm (3.05 eV) and 472 nm (2.63 eV). The broad wing at the red side of the 472 maximum disappears and is attributed to the I form which is not populated at this low temperature. The 0-0 transitions of all three forms have been located in hole-burning spectroscopy experiments [7].

To compute the spectrum of wild-type GFP, we construct a representative configuration of the chromophore-protein structure. As we compare to low-temperature (T=1.6 K) spectroscopy experiments, it is a reasonable choice to evaluate the theoretical spectrum of a single conformation obtained in a simulated annealing procedure instead of following the more costly route to compute the average spectrum over several snapshots of a molecular dynamics (MD) simulation at the experimental temperature. Moreover, we will see below that the chromophore is kept rather rigidly within its binding site by a complex network of hydrogen bonds. In this Section, we describe how this representative conformation for the neutral A and anionic I and B forms of wild-type GFP is constructed starting from the available X-ray structures of neutral wild-type GFP and of the GFP mutant S65T where the mutations stabilize the anionic B form against the A form of GFP.

4.1.1 The neutral form

The starting structure for the construction of the neutral form of the protein is the X-ray structure [71] at 1.90 Å resolution (entry 1GFL in the Protein Data Bank [72]). During the crystallization process precedent to the X-ray scanning, the proteins pack together to form dimers so that the X-ray structure with code 1GFL contains not one but two GFP proteins. In our simulations of the neutral and anionic I forms, we keep the dimer for simplicity. The setup and a fast preliminary MM equilibration is performed using the Amber suite of programs [43] and the Amber force field is used in all

MM simulations. The structure is then refined in a simulated annealing run within QM/MM.

Equilibration of GFP with MM

Since X-ray diffraction experiments do not distinguish hydrogen atoms, the positions of the hydrogens are not given in the X-ray structure of GFP and must be added when setting up the simulation. This is in part automatically done using the module *Xleap* within the Amber package as most of the standard aminoacids exist in a given ionization state at neutral pH. Exceptions are the glutamine/glutamic acid, aspartic acid, lysine and histidine which can have different protonation states. Their charge state is assigned based on their hydrogen bonding configuration, or chosen as they most frequently appear in nature. Of these aminoacids, only the histidine and the glutamic acid are present in the hydrogen-bond network surrounding the chromophore. Since we do not perform long MD simulation but only refine the structure by simulated annealing, we focus our attention on the protonation state of the histidines and the glutamic acids in the binding site, and leave the Amber package to assign the protonation states of the other residues. Based on the most likely hydrogen-bonding configuration of the chromophore, the histidine residues numbered 25, 148, 181, 199 and 217 are protonated at their δ nitrogen while the remaining hystidine residues are protonated at their ϵ nitrogen. For the protonation of the glutamic acids, particular attention is given to Glu-222 since this residue is involved in the proton shuffle between the neutral A and the anionic forms. As explained later, experiments and theoretical work indicate that Glu-222 is anionic in the A state while is the proton receptor in the anionic I and B form of GFP. We follow this criterion in our setup.

To simulate the protein *in vivo*, the protein is then set in a physiological solution, with a saline concentration of 0.15 M. The protein is placed at the center of a cubic MM simulation box, surrounded by 12 Å of TIP3P [73] water molecules in each direction. As periodic boundary conditions are applied, the box is sufficiently large to avoid interactions between the images of the protein. Finally, counter-ions are added to the water solution to achieve physiologic concentration and to ensure that the cell is completely neutral. As the total sum of the partial charges of the protein is not zero but equals -12, we added more positive ions (63 Na^+) than negative ones (51 Cl^-) to have a simulation box with zero total charge. The total simulation box contains around 70000 atoms and its size is roughly $83 \times 97 \times 82$ Å^3. In Fig. 4.2, we show part of the simulation cell comprising the protein and a few Å of the surrounding water.

We first perform a classical MM equilibration of the hydrogens and the waters using the module *Sander* of the Amber program in order to start the more costly QM/MM simulations with these many degrees of freedom in a stable configuration. During the relaxation, the coordinates of the heavy atoms of the protein are kept fixed to preserve a protein structure close to the original crystallographic one, and because the force field parameters for the chromophore are not available in the Amber library as the chromophore is not a standard aminoacid. The chromophore force field parameters are available at the Charmm level [74] but their use would then be incompatible with the subsequent QM/MM simulations.

Even though the heavy atoms of the chromophore are kept fixed, we need to assign a force field to the chromophore as its atoms interact with the hydrogens of the protein, the solution waters, and the ions which are being equilibrated. Fortunately, as the chromophore is kept fixed, we do not need a very accurate force field and we can proceed to generate one for our particular purpose as follows. We determine the partial charges of the atoms of the chromophore by computing a quantum mechanical electrostatic potential within Hartree-Fock using a 6-31G* basis and the Gaussian03 [53]

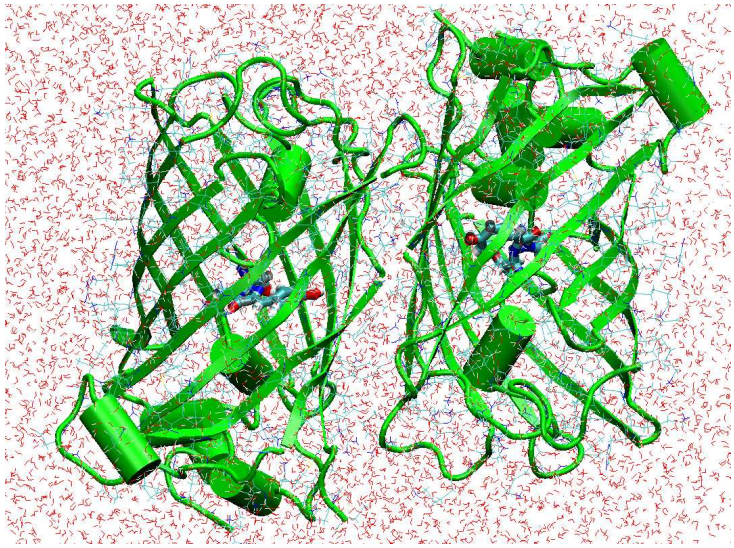

Figure 4.2: Model of the neutral A form of wild-type GFP in water solution. The dimerized structure is shown together with the chromophores in the protein cavities. The diameter of the barrel of a single GFP protein is approximately 24 Å and the height 42 Å. The distance between the two proteins is approximately 5 Å. The two chromophores do not interact as they are shielded by the protein barrel and are at a distance of about 17 Å.

code. The partial charges are then fitted to reproduce this potential on a large number of grid points covering the region around the chromophore. We assign the bond, angle, and dihedral parameters to each atom of the chromophore based on its position and chemical role following Ref. [75]. We know that this force field is not sufficiently accurate to describe the intra-molecular interactions of the heavy atoms of the chromophore since an energy relaxation test performed for this chromophore model does not yield a stable structure. However, we consider it sufficient to give an approximate description of the interaction of the chromophore atoms with the rest of the system.

To restrain the protein coordinates, we find that simply locking the protein coordinates leads to an unstable simulation. Therefore, we introduce a restoring potential, $k\,\Delta x^2$, where Δx is the displacement of the atomic coordinates respect to the reference X-ray structure. Finally, the MM equilibration is performed in the following three steps:

1) 1000 steps of energy minimization using the steepest descent algorithm in the first 10 steps, and the conjugate gradients algorithm afterwards.
2) 33 ps of classical isothermal and isobaric molecular dynamics at 300 K and 1 atm. We use an isotropic pressure scaling and the Andersen thermostat to couple the temperature to an external bath. We consider the system equilibrated when the temperature fluctuations are less than 5% and the density is constant within 2%. The time step is 0.0005 ps.
2) 3500 steps of energy minimization.

Finally, an equilibrated structure is obtained with a norm of the atomic forces of 0.1 kcal/mol/Å.

Simulated annealing within QM/MM

To obtain an accurate description of the structure of the chromophore binding site and of the internal geometry of the chromophore which was left so far unrelaxed, we start from the structure obtained in the preliminary MM equilibration and perform a simulated annealing procedure within QM/MM. In the QM/MM relaxation, the atoms of the chromophore are treated quantum mechanically and the rest of the protein and the solvent are treated classically. The QM/MM boundary is set through the single C_{OOH}—C_α bond of Phe-64 and the single N—C_A bond of Val-68 as shown in Fig. 4.3. We note that, as we keep the dimer structure of the original crystallographic structure, we only treat quantum mechanically the chromophore of one of the two GFP barels and leave the other chromophore at the fixed crystallographic coordinates.

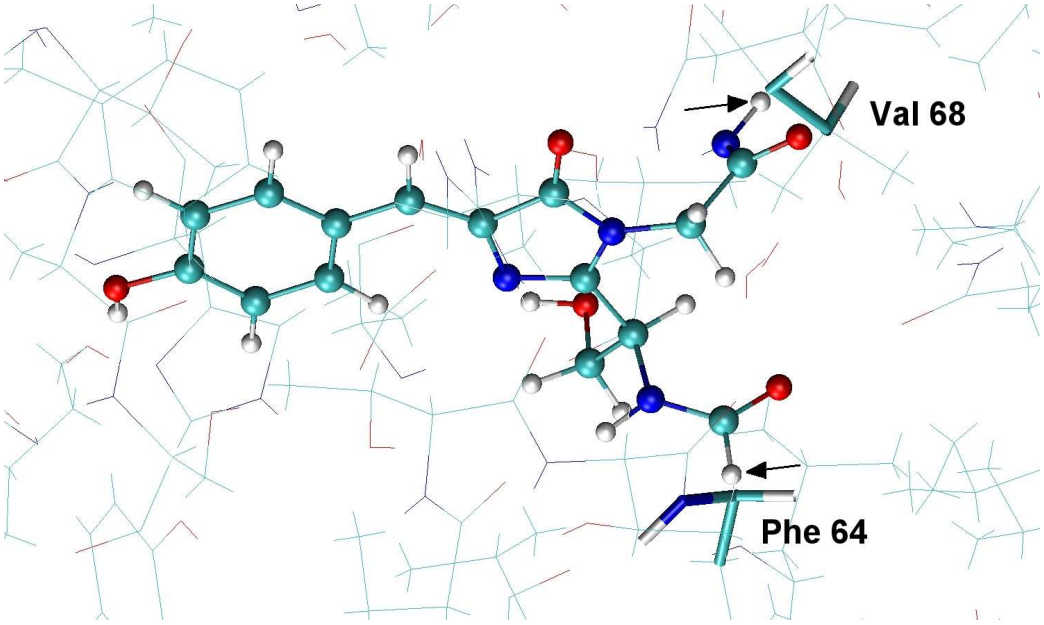

Figure 4.3: QM/MM partitioning for the A form of wild-type GFP. The QM/MM boundary is set through the single C_{OOH}—C_α bond of Phe-64 and the single N—C_A bond of Val-68. The two arrows indicate the two hydrogen-link atom between the QM and the MM. The MM part closest to the QM atoms is represented with cylinders.

The QM/MM simulations are performed with the density functional theory CPMD [46] code using the PBE generalized gradient approximation functional [52] and a plane wave cutoff of 70 Ry. The size of the supercell used for the QM calculations is roughly $21 \times 15 \times 13$ Å^3, and the chromophore is approximatively in the xy plane. A time step of about 0.7 fs is used.

The simulated annealing is performed through a molecular dynamics simulation where the velocities are rescaled at each step by a factor f. To quench the system and obtain a converged structure of the chromophore binding site, we perform:

1) 1.40 ps of molecular dynamics with $f = 0.999$.
2) 0.35 ps of molecular dynamics with $f = 0.99$.

At the end of the quenching, the amplitude of the oscillations for the bond lengths along the chromophore are of the order of 0.005 $\AA$, and thus negligible.

In Fig. 4.4, we show the final structure of the binding site of the neutral A form with the closest residues to the chromophore which are most relevant for either of the three forms of wild-type GFP. The A form is character-

ized by a hydrogen-bond network connecting the oxygen of the chromophore to the carboxylate of the negative Glu-222 through the water W1 and Ser-205. As we discuss below in Section 4.1.2, the accepted picture is that this hydrogen-bond chain provides the path for the proton leaving the photoexcited chromophore and being transferred to Glu-222. We also observe that the negatively charged Glu-222 is stabilized by a hydrogen bond donated by the oxygen of the tail of the chromophore (Ser-65). The stabilization of the anionic Glu-222 is crucial for the existence of the A form as perturbation of the hydrogen-bond network surrounding this residue leads to destabilization of the A state in favor of the B state.

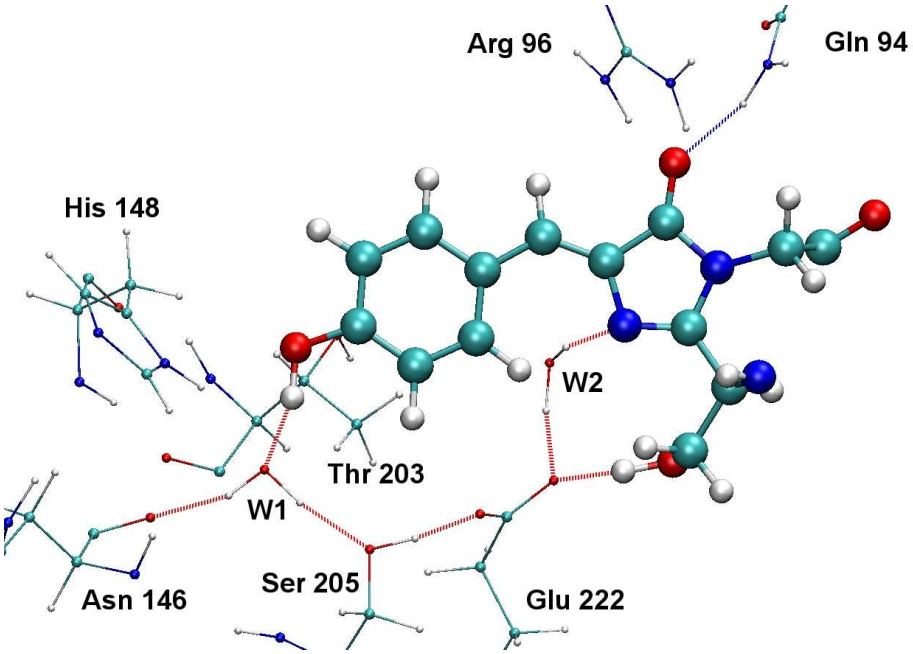

Figure 4.4: Binding site of the neutral A form of GFP. The residues closest to the chromophore which are relevant in either the A, B, or I form are shown. The hydrogen bonds are drawn if the bond length is less than 3 Å and the donor-hydrogen-acceptor angle is less than 30°. Glu-222 is negatively charged in the A form and is the acceptor of the proton leaving the hydroxyl group of the phenolic ring upon photoexcitation.

4.1.2 The intermediate form

Both the anionic intermediate and B forms differ from the neutral form as the proton of the hydroxyl group of the phenolic ring has left the chromophore

and has been transferred to the sourrounding environment. For both forms, we assume that the proton leaving the phenolic ring is transferred to the negatively charged residue Glu-222 which becomes neutral upon protonation.

The identification of the residue accepting the proton leaving the chromophore has been a subject of debate for several years. Already in the early models proposed after structural and spectroscopic analysis of wild-type GFP and some of its mutants [76–78], it was suggested that Glu-222 is negatively charged in the A form and the acceptor of the proton upon photoexcitation. This picture was for instance put forward by Ormö *et al.* [76] in their X-ray characterization of the S65T mutant where the B form is stabilized by mutation of the side chain of the chromophore at position 65, which donates a hydrogen bond to Glu222 in the A form of wild-type GFP (see Fig. 4.4). This transfer mechanism was however doubted by successive FTIR (Fourier transform infrared spectroscopy) experiments [79], where Glu-222 appears to be protonated in both the neutral and the anionic form as the feature corresponding to the change in its protonation is absent in the FTIR difference spectrum. However, theoretical calculations of the vibrational frequencies of both forms [80] allowed to reinterpret the FTIR results by showing their compatibility with a protonation state of Glu-222 in the B form different from the one in the A form. Finally, the good agreement between classical MD calculations [69] and the X-Ray structure of the S65T mutant where the chromophore is stabilized in the anion form with a protonated Glu-2222 also supports the picture of a proton transfer mechanisms from the chromophore to Glu-222 in wild-type GFP.

Finally, we note that the proton can be added to Glu-222 in the *syn*- or the *anti*-configuration. In the theoretical work by Tozzini and coworkers [69], it was observed that, during the classical molecular dynamics simulations at 298 K, the hydrogen jumps between the two configurations but that the *anti*-conformation is the one occupied for longer times. Therefore,we protonate the Glu-222 residue with the proton in the *anti*-configuration. In Fig. 4.5, we show the anionic intermediate form with the chromophore and the most important residues which are believed to be involved in the proton transfer mechanism.

For the calculation on the I form of the chromophore we followed the same procedure as described for the A form. A QM/MM simulated annealing procedure is performed on the structure following these steps:

1) 3 ps of molecular dynamics with $f = 0.999$.
2) 1 ps of classic molecular dynamics at constant volume at 300 K. Only the residues involved in the hydrogen-bond network of Fig. 4.5 are moved while the chromophore and the rest of the protein is kept fixed.

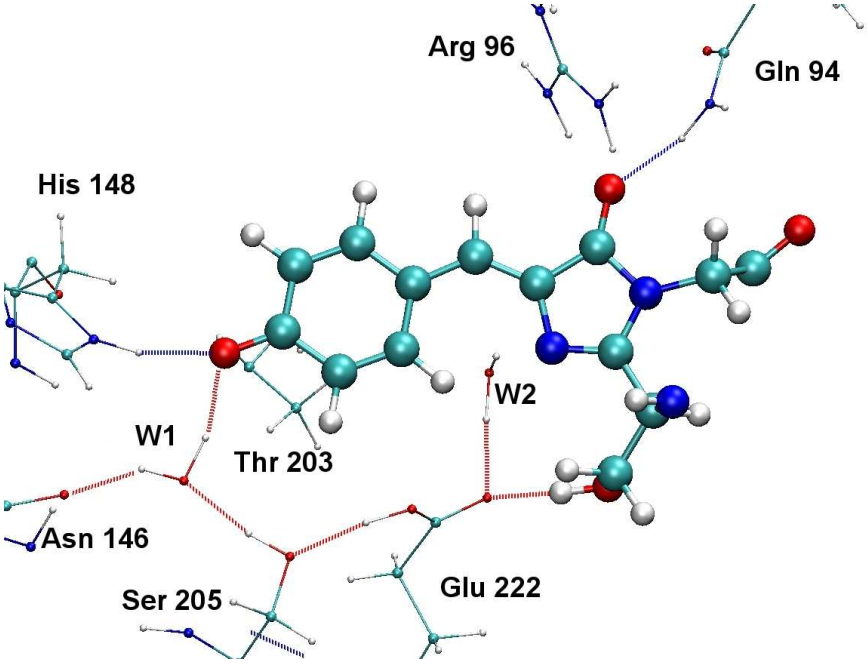

Figure 4.5: Binding site of the anionic I form of GFP. The residues closest to the chromophore which are relevant in either the A, B, or I form are shown. The hydrogen bonds are drawn if the bond length is less than 3 Å and the donor-hydrogen-acceptor angle is less than 20°. The chromophore is deprotonated and Glu-222 is commonly considered as the acceptor of the proton leaving the hydroxyl group of the phenolic ring.

This allows a faster rearrangement of the residues interested in the proton-shuffle.

3) 0.28 ps of molecular dynamics with $f = 0.999$.

In Fig. 4.5, the final structure of the binding site of the I form is shown with the closeby residues which are the most relevant for either of the three forms of wild-type GFP. The hydrogen-bond network connecting the phenolic oxygen to Glu-222 is shown. In the anionic form, His-148 donates a hydrogen bond to the phenolic oxygen which is now deprotonated while Glu-222 is neutral. The electrostatic potential generated by the MM atoms of the protein and the solvent on the QM system is shown in Fig. 4.6. As expected, the potential is positive and therefore attractive for electrons around the positive Arg-94 and the hydrogen atom of Gln-94 which creates a hydrogen bond to the oxygen of the imidazole ring. It is also positive along the hydrogen-bond network running from the phenolic oxygen through the W1 water and Ser-205 to Glu-222.

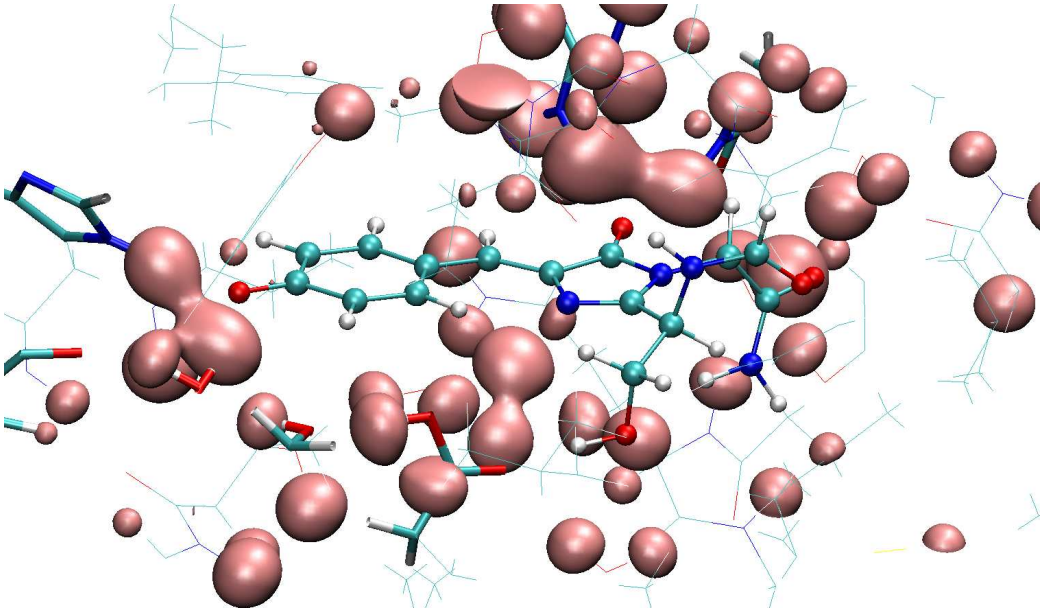

Figure 4.6: Electrostatic potential created by the MM atoms on the QM system. The residues which hydrogen bond to the chromophore are shown in a thicker representation (also compare with Fig. 4.5). An isosurface of 0.2 a.u. is shown in pink while no negative isosurface of -0.2 a.u. is close to the chromophore.

4.1.3 The B form

The construction of a model for the B form of wild-type GFP is more complex as the protein has been only crystallized in the neutral A form (entry 1GFL in the Protein Data Bank). The B form differs from the neutral one not only in the protonation of the chromophore and the Glu-222 residue but also because various residues in the binding pocket of the chromophore have a different, more stable conformations. The X-ray structure is however available for mutants of wild-type GFP where the protein is not in its original sequence but some mutations have been done to enhance the stability of the anionic B form. The environment of these mutants apart from the particular mutations is believed to be closer to the one of wild-type GFP in the B form.

The basic idea in the construction of the model for the B form is to start from the crystallographic structure of one of these mutants and "undo" the mutations to restore the protein sequence as in its wild-type expression. Following the theoretical work by Nifosi and Tozzini [69], we start from the crystallographic structure of mutant S65T (code 1EMG in the Protein Data Bank [72]). This X-ray structure shows two main differences with the wild-

type structure as in the 1GFL. First, the S65T substitution is present where the aminoacid serine at position 65 (Ser-65) is substituted with a threonine (Thr-65). Second, the biologic unit is now a monomer.

To restore the wild-type structure, we perform a homology modelling. To restore the aminoacid Ser-65, the methyl group of Thr-65 is substituted by a hydrogen atom, and the resulting Ser-65 oxygen tail is flipped in the other direction towards the close Glu-222 residue to form a hydrogen bond. This structure is then ready for the classic MD simulation for the preliminary equilibration and the subsequent QM/MM dynamics. In Fig. 4.7 we show the hydrogen bond network for the B form of the protein after our homology modeling construction as compared to the neutral and the intermediate anionic form.

Starting from the structure constructed by homology, we proceed as in the case of the neutral form. Using the Amber suite of programs, we add the missing hydrogens, the waters and the counterions. For the B form, we only add 10 Å of TIP3P water molecules in each direction to avoid the occurrence of uninfluential but unpleasant vacuum bubbles at the edges of the simulation cell which were present in the simulation cell of the neutral and intermediate forms. The total charge of the protein is -6, the number or counterions added is 28 Na^+ and 22 Cl^-, and the box dimensions are approximately $65 \times 79 \times 78$ Å^3 for a system of about 31000 atoms. In the B form, the number of atoms in the cell is approximately half the number for the A or I form because here the crystallized form is a monomer. As before, we equilibrate the hydrogens and the waters with a classic MD approach, and we then perform a simulated annealing run with the QM/MM approach. The procedure followed was:

1) 1.4 ps of molecular dynamics with $f = 0.999$.
2) 0.35 ps of molecular dynamics with $f = 0.997$.
3) 0.35 ps of molecular dynamics with $f = 0.995$.
4) 0.35 ps of molecular dynamics with $f = 0.991$.
5) 0.35 ps of molecular dynamics with $f = 0.987$.

The annealing factor is slowly decreased to allow a smoother quenching of the system. The main hydrogen-bond distances are smoothly dumped during this slow quenching until a temperature less than 0.02 K is reached.

In Fig. 4.7, we show the optimal binding site of the chromophore as obtained in our QM/MM simulation. The hydrogen-bond network surrounding the chromophore is very similar to the one of the I form (see Fig. 4.5) with a chain of hydrogen bonds running from the chromophore to Glu-222 and

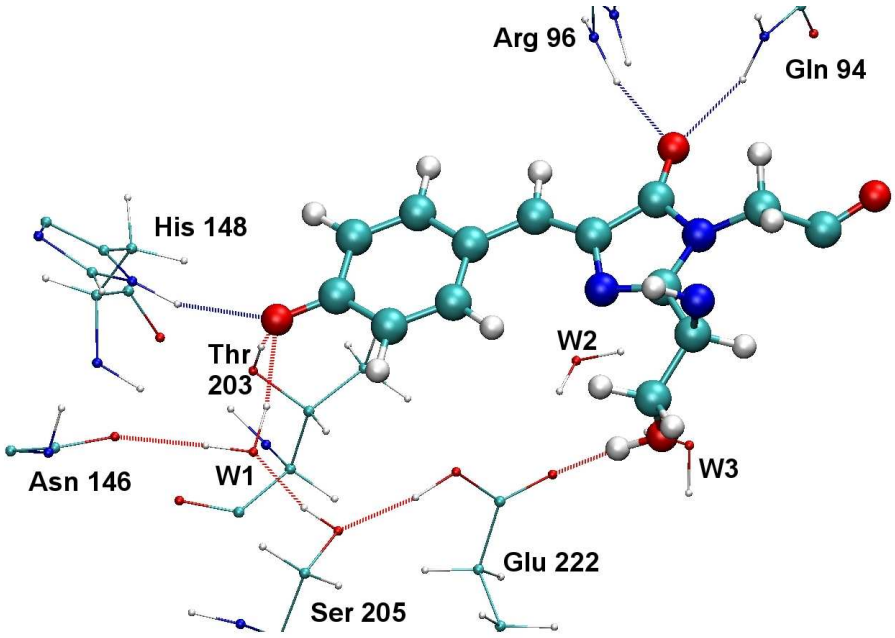

Figure 4.7: Binding site of the anionic B form of GFP. The residues closest to the chromophore are shown. The hydrogen bonds are drawn if the bond length is less than 3 Å and the donor-hydrogen-acceptor angle is less than 30°. The oxygen of the phenolic ring is deprotonated and the Glu-222 residue becomes neutral as the proton acceptor. The position of Thr-203 is different than in the I form and forms a hydrogen bond with the chromophore.

back to the chromophore. In addition, Thr-203 is now positioned to further stabilize the negative charge on the chromophore as it donates a hydrogen bond to the phenolic oxygen. Since this additional hydrogen bond further stabilizes the electronic ground state, the absorption maximum of the B form is expected to be blue shifted with respect to the I form.

4.2 Structural analysis of the models

To analyze the structural features of the models of the neutral and anionic forms of wild-type GFP obtained in our QM/MM calculations, we focus on the binding site of the chromophore as these local properties will predominantly determine its excitation spectrum. For the labeling of the atoms of the chromophore, we refer the reader to Fig. 4.8 where the heavy atoms are numbered starting from the phenolic oxygen along the top ridge of the

phenol, through the bridge and around the imidazole ring.

The internal structure of the chromophore is expected to play a very important role in tuning the excited state properties of the chromophore. In particular, the degree of bond-length alternation in the conjugate chain running through the chromophore is correlated to the size of the gap between the ground state and the lowest π-bonding to antibonding excitation with a stronger bond-length alternation yielding a larger gap. This fact can be easily understood by considering the limiting case of an infinite chain of carbon atoms, where the gap is zero in the absence of bond-length alternation and opens as the chain dimerizes. In the particular case of GFP, the existence of correlation between gap and bond-length alternation has been largely verified through the analysis of several GFP mutants for which the X-ray structure is available, and the construction of simplified theoretical models of the binding site of the GFP chromophore [69].

The protein environment and, in particular, the residues in the binding site of the chromophore can affect the spectral response of the chromophore in a dual manner. They can in principle tune the internal geometrical structure of the chromophore as well as act on the excitations more directly as some close residues may be charged or form hydrogen bond to the chromophore. The hydrogen-bond network in the active site is in fact rather different in the three forms of GFP where residues play different roles in stabilizing either the ground state (opening the gap) or the excited state (closing the gap). The role of the positively charged Arg-96 in close proximity to the GFP chromophore will also be discussed at length below.

We begin with a detailed analysis of the neutral A form of GFP as, only for this form, a comparison with the X-ray structure is possible [71]. The anionic B form is constructed by homology starting from the X-ray structure of the mutant S65T but a direct comparison with the crystallographic data of the mutant would be misleading as the positions of some close residues are rather different. In Table 4.1, we summarize the main structural features of the chromophore of the neutral A form of GFP within the protein and in vacuum, and compare with the X-ray structure at 1.90 Å resolution used as starting geometry. We first note that the structural properties of the chromophore optimized in the protein and in vacuum are remarkably similar. Both chromophores are essentially planar, and the presence of the protein only yields an average bond-length shortening of 0.01 Å without affecting the bond-length alternation of the conjugate chain of the chromophore. When comparing the internal bonds of the chromophore resulting from our QM/MM simulation to the initial X-ray structure, we see that the agreement is rather good with a root mean square deviation of 0.04 Å, and a maximum deviation of only 0.07 Å. The hydrogen-bond lengths between

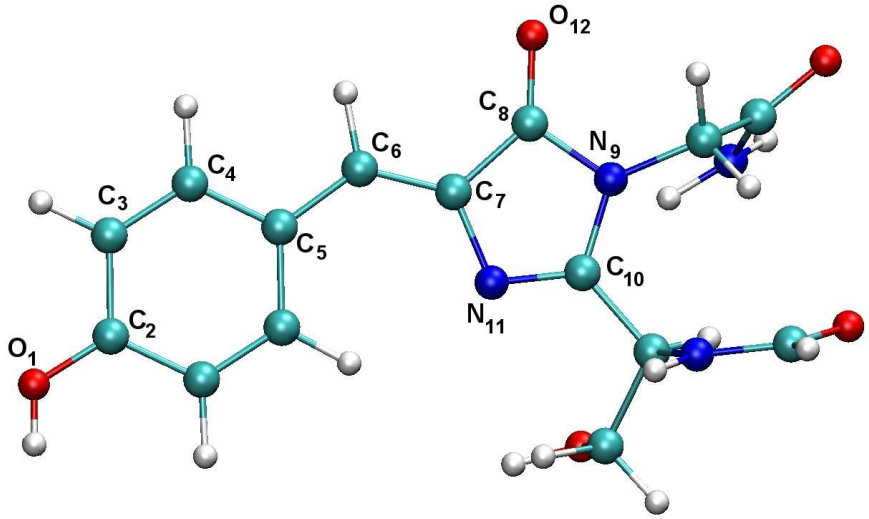

Figure 4.8: Atom numbering used for the chromophore of neutral GFP.

hydrogen donor and acceptor are shown for residues in close proximity to the chromophore and are also found in very good agreement with the X-ray structure.

In Fig. 4.9, we show the bond lengths along the chromophore of the neutral A and the anionic I and B forms of GFP obtained in our DFT/PBE QM/MM calculations and compare them with the the structures optimized in vacuum. As in the case of the neutral chromophore, the bond lengths of the anionic I and B forms do not dramatically differ from the values we obtain when the anionic chromophore is optimized in vacuum. Slight differences are observed in the degree of bond alternation close to the central bridge as the difference between the two central bond lengths, C_5—C_6 and C_6—C_7, is smaller in vacuum than in the protein. Overall, we can however conclude that, for all three forms of wild-type GFP, the protein environment acts to preserve an internal structure of the chromophore which is not dramatically altered with respect to the one in vacuum.

As in the comparison of the gas-phase neutral and anionic chromophores in Chapter 3, it is easier to understand the geometrical changes with the charge state of the chromophore if we show again the two resonant forms of the anionic chromophore in Fig. 4.10. In the benzenoid form, the negative charge is localized on the phenolic oxygen and this bond structure is therefore also characteristic of the neutral chromophore. Upon deprotonation, the quinonoid form is also accessible where the negative charge has migrated to

	X-ray	Protein	Vacuum
Bond length (Å)			
O_1—C_2	1.40	1.36	1.37
C_2═C_3	1.39	1.40	1.41
C_3—C_4	1.37	1.38	1.40
C_4═C_5	1.38	1.42	1.43
C_5—C_6	1.40	1.44	1.45
C_6═C_7	1.41	1.37	1.38
C_7—C_8	1.49	1.47	1.50
C_8—N_9	1.32	1.39	1.43
N_9—C_{10}	1.35	1.40	1.39
C_{10}═N_{11}	1.34	1.31	1.33
N_{11}—C_7	1.45	1.40	1.41
C_8═O_{12}	1.22	1.25	1.23
Dihedral angle (°)			
D($C_4C_5C_6C_7$)	173.9	176.8	179.0
D($C_6C_7N_{11}C_{10}$)	178.9	177.1	179.6
Hydrogen-bond length (Å)			
Arg-96(N) ··· CHR(O_{12})	2.71	2.76	–
Gln-94(O) ··· CHR(O_{12})	3.08	2.86	–
CHR(O_1) ··· W1(O)	2.64	2.56	–
His-148(N) ··· CHR(O_1)	3.27	3.21	–

Table 4.1: Structural parameters for the neutral A form of GFP obtained in our QM/MM simulations as compared to the X-ray structure used as initial configuration. We list the most representative bond lengths and dihedral angles of the chromophore, and the hydrogen-bond distances between the hydrogen donor and acceptor (D ··· A) for close residues to the chromophore (CHR). We also list the values for the chromophore optimized in vacuum. See Fig. 4.8 for the labeling of the atoms of the chromophore. For reference, the bonds of the central carbon bridge are C_5—C_6 and C_6═C_7.

the imidazole oxygen.

When comparing the neutral and the anionic forms, we see that the largest difference occurs in proximity of the hydroxyl oxygen of the phenolic ring. After deprotonation of the hydroxyl group, the oxygen-carbon bond, O_1—C_2, looses its single-bond character and significantly shortens by about 0.1 Å when going from the neutral to the anionic form. The shortening is slightly

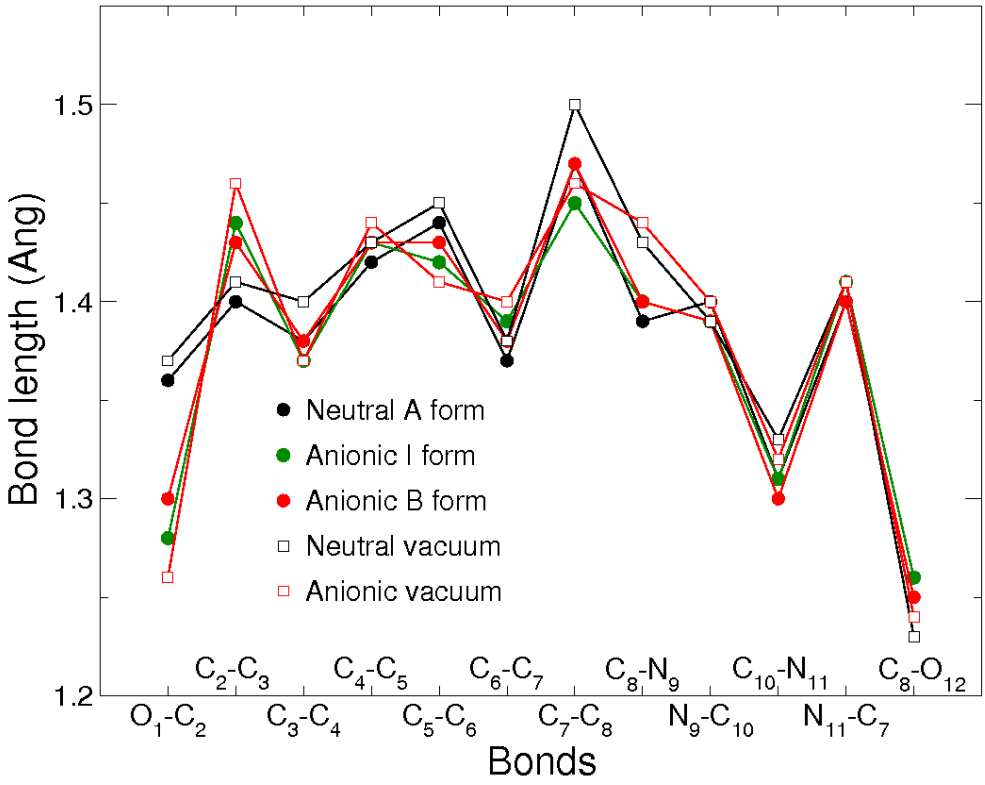

Figure 4.9: Structural properties of the chromophore of the neutral A, and the anionic I and B forms as obtained in our QM/MM simulations. We also show the neutral and anionic structures optimized in vacuum within all-electron DFT/BLYP with a cc-pVTZ basis. The bonds of the central carbon bridge are C_5—C_6 and C_6—C_7.

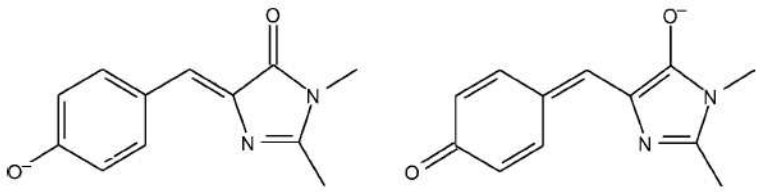

Figure 4.10: Scheme of the two resonant forms of the anionic chromophore: Benzenoid (left) and quinonoid (right).

smaller in the B form as compared to the I form as the negative charge on the oxygen is further stabilized in the B form by the position of Tyr-203, which donates an additional hydrogen bond to the oxygen. As a consequence, while the neutral model is characterized by a more marked aromaticity of the phenolic ring (with more similar bond lengths between all carbon atoms of the ring), the aromaticity of the phenolic ring of the anionic forms is reduced, yielding a quinoid structure. The effect of deprotonation is also visible for the bonds in proximity of the central bridge of the chromophore where the degree of bond alternation is reduced in the anionic as compared to the neutral form. Interestingly, we note that the difference between the neutral and the anionic bond lengths is overall larger in vacuum: For instance, the two central bonds of the carbon bridge in the anionic species have almost the same lengths, the difference being less than 0.01 Å, while in the neutral species, these two bonds differ by as much as 0.07 Å. In the protein, the same differences for the I and the neutral form amount to 0.04 and 0.07 Å, respectively. By inspecting the other bonds along the conjugated chain, it appears that the protein environment acts to partially compensate the change in protonation state, and keeps the chromophore in a more similar structural conformation in the two protonation states with respect to vacuum.

In Fig. 4.12, we compare the geometrical properties of the chromophore of the three forms of GFP with the results of other simulations available in the literature. We first focus on the comparison with the work by Marques *et al.* [12] who construct their protein models within a DFT/LDA QM/MM approach. It is important to understand the structural features of their models as they most significantly differ from ours as well as from other calculations reported in the literature. Moreover, we will show later that, using their DFT/LDA QM/MM structures, Marques *et al.* can claim an excellent agreement between TDDFT and the experimental absorption spectra of wild-type GFP.

For the neutral A form, the structural agreement of our model with the DFT/LDA calculations by Marques *et al.* [83] is reasonable while, for the anionic I form, the models are significantly different. In the neutral form, Marques *et al.* find a slightly more marked bond-length alternation along the chromophore, and the same structure is essentially preserved when moving to their anionic I form, with the exception of the oxygen-carbon bond of the phenolic group, O_1—C_2, which is significantly shorter in the I form in agreement with our calculations. In particular, the carbon bond alternation in the central carbon bridge is 0.1 Å in the neutral form and 0.09 Å in the I form, compared to 0.07 Å and 0.04 Å in our calculations. Another evident difference with our results is the bond length of the subsequent single carbon bond of the imidazole ring, C_7—C_8. In their LDA/DFT model, this bond

length is longer by 0.05 Å and 0.07 Å than the values we obtain in the neutral and anionic I form, respectively.

The difference between our model and the one by Marques *et al.* cannot be attributed to the use of LDA instead of the generalized gradient approximation PBE as the exchange and correlation functional employed when relaxing the chromophore geometry. We showed in our studies of model chromophores in vacuum that the use of LDA yields essentially the same structural parameters as PBE, BLYP or B3LYP (see Fig. 3.3). A closer analysis of the protein structures of Marques *et al.* reveals instead that the different geometry of the chromophore is likely due to particular choices made in the setup of the protein environment. The most striking feature in their models is that *all* histidine protein residues in both the neutral and I forms are protonated both at the δ and ϵ nitrogens, and carry therefore a positive net charge. As most histidines are far from the chromophore site, their charge state will not significantly affect the chromophore geometry. However, His-148 is close to the chromophore and even creates a hydrogen bond with the chromophore in the anionic state (Figs. 4.5 and 4.7). Therefore, since protonating both nitrogen of His-148 corresponds to placing an additional positive charge in close proximity to the chromophore, we may expect a noticeable effect on the structural parameters of the system. For clarity, we note a rather misleading feature of the work by Marques *et al.* as the structural coordinates made available by the authors do not correspond to the figure which appears in their paper where His-148 is drawn as only protonated at the δ nitrogen. We also note that Marques *et al.* construct the I form by deprotonation of the neutral form but always refer incorrectly to this structure as the B form.

To proceed, we need to motivate why we believe our model to be a closer representation of wild-type GFP in its various protonation forms than the one by Marques *et al.*. Our model corresponds to one accepted in the literature and its validity is mostly supported by the X-ray characterization of the structures of wild-type GFP and other mutants. The β barrel in GFP appears to be partially perturbed around the phenolic end of the chromophore in proximity to His-148. The β strand that covers the chromophore moves around His-148 and the backbone of one strand from residue 144 to 150 is not directly hydrogen bonded to the adjacent backbone between residues 165 and 170. The two backbones appear instead to be held together by forming hydrogen bond with the imidazole ring of His-148, with the backbone nitrogen of Arg-196 being a donor to nitrogen ϵ of His-148.

Accordingly, as shown in Fig. 4.11, we do not protonate the nitrogen ϵ of His-148 but allow it to be a hydrogen-bond acceptor for the backbone nitrogen of Arg-168. With this assumption, we obtain a structure for the neutral form with an Arg-168(N)$\cdots$His-148(N) distance of 3.05 Å in close

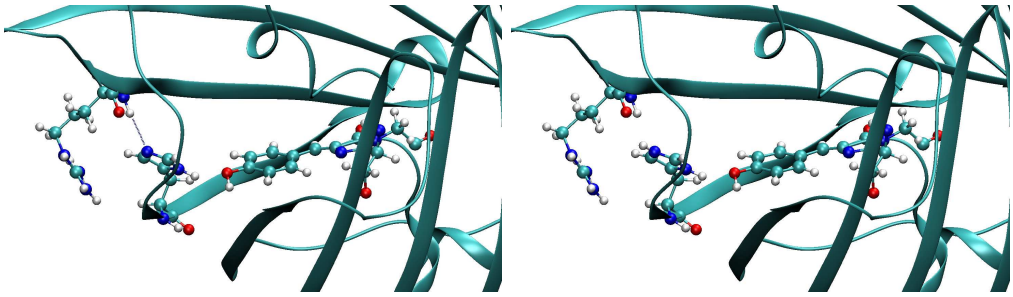

Figure 4.11: Particular of the structure of the I form for our model (left) and the one by Marques *et al.* [12] (right). The chromophore, the residues His-148 and Arg-168, and part of the β barrel are shown. Note the difference in the protonation of the nitrogen ϵ of His-148 which, in our model, is a hydrogen-bond acceptor for the backbone nitrogen of Arg-168.

agreement with the X-ray value of 2.94 Å. On the other hand, Marques *et al.* find a significantly larger nitrogen-nitrogen distance of 3.40 Å since their His-148 is protonated at both nitrogens and cannot be a hydrogen bond acceptor for Arg-168. Moreover, as shown in Table 4.2, the positive His-148 in the models by Marques *et al.* is significantly closer to the chromophore with a His-148(N)$\cdots$CHR(O_1) distance which is 0.4 Å shorter than the value in our as well as the X-ray structure. Finally, the bond lengths along the chromophore are significantly closer in our model to the X-ray values than the ones computed by Marques *et al.* We therefore believe that the overall better agreement of our structure with the crystallographic data is strong evidence that our model is most likely a closer representation of the real structure of GFP than the one where His-148 is protonated at both nitrogens.

In Fig. 4.12, we also compare the structure of the chromophore that we obtain in the three forms of GFP with other models available in the literature. We find that the structures of our chromophore for the neutral A and anionic B forms are in close agreement with the simulations denoted as (MM + QM) by Laino *et al.* [81]. These authors construct the missing force field for the chromophore to perform a first MM relaxation of the protein and, subsequently, refine the structural parameters by performing a DFT/BLYP optimization of the chromophore binding site only keeping the close residues with fixed heavy atoms at the boundary. The structure of the chromophore in our model is remarkably similar to the one by Laino *et al.* [81] for both the neutral and the anionic B form, and the hydrogen-bond distances of the chromophore to the closeby residues are also in reasonable agreement (see Table 4.2). A detailed comparison with the hydrogen-bond distances of the (MM + QM) model is however not proper as only few residues were kept

in their QM optimization. Even though all distances listed in Table 4.2 do not significantly change in the QM refinement, an exception is the Thr-203(O)$\cdots$CHR(O_1) distance which in their MM simulation is 2.80 Å but becomes significantly larger, 3.28 Å, in the subsequent partial QM optimization.

Finally, our results for the anionic I and B forms are in reasonable agreement with the CASSCF/MM calculations by Sinicropi *et al.* [13] who relax the chromophore structure within CASSCF together with the position and the orientation of three closeby classical waters. The structure of the remaining of the protein in the neutral A form is kept to the original crystallographic coordinates while the I form is obtained starting from the A form and manually reorienting the residues Ser-205 and Glu-222 to form the expected hydrogen bond network (see Fig. 4.5). The B form is derived from the I form by relaxing Thr-203 in a conformation to form hydrogen bond with the phenolic oxygen of the chromophore (see Fig. 4.7). Despite the simplicity of their QM/MM embedding scheme and lack of complete relaxation, the few available hydrogen-bond distances are in agreement with the values we find, as shown in Table 4.2. For the chromophore, we find that the degree of bond-length alternation is slightly larger in CASSCF as compared to the DFT calculations, but the overall agreement is rather good.

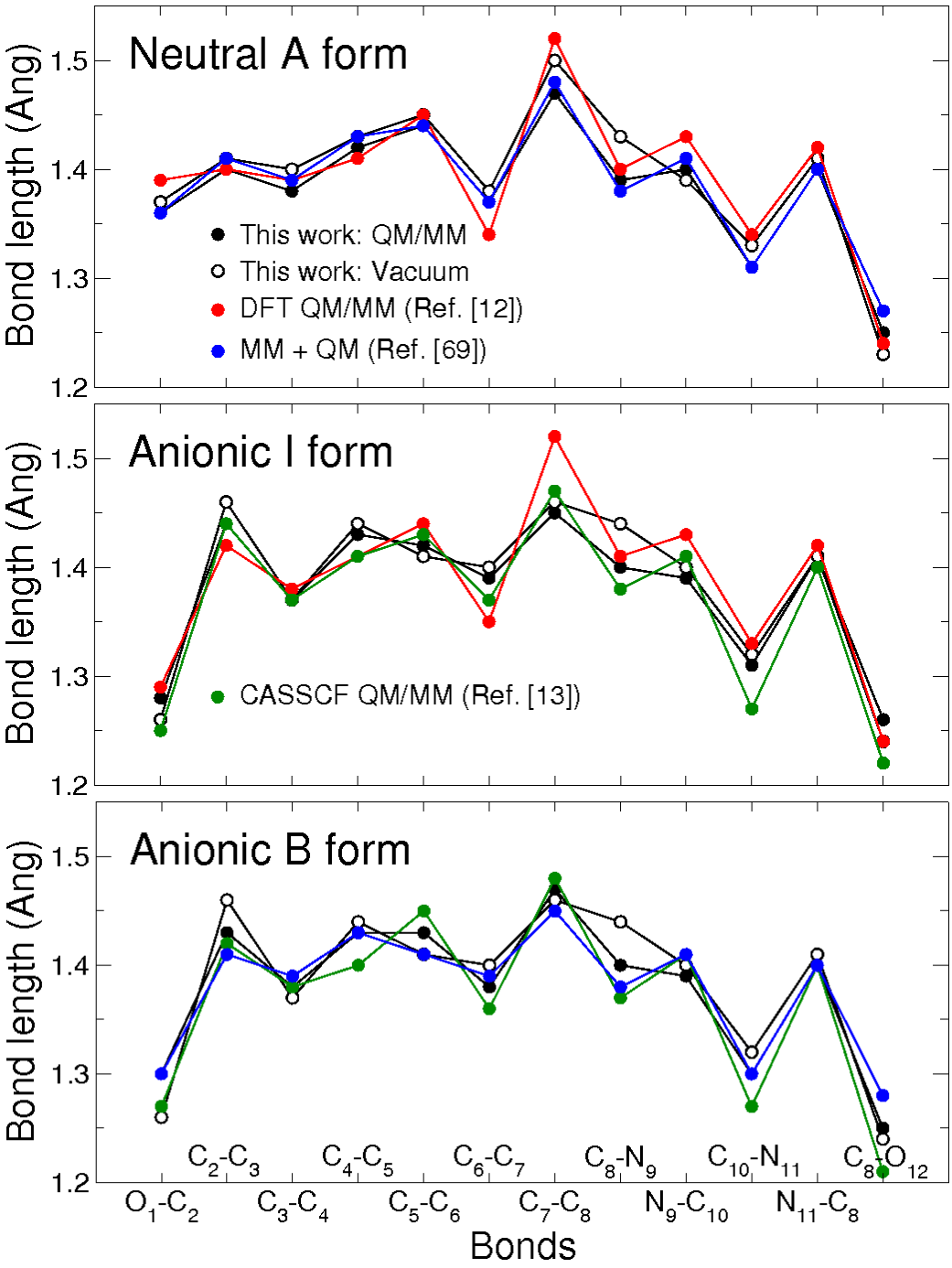

Figure 4.12: Bond lengths of the chromophore of the A, I, and B forms of GFP. We compare our QM/MM results to DFT/LDA [12] and CASSCF QM/MM [13] simulations, and to MM simulations followed by partial QM relaxation [69]. The neutral and anionic geometries optimized in vacuum are also shown. The bonds of the central bridge are C_5—C_6 and C_6—C_7.

	Arg-96(N) ⋮ $CHR(O_{12})$	Gln-94(N) ⋮ $CHR(O_{12})$	$CHR(O_1)$ ⋮ W1	His-148(N) ⋮ $CHR(O_1)$	Thr-203(O) ⋮ $CHR(O_1)$
			Neutral A form		
This work	2.76	2.86	2.56	3.21	–
DFT QM/MM	2.74	3.18	2.77	2.81	–
MM + QM	2.71	3.05	2.61	N/A	–
X-ray	2.71	3.08	2.64	3.22	–
			Anionic I form		
This work	2.77	2.85	2.78	2.96	–
DFT QM/MM	2.72	3.31	2.83	2.67	–
CASSCF QM/MM	N/A	N/A	2.71	N/A	–
			Anionic B form		
This work	2.76	2.95	2.79	3.02	2.76
CASSCF QM/MM	N/A	N/A	2.69	N/A	N/A
MM + QM	2.63	3.00	2.75	2.80	2.80[a]

[a] 2.80 Å is the MM while 3.28 Å the (MM + QM) value [82].

Table 4.2: Most representative hydrogen-bond distances in Å between the hydrogen donor and acceptor (D $\cdots$ A) for close residues to the chromophore (CHR). We compare our DFT/PBE QM/MM results to other models available in the literature as the DFT/LDA [12] and CASSCF QM/MM [13] simulations and the MM simulations followed by partial DFT/BLYP relaxation (QM+MM) [69]. Note that, for the bond $CHR(O_1) \cdots$ W1, the chromophore is a donor in the neutral A form and an acceptor in the I and B forms, and that Thr-203(O) only forms a hydrogen bond to the chromophore in the B form.

4.3 TDDFT/MM absorption spectra

We compute the vertical low-lying singlet excited states of the chromophore of GFP in the protein environment for the structures of the A, I and B forms presented in the previous Section using a TDDFT/MM approach and the Gaussian03 code [53]. We perform all-electron linear-response TDDFT calculations in the presence of the external field of the classical protein environment. Within Gaussian03, the MM atoms are treated as unscreened point charges which we place at the positions corresponding to the optimal geometry that we obtained in our QM/MM simulations. For the MM atoms, we use the same partial charges as in the QM/MM simulation. We employ the BLYP and B3LYP exchange-correlation functionals and a cc-pVTZ Gaussian basis set. We always compute at least the lowest five solutions as the state with the largest oscillator strength is not always the lowest.

The linear-response TDDFT results for wild-type GFP in its three forms are summarized in Table 4.3. We compute two sets of excitation energies corresponding to the chromophore in the protein environment (Protein) and to the same chromophore without the surrounding protein (Vacuum). Comparing the excitation energies with and without the protein environment allows us to access the polarization effects of the protein on the electronic states of the chromophore. Moreover, a comparison with the excitations for the chromophore geometry optimized in vacuum (see Chapter 3) reveals how the changes in the structure of the chromophore due to the presence of the protein affects the excitation energies. In all cases, we list the lowest two singlet excitations, minus the Kohn-Sham energy of the highest occupied molecular orbitals (HOMO) as the DFT ionization continuum begins at minus the value of the HOMO orbital energy [84] and the Kohn-Sham gap between the lowest unoccupied and the highest occupied molecular orbitals.

Comparison with experiments

We first give a general look at the results focusing on the comparison with the experimental absorption maxima [85, 86]. For all three forms and both exchange-correlation functionals, the excitation energy with the largest oscillator strength is the lowest singlet state with a dominant $\pi \rightarrow \pi^*$ (HOMO $\rightarrow$ LUMO) character. For the neutral A form, the BLYP and the B3LYP excitation energies in the protein nicely bracket the experimental value with BLYP giving an energy 0.03 eV lower and B3LYP 0.18 eV higher than the experiments. The oscillator strength (about 0.6) and the HOMO $\rightarrow$ LUMO contribution to the character of the excitation is comparable when the BLYP and B3LYP functionals are used. The ionization threshold for both function-

als is significantly higher than the excitation energies, not to pose a problem. The effect of polarization by the protein environment is small and only amounts to a red shift of about 0.04 eV in the excitation energy. We may therefore regard this result as a success of linear-response TDDFT in describing the experimental spectrum of the neutral form of GFP but we will return to this point later when interpreting the TDDFT results for all three forms of wild-type GFP.

The TDDFT excitations for the anionic forms of GFP are instead not in good agreement with the experimental values. For the I form, BLYP and B3LYP overestimate the experimental excitation energy by 0.37 and 0.55 eV, respectively, yielding excitation energies not too dissimilar to the neutral case. By comparing with the excitations computed in the absence of the protein, we note that the inclusion of the protein environment yields a blue shift of 0.05 and 0.18 eV for the BLYP and B3LYP functionals, respectively, that is, a shift in the opposite direction than for the neutral form. More importantly, we observe that the inclusion of the protein significantly raises the DFT ionization threshold so that the excitations in the protein are 1.06 and 1.68 eV lower than the BLYP and the B3LYP ionization threshold, respectively. Therefore, while the meaning of the excitation energies in vacuum is questionable as they lie above the ionization threshold, the problem represented by the general underestimation of the ionization threshold in DFT disappears when the protein is included.

For the B form, the agreement of TDDFT with experiment is equally poor as for the I form. TDDFT/BLYP and B3LYP overestimate the experimental value by 0.30 and 0.49 eV, respectively. Also for the B form, the inclusion of the protein environment yields a blue shift 0.14 and 0.45 eV for the BLYP and B3LYP functionals, respectively. Differently from the case of the I form, only the use of the B3LYP functional yields an excitation in the protein well below the excitation threshold while BLYP gives an excitation in the protein which is 0.42 eV higher than the BLYP ionization threshold. Moreover, within TDDFT/BLYP, the two lowest singlet excitations are nearly degenerate with a comparable oscillator strength (0.43 and 0.32) and character of the excitations. This degeneracy is lifted when using the B3LYP functional which moves the second singlet 0.57 eV higher than the lowest singlet. For both anionic forms, the use of the B3LYP functional always yields a lowest state with a higher oscillator strength and a more definite single-excitation (HOMO $\rightarrow$ LUMO) character than when the BLYP functional is employed.

On the positive side, we note that TDDFT correctly predicts the expected blue shift in going from the I to the B form. In the B form, the ground state is further stabilized with respect to the excited state by the additional hydrogen bond of the phenolic oxygen of the chromophore with Thr-203, and one would

expect a larger excitation energy than in the I form. In particular, BLYP and B3LYP predict the excitation energy of the B form to be higher than the I form by 0.06 and 0.07 eV, respectively. With the available experimental data, we can only estimate an upperbound (0.13 eV) to the difference between the excitations of the B and the I form: For the I form, the location of the absorption maximum is not available and we therefore use the energy of the 0-0 transition, which is smaller than the absorption maximum by an amount equal to the Stokes' shift.

Finally, we note that the mixed agreement we find between TDDFT and experiments is at odds with the good performance observed by Marques *et al.* [12] who obtain 3.05 eV and 2.65 eV as the TDDFT/LDA absorption maximum for the A and the I form, respectively. We believe that the positive picture emerging from their work is coincidental and due to the fact that their model for the I form is different from our structure as discussed in Section 4.2. For both the neutral A and the anionic I form, they compute the TDDFT excitations of the protein chromophore without the protein environment. As their chromophore for the A form is not too dissimilar from our structure, we know from our calculations that computing the excitation without the protein environment will indeed yield a TDDFT excitation very close to the experimental value. On other hand, the chromophore of their I form is significantly different from our geometry due to how they protonate all histidine residues in the protein. Apparently, this geometry in vacuum gives a TDDFT excitation close to the experimental value for the I form. For clarity, we point out again that even though they build the I form by deprotonation of the neutral form, they refer to their structure as the B form and therefore find a particularly good agreement with experiments as they compare with the absorption maximum of the B form (2.63 eV).

Geometrical and electrostatic effects of the protein environment

In Table 4.4, we summarize the TDDFT excitation spectra for the chromophore optimized in the gas phase and in the protein environment for the three forms of wild-type GFP, and compute the spectral shift, $\Delta E_{protein}$, due to the interaction of the chromophore with the protein. The spectral shift can be decomposed in two contributions due to the geometrical change in the chromophore geometry and to the electrostatic effect of the environment on the excitation:

$$\Delta E_{protein} = \Delta E_{geom} + \Delta E_{elect} \, .$$

The geometrical shift (ΔE_{geom}) is computed as the difference of the excitation of the protein chromophore without the protein environment and the

excitation of the chromophore optimized in the gas phase. The electrostatic shift (ΔE_{elect}) is obtained as the difference of the excitation energies of the protein chromophore with and without the protein environment.

For the neutral A form, BLYP and B3LYP yield a total red shift due to the chromophore-protein interaction of -0.08 and -0.19 eV, respectively. For both functionals, the shift due to the electrostatic interaction with the MM atoms is very small (less than -0.04 eV) while B3LYP sees a larger red shift due to geometrical changes than the one obtained with the BLYP functional. As we have seen in the previous Section, the degree of bond alternation for the neutral and the anionic forms is more marked for the chromophore optimized in vacuum than in the protein. As expected, this translates in a red shift in the excitation energies when going from the gas-phase chromophore to the protein chromophore, a shift which is significantly larger (-0.16 eV) when using the B3LYP functional.

For the anionic I form, BLYP and B3LYP predict no protein shift (0.01 eV) as the geometrical red shift and the electrostatic blue shift cancel. The reference protein shift of -0.09 eV is estimated as the difference between the energy of the 0-0 transition in the protein [7] and the absorption maximum of a smaller chromophore in the gas phase [58]. Therefore, it represents a lowerbound to the true shift, as the excitation of a larger chromophore in the gas phase (with the same dimensions as the one optimized in the protein) will be smaller than the excitation of the smaller model chromophore, while the absorption maximum of the I form is larger than its 0-0 transition. In the anionic B form, BLYP and B3LYP yield comparable small blue shifts of 0.07 and 0.08 eV, respectively. For both functionals, the geometrical red shift is very close to the values obtained for the anionic I form since the I and B chromophores have very similar geometries. However, due to the further stabilization of the ground state in the B form discussed above, the electrostatic blue shift is larger in the B than the I form, yielding a net protein blue shift. The experimental estimate of the shift is 0.04 and again represents a lowerbound as we are using the absorption maximum for a smaller model chromophore to estimate the excitation in the gas phase.

Understanding the mixed performance of TDDFT

The mixed success of TDDFT, which appears to be accurately describing the excitation of the neutral form but not of the anionic forms of wild-type GFP, may be due to several factors which we now analyze separately. The first obvious culprit is the use of TDDFT in the adiabatic approximation and we will therefore try to understand if any of its commonly known problems can also affect our results. However, we need to keep in mind that, whether

there is a problem with TDDFT, it must only affect the anionic and not the neutral form as the neutral appears to be correctly described by adiabatic TDDFT.

We begin with the underestimation by DFT of the ionization threshold and, in Table 4.3, list for all systems minus the Kohn-Sham energy of the HOMO orbital, which indicates the start of the TDDFT continuum. We observe that for the neutral chromophore with and without the protein environment, the ionization threshold is well above the lowest excitation energy and therefore does not pose a problem. On the other hand, as in the case of the anionic model chromophores in the gas phase (Chapter 3), the TDDFT excitation energies of the I and B anionic chromophores in vacuum lie above the ionization threshold and their meaning is therefore questionable. The inclusion of the protein environment significantly raises the ionization threshold of the I form above the lowest singlet excitation energy while, for the B form, only the use of the B3LYP functional brings the ionization threshold above the relevant excitation energy. However, even though the TDDFT continuum is now higher than the excitation, TDDFT overestimates the experimental excitation energies of the I and B forms by 0.4-0.5 eV. We also note that the TDDFT error in the protein is comparable to the error in the gas phase for the smaller anionic chromophore, where TDDFT overestimates the excitation by 0.26 and 0.46 eV when using BLYP and B3LYP, respectively. Therefore, the position of the ionization threshold does not appear a relevant factor in understanding why the performance of TDDFT is significantly poorer in describing the anionic than the neutral form of GFP.

Adiabatic TDDFT is also known to perform rather poorly if the excitations are characterized by significant charge transfer. To understand whether this is a problem, we can compare the TDDFT excitations and the Kohn-Sham eigenvalue differences since these two quantities will become equal if the excitation is characterized by charge transfer. We note that this analysis holds when no exact exchange is used in the functional as only in this case the eigenvalue differences are a close representation of excitation energies since all orbitals see the same potential and therefore a constant number of electrons. In the presence of exact exchange, the virtual orbitals see a different potential and a different number of electrons than the occupied ones, and for instance the HOMO-LUMO gap will be closer to the difference between ionization potential and electron affinity.

In the three forms of GFP, the singlet TDDFT excitations with the largest oscillator strength are the lowest in energy and have a dominant HOMO $\rightarrow$ LUMO character, and should therefore be compared with the Kohn-Sham HOMO-LUMO gaps which are also listed in Table 4.3. For the neutral A and anionic I forms, the BLYP HOMO-LUMO gap is always smaller by about

1 eV than the lowest TDDFT excitation. Therefore, since TDDFT significantly corrects the Kohn-Sham gap, we can infer that the excitations are not characterized by strong density transfer. For the B form, the comparison with the BLYP Kohn-Sham gap is not as straightforward since the two lowest excitations are nearly degenerate at 2.93 and 2.94 eV, and dominated by the (HOMO-3) → LUMO contribution even though the HOMO → LUMO transition is rather large: The HOMO-LUMO BLYP gap is 1 eV smaller than the lowest excitation while the (HOMO-3)-LUMO eigenvalue difference is equal to 2.90 eV and therefore very close to the excitation energy. This could be a sign of potential problems of the TDDFT description of the B form. In Fig. 4.13, we also show the difference between the ground state density and an estimate of the TDDFT excited state density for the neutral A and the anionic I form with and without the protein environment. The anionic excited state density computed without the protein environment appears to extend in the tails of the chromophore as compared to the ground state one. However, this slight charge transfer to the tails disappears when the protein environment is included and no striking difference can be observed between the neutral and the anionic case.

In summary, the indicators given by the position of the excitation with respect to minus the HOMO Kohn-Sham eigenvalue and to the Kohn-Sham eigenvalue difference are rather similar in the neutral and the anionic forms (in particular the I form) of wild-type GFP and it is therefore not evident why TDDFT should give an excellent agreement with experiments for the neutral form but not the anionic forms. A possible reason for the failure of TDDFT could be that the excitation has a significant double- or higher-excitation character in the case of the anion but not in the neutral form. We come back to this point when analyzing the quantum Monte Carlo results in the next Section as these results seem to indicate that the character of the excitation is rather similar in the two cases. Finally, it is important to stress that, when computing the excitation energies with a QM/MM scheme, the MM atoms only affect the electronic states of the QM part via an electrostatic polarization field. Therefore, all the residues which are hydrogen bonded to the chromophore cannot in reality forming a proper "bond". In Section 4.5, we analyze the effect of extending the QM part of the simulation to include relevant closeby residues.

Figure 4.13: Difference between the DFT/B3LYP ground and excited state densities for the protein chromophore of the neutral A form (top) and the anionic I form (bottom) without (left) and with (right) the protein environment. The isosurface corresponds to a value of -0.001 in red and $+0.001$ in blue. A negative (red) value coresponds to the excited state density being larger than the ground state one.

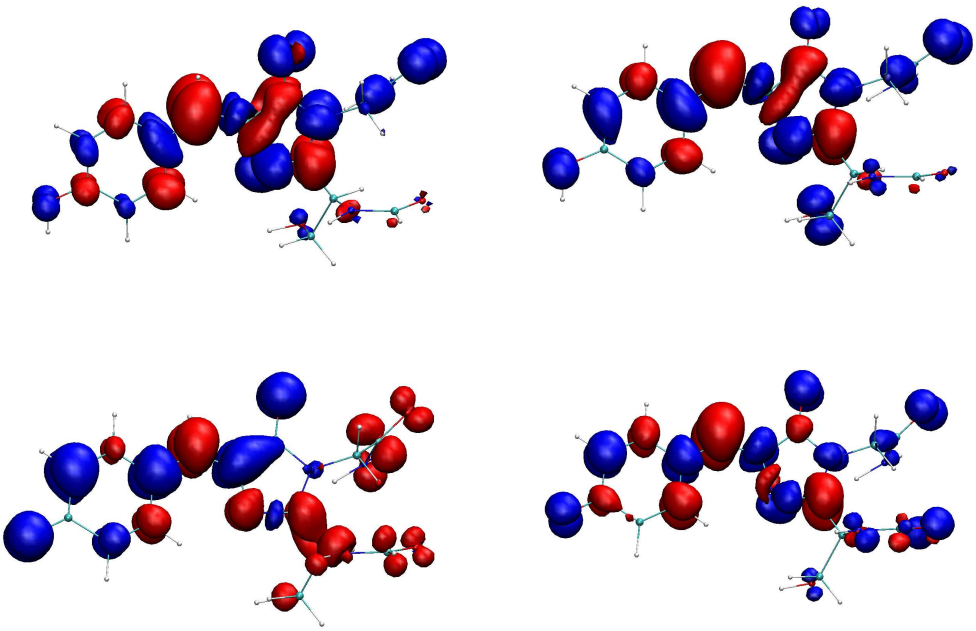

Table 4.3: TDDFT/BLYP and B3LYP excitation energies (eV) and oscillator strengths (in parenthesis) for the A, I, and B forms of GFP with (Protein) and without (Vacuum) the protein environment, computed using a cc-pVTZ basis. The geometries are optimized in the presence of the protein. The dominant electronic transitions and their contributions in parenthesis (if > 0.1) are also listed. Only the lowest two singlet excitations are given together with minus the Kohn-Sham energy of the highest occupied molecular orbitals ($-\epsilon_{\mathrm{HOMO}}$), which corresponds to the ionization threshold in DFT. Also the Kohn-Sham gap between the lowest unoccupied and the highest occupied molecular orbitals ($\Delta\epsilon_{\mathrm{HL}}$) is listed.

	Vacuum		Protein	
	BLYP	B3LYP	BLYP	B3LYP
	Neutral A form (Expt. 3.05 eV)			
$S_0 \rightarrow S_1$	3.07(0.63)	3.26(0.58)	3.02(0.65)	3.23(0.64)
	H→L(0.47)	H→L(0.58)	H→L(0.45)	H→L(0.58)
	H-1→L(0.30)	H-1→L(0.19)	H-1→L(0.23)	H-2→L(0.20)
	H-4→L(0.12)	H-2→L(0.12)	H-3→L(0.21)	
$S_0 \rightarrow S_2$	3.19(0.02)	3.66(0.22)	3.14(0.01)	3.48(0.18)
	H-3→L(0.69)	H-2→L(0.67)	H-4→L(0.69)	H-2→L(0.66)
	H-4→L(0.12)	H→L(0.15)		H→L(0.17)
$\Delta\epsilon_{\mathrm{HL}}$	2.10	3.36	2.05	3.29
$-\epsilon_{\mathrm{HOMO}}$	5.10	6.01	5.57	6.47
	Anionic I form (Expt. ≈2.50 eV)			
$S_0 \rightarrow S_1$	2.82(0.60)	2.87(0.67)	2.87(0.71)	3.05(0.96)
	H→L(0.43)	H→L(0.54)	H→L(0.52)	H→L(0.58)
	H→L+2(0.29)	H→L+1(0.31)	H-2→L(0.20)	
	H→L+1(0.27)			
$S_0 \rightarrow S_2$	3.03(0.12)	3.19(0.28)	3.08(0.12)	3.48(0.001)
	H→L+2 (0.63)	H→L+1(0.62)	H-2→L(0.60)	H-1→L(0.68)
	H→L(0.16)	H→L+2(0.15)	H-4→L(0.19)	H-2→L(0.14)
	H→L+4(0.13)		H-5→L(0.19)	
$\Delta\epsilon_{\mathrm{HL}}$	1.71	2.79	1.82	2.92
$-\epsilon_{\mathrm{HOMO}}$	1.05	1.79	3.93	4.73

Continues on the next page

Continues from the previous page

	Vacuum		Protein	
	BLYP	B3LYP	BLYP	B3LYP
	Anionic B form (Expt. 2.63 eV)			
$S_0 \rightarrow S_1$	2.79(0.57)	2.87(0.75)	2.93(0.43)	3.12(0.90)
	H→L(0.43)	H→L(0.55)	H-3→L(0.46)	H→L(0.59)
	H→L+2(0.30)	H→L+1(0.25)	H→L(0.39)	
	H→L+1(0.26)		H-1→L(0.14)	
	H→L+5(0.10)		H→L+1(0.11)	
$S_0 \rightarrow S_2$	3.01(0.13)	3.23(0.18)	2.94(0.32)	3.69(0.04)
	H→L+2(0.62)	H→L+1(0.65)	H-3→L(0.53)	H-1→L(0.67)
	H→L (0.17)	H→L(0.16)	H→L(0.34)	H-3→L(0.10)
	H→L+3(0.15)		H-1→L(0.11)	
$\Delta\epsilon_{\mathrm{HL}}$	1.69	2.76	1.92	3.05
$-\epsilon_{\mathrm{HOMO}}$	1.06	1.81	2.51	3.31

Table 4.4: TDDFT/BLYP and B3LYP excitation energies in eV computed for the protein chromophore without ($CHR_{protein}$) and with (GFP) the inclusion of the protein environment, and for the chromophore optimized in the gas phase (CHR_{gas}). The spectral shifts due to the geometrical changes in the chromophore ($\Delta E_{geom} = CHR_{protein} - CHR_{gas}$) and to the electrostatic effect of the environment ($\Delta E_{geom} = GFP - CHR_{protein}$) are listed together with the total protein shift, $\Delta E_{protein} = \Delta E_{geom} + \Delta E_{elec}$.

	CHR_{gas}	$CHR_{protein}$	GFP	ΔE_{geom}	ΔE_{elec}	$\Delta E_{protein}$
			Neutral A form			
BLYP	3.10	3.07	3.02	-0.03	-0.05	-0.08
B3LYP	3.42	3.26	3.23	-0.16	-0.03	-0.19
Expt.	–	–	3.05	–	–	–
			Anionic I form			
BLYP	2.86	2.82	2.87	-0.04	0.05	0.01
B3LYP	3.04	2.87	3.05	-0.17	0.18	0.01
Expt.	≈2.59	–	2.50	–	–	-0.09
			Anionic B form			
BLYP	2.86	2.79	2.93	-0.07	0.14	0.07
B3LYP	3.04	2.87	3.12	-0.17	0.25	0.08
Expt.	≈2.59	–	2.63	–	–	0.04

4.4 QMC/MM excitation energies

We now compute the vertical excitations of the three forms of wild-type GFP in quantum Monte Carlo. As for the chromophore models in vacuum, we only compute the bright $\pi \rightarrow \pi^*$ (HOMO $\rightarrow$ LUMO) transition, which has the largest oscillator strength and therefore corresponds to the maximum absorption of GFP. The QMC excitations are given by the difference of the excited state S_1 and the ground state S_0 total energy, which are now computed in the presence of the MM protein environment. To perform a quantum Monte Carlo calculation in the presence of the protein environment, we need to include the environmental effects in the actual QMC calculation as well as in the construction of the starting many-body wave function.

The inclusion of a classical MM environment in the QMC calculation is straightforward as it amounts to include an additional external potential which we simply represent on a grid encompassing the QM component. The potential is computed using the positions of the MM atoms as obtained in the DFT/PBE QM/MM calculation to setup the protein model, and the partial charges of the MM atoms are screened as in the CPMD code. For the setup of the determinantal component of the QM wave function, we employ the GAMESS code which allows us to include the MM atoms as screened point charges via the effective fragment potential module of GAMESS. We only consider the MM atoms which lie within $R_c = 12$ $\mathring{A}$ of the QM/MM boundary and leave the positions of the MM atoms at the optimal coordinates of the DFT/PBE QM/MM calculation. As GAMESS only allows a particular way to screen the point charges which is different than the one in the CPMD code, we verify below that the use in QMC of the MM potential compatible with the screening in GAMESS yields equivalent QMC/MM excitation energies than when using the CPMD screening.

All QMC calculations are performed with Hartree-Fock semi-relativistic energy-consistent pseudopotentials specifically constructed for use in QMC, and the corresponding cc-pVDZ basis sets [67]. We use Jastrow-Slater wave functions where the determinantal component is determined within GAMESS in the presence of the MM atoms and the Jastrow correlation factor is subsequently optimized by energy minimization using the Hamiltonian with the additional potential due to the MM environment. In the Jastrow factor, we only include electron-electron and electron-nucleus correlation terms as the additional electron-electron-nucleus terms would significantly slow down the calculations. Moreover, in Chapter 3, we have seen for the small anionic model chromophore that neglecting the three-body terms in the Jastrow factor does not significantly affect the DMC excitation energies. For the determinantal components of the ground and excited states, we use the four

Table 4.5: Vertical excitation energies (eV) for the neutral A and the anionic I and B forms of wild-type GFP computed within VMC, DMC, and TDDFT/BLYP and B3LYP. The experimental energies correspond to the absorption maxima for the A and B forms and the 0-0 transition for the I form. The CASPT2 results are from Ref. [13]. For the QMC results, the statistical error on the last figures is given in parenthesis.

	A form	I form	B form
Excitation energy (eV)			
Expt.	3.05	2.50	2.63
VMC	3.96(09)	3.34(09)	3.71(9)
DMC	3.66(11)	3.16(11)	3.15(12)
TDDFT/BLYP	3.02	2.87	2.93
TDDFT/B3LYP	3.23	3.05	3.12
CASPT2 [13]	–	2.65	2.81

determinants resulting from a two-state SA-CASSCF(2,2) calculation with equal weights. Again, this choice for the wave function is both motivated by the necessity to reduce the computational demands of the calculation, and by the observation for the small anionic model chromophore that the inclusion of a larger active space as well as the reoptimization of the orbitals in the presence of the Jastrow factor only affects the DMC excitation energy by at most 0.1-0.2 eV.

In Table 4.5, we summarize the VMC, DMC and TDDFT vertical excitations for the neutral A and the anionic I and B forms of wild-type GFP and compare them with experiments as well as CASPT2 results available in the literature [13]. We observe that the DMC excitation energies are always lower than the VMC values as in the case of the model chromophores in vacuum, but remain too high as compared to experiments. For all three forms, DMC overestimates the excitation energies by as much as 0.5-0.6 eV while the available CASPT2 results appear to be in significantly better agreement with experimental values.

For the anionic I and B forms, the DMC excitation is higher than experiments by roughly the same amount as in vacuum where, for consistency, we are considering the gas-phase excitation computed with the same type of wave function, that is, a SA-CASSCF(2,2) determinantal component with no reoptimized orbitals. Based on our experience in the gas phase, we may

expect that reoptimizing the determinantal component in the presence of the Jastrow factor will affect the excitation energy by a small amount of the order of 0.1-0.2 eV and therefore not sufficient to bring the DMC excitations in agreement with experiments. Again similarly to case of the gas-phase anion, we find that the DMC excitation is reasonably close to the linear-response TDDFT excitation. For the neutral form, the DMC excitation energy is also overestimated as compared to experiments by roughly the same amount as in the anionic forms and therefore in disagreement with the the TDDFT value which instead agrees perfectly with the experimental absorption maximum.

On the positive side, the shift in the excitation energy following deprotonation of the neutral form is well reproduced by DMC while TDDFT fails in describing how the energy correlates with the charge state of the chromophore. The DMC difference between the neutral and the anionic I form is roughly (0.50 ± 0.16) eV, as compared to the experimental value of 0.50 eV. The DMC difference between the neutral and the anionic B form is instead (0.49 ± 0.16) eV while the experimental shift is 0.42 eV.

Understanding the mixed performance of quantum Monte Carlo

Since this work represents the first application of quantum Monte Carlo methods to the computation of the excitations of a large biosystem via a mixed quantum/classical approach, we are still in the process of understanding the reasons of its mixed performance. It is certainly positive and extremely important that DMC can well reproduce the difference between the excitation energies of the various forms of GFP, especially since TDDFT fails to see significant differences between the different charge states of the chromophore. However, an absolute error of 0.5-0.6 eV is certainly too large and at variance with the good performance we have previously observed with the present QMC approach when describing the excitation energies of small prototypical gas-phase molecules [61–64].

There are essentially three possible errors in a DMC calculation, the time-step error, the localization error in treating the non-local pseudopotential and, finally, the fixed-node error. None of this errors is actually small and we always rely in their cancellation when computing the difference of total energies. As the shift between the neutral and anionic forms are well reproduced, the source of error must be affecting all three forms in the same manner by raising the absolute value of all excitations. The impact of the time-step and localization errors have been already analyzed in detail for the small model chromophore in vacuum. The system is now slightly larger and we are using the same time step of 0.055 a.au. as for the smaller model chromophores. However, the algorithms we are using for the DMC calculation

is characterized by small time-step errors and in going to the larger system, we have added the "same" type of atoms (C, N, O and H) and consequently the "same" type of electrons. Therefore, if the time-step extrapolated excitation energy is indistinguishable within statistical error from the excitation at $\tau = 0.055$ a.u. for the smaller chromophore, we expect the same to be true for the protein chromophores. The same analysis holds for the impact of the localization error on the excitations. For the smaller chromophore, the use of a three-body Jastrow factor does not significantly affect the DMC excitation energies within statistical error, so the effect of localizing the non-local potential appears small. Moreover, the DMC excitation energy of the small chromophore is also unchanged when a different algorithm beyond the localization approach is employed. Therefore, as the protein chromophore is rather similar to the smaller chromophore, we would also expect that time-step error and localization error are under control.

The most serious source of error is the fixed-node error which can only be controlled by using better many-body wave functions. In the small anionic chromophore, all attempts to use more sophisticated wave functions have only marginally lowered the DMC excitation energy by 0.1-0.2 eV with respect to the result obtained using the simplest unoptimized SA-CASSCF(2,2) wave function. For the protein system, we find that using the same simple ansatz for the many-body wave function yields excitation energies which are all overestimated with respect to experiments by 0.5-0.6 eV. If the source of error were the fixed-node error, the fact that the error is the same for both the anionic and the neutral forms would likely mean that the multi-configurational nature of the excitation is the same for both charge states and that we are missing some important static correlations in our wave function. This is however at odds with the observation that linear-response adiabatic TDDFT appears to be able to describe the neutral but not the anionic form. In the previous Section, when trying to understand the mixed performance of TDDFT, we attributed the problems of TDDFT with the anionic form to the possible double- and higher-excitation character of this excitation as charge-transfer or underestimation of the ionization threshold do not seem to play a role. Now, from the DMC results with the SA-CASSCF(2,2) wave function, we can infer that, whatever the multi-configurational nature of the excitation, it should be rather similar for both charge states, a fact which leaves unexplained the mixed success of TDDFT. Therefore, what is the source of error in the QMC calculations? Certainly, the most likely candidate remains the fixed-node error and, also in the protein, additional extensive (and computational costly) studies must be performed with larger active spaces in the determinantal component of the wave functions to really ensure that the CASSCF(2,2) ansatz is not too simplistic.

Another obvious source for our problems is the QM/MM model itself, not being representative of the real protein structure of wild-type GFP and/or of the chromophore-protein interaction. We have already seen how small variations in such a complex model can yield very different excitation energies as in the study by Marques *et al.* who find perfect agreement of adiabatic TDDFT with experiments by protonating (we believe) incorrectly a closeby residue to the GFP chromophore. To understand whether our protein structure is realistic, one could study different mutants as the availability of more theoretical data would allow us to better understand correlations between models and performance of TDDFT and QMC. This rather challenging and long route has not been undertaken in this thesis.

It is instead conceptually simpler even though computationally rather costly to check whether the protein-chromophore interaction is poorly described in our QM/MM model. In the calculations of the electronic states of the QM chromophore in the MM environment, the interaction between the QM and the MM part is purely electrostatic and the environment is not polarizable in response to the excitation of the quantum chromophore. Moreover, as already pointed out above, the complete quantum nature of hydrogen bonds cannot be described in terms of electrostatic interactions only. To understand the limitation of our QM/MM description of the protein-chromophore interaction, we can simply enlarge the QM part to include close residues. One first step in this direction is described in the next Section.

4.5 Enlargement of the QM part

We focus here on the effect of enlarging the QM part of our QM/MM calculation of the I form of wild-type GFP. In Fig. 4.14, we show how we enlarge the QM part of our system by including the residues Arg-96 and Gln-94 which are hydrogen bonded to the oxygen of the imidazole ring. We choose to include these two residues as their presence stabilizes the excited state more than the ground state since, in the excited state, charge is displaced from the phenolic to the imidazole ring. We therefore expect that, if their approximate MM treatment is responsible for the poor agreement of the theoretical excitations with experiments, a correct inclusion in the QM part of these two residues should lower the excitation in the direction of the experimental numbers. Also we note that residue Arg-96 is positively charged and has therefore an important role in stabilizing the negative charge of the anionic chromophore.

The structure of the enlarged QM/MM system is again equilibrated within DFT/PBE QM/MM with the CPMD code. Starting from the structure of the I form obtained in Section 4.1.2, we perform the following steps:

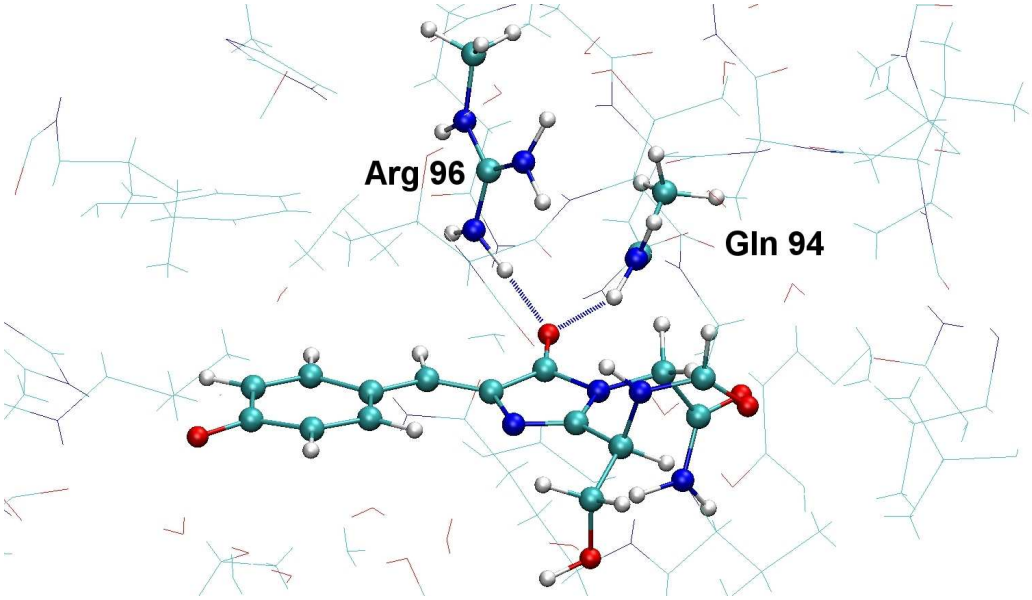

Figure 4.14: Enlarged QM structure of the I form of wild-type GFP. The original chromophore as well as residues Arg-96 and Gln-94 are emphasized and are now included in the QM calculation.

1) 0.7 ps of molecular dynamics at constant energy (NVE). The temperature monotonically grows and the QM energy becomes lower.
2) 1 ps of molecular dynamics with $f = 0.99$ of the MM system only with the QM atoms at fixed positions.
3) 0.07 ps of molecular dynamics with $f = 0.999$.
4) 0.14 ps of molecular dynamics with $f = 0.995$.

At the end of the simulation, the bond lengths along the chromophore are oscillating less than 0.0005 Å, and the hydrogen-bond distances between the imidazole oxygen and hydrogens of Arg-96 and Gln-94 vary less than 0.005 Å. In figure 4.14, we show the behavior of the hydrogen-bond distances between the imidazole oxygen and the hydrogen donated by Arg-96(N) and Gln-94(N) during the first 0.05 ps of the equilibration process. The starting structure is the optimal structure obtained with the two residues treated as classical, where the two bonds lengths determined via an effective electrostatic interaction between the QM oxygen and the MM hydrogens are rather similar. Interestingly, as the two residues are included in the QM part and the hydrogen bonds treated fully quantum mechanically, the two bonds immediately start to differ. The bond of the oxygen with the hydrogen from the

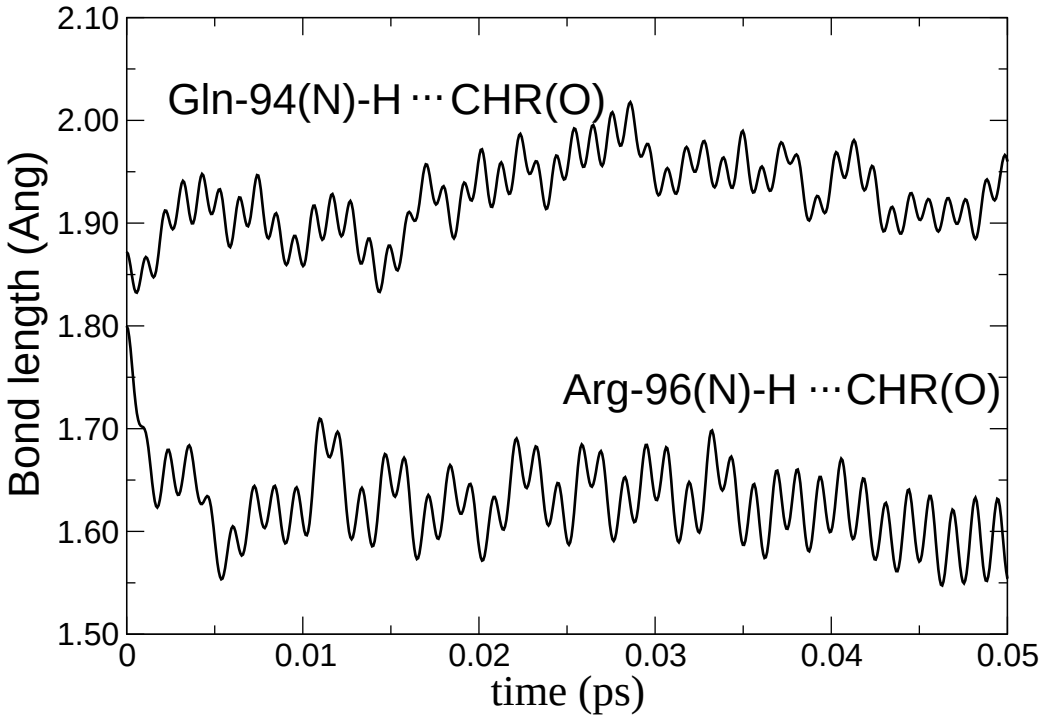

Figure 4.15: Bond lengths in Å between the imidazole oxygen and the hydrogen of the hydrogen-donor nitrogens of Arg-96 and Gln-94 which are included in the QM system. The first 0.05 ps of molecular dynamics are shown. The starting configuration is the optimal QM/MM structure obtained with the two residues treated as classical.

positive Arg-96 is stronger and the corresponding distance becomes shorter while the distance to Gln-94 lengthens. At the end of the dynamics run (not shown in the figure), the oxygen distance from the hydrogen of Arg-96 and Gln-94 settles at 1.66 and 1.88 Å, respectively. The Arg(N)$\cdots$O$_{12}$ and Gln(N)$\cdots$O$_{12}$ distances become 2.71 and 2.86 Å, respectively, that is, only different by -0.06 and 0.01 Å than the optimal values obtained when the two residues are treated classically. The distances along the chromophore are less affected by the inclusion of the two residues in the QM part of the system. For instance, the central bonds of the carbon bridge become 1.41 and 1.40 Å, as compared to the starting distances of 1.42 and 1.39 Å.

In Table 4.6, we show the TDDFT/MM excitation energies using the BLYP and the B3LYP functionals which should be compared to the corresponding TDDFT energies of Table 4.3. The reduction in the excitation energy is very small and equal to 0.07 and 0.03 eV for the BLYP and B3LYP

Table 4.6: TDDFT/MM excitation energies in eV for the I form of wild-type GFP where the QM part is enlarged to include Arg-96 and Gln-94. The results obtained with the BLYP and B3LYP functionals are shown.

	BLYP	B3LYP
$S_0 \rightarrow S_1$	2.80(0.57)	3.02(0.97)
	H→L(0.46)	H→L(0.58)
	H-2→L(0.39)	
$S_0 \rightarrow S_2$	3.00(0.26)	3.43(0.0023)
	H-2→L(0.53)	H-1→L(0.69)
	H→L(0.28)	
	H-3→L(0.17)	
	H-4→L(0.16)	
$\Delta\epsilon_{HL}$	1.82	2.91
$-\epsilon_{HOMO}$	4.14	4.93

functional respectively. For this enlarged QM/MM setup, we also perform quantum Monte Carlo calculations using a SA-CASSCF(2,2) wave function and including the MM environment. We find a DMC excitation energy of 3.2(1) eV which is compatible with the DMC result of 3.16(11) obtained with the smaller QM system. Therefore, even though enlarging the QM subsystem to include the residues Arg-96 and Gln-94 reduces the corresponding hydrogen bond distances to the imidazole oxygen, this geometrical changes are not sufficient to bring a significant shift in the TDDFT and DMC excitation energies.

4.6 Conclusions

We constructed the protein models of the neutral A and the two anionic I and B forms of wild-type GFP using a DFT QM/MM approach. For the neutral form, we started from the available X-ray structure while we obtained the I form by deprotonating and further relaxing the optimal neutral A form. The starting structure of the B form was built by homology from the X-ray structure of the mutant S65T of wild-type GFP.

We first performed a thourough structural analysis of the neutral and anionic forms comparing to available crystallographic data as well as other theoretical studies available in the literature. While all evidence points at

the correctness of our protein models, we find that the DFT QM/MM calculations by Marques *et al.* [12] yield structures significantly different from ours since these authors have incorrectly described the binding site of the chromophore by wrongly protonating a closeby residue. Since the incorrect description of the residues surrounding the chromophore affects its geometry and consequently its response to light, it follows that the perfect agreement of the corresponding TDDFT spectra for the isolated chromophore with the experiments is purely coincidental. This surprising outcome shows how difficult it is to correctly describe a complex biosystem and how easy to be mislead in believing the corretness of a given model when comparing to relatively few experimental numbers.

Starting from our optimal structures, we have computed the vertical TDDFT excitations of the chromophore of GFP in the protein environment for the neutral A, and the two anionic I and B forms. We find that adiabatic TDDFT appears to be accurately describing the excitation of the neutral form but significantly overestimates the excitations of the anionic forms by 0.3-0.4 eV depending on the functional. Consequently, the TDDFT shift in the excitation upon deprotonation is not correctly reproduced since TDDFT gives very similar excitations energy for all three forms of GFP. We tried to analyze the reasons of this mixed performance but it is not evident why TDDFT should give an excellent agreement with experiments for the neutral but not the anionic forms, in particular the I form. The excitations of these systems do not appear to be characterized by particular charge transfer and the underestimation of the DFT ionization threshold is not an issue once the chromophore is embedded in the protein environment. A possible explanation is that the excitation of the anionic form has a significant double- or higher-excitation character as compared to the neutral case. However, QMC calculations obtain a comparable description of the excitations of all three protein forms using the same, simple form of wave function, which indicates that, whatever the multi-configurational nature of the excitation, it should be rather similar for both charge states.

Finally, we have explored for the first time the use of QMC in describing the excitations of a chromophore in its protein environment and performed QMC/MM calculations of the excitation energies of the three forms of wild-type GFP using for the moment only a simple wave function. While the DMC excitation energies of both the neutral and the anionic forms are significantly higher than experiments by as much as 0.5-0.6 eV, the experimental shift between the different charge states of the chromophore is well reproduced by DMC. Even though it is reassuring that DMC can describe relative energies correctly, the DMC overestimation of the excitation energies is rather large and we have only begun to investigate the possible reasons for this error

such as shortcomings in the QM/MM description of the chromophore-protein interaction.

Chapter 5

Anion-π and π-π cooperative interactions

5.1 Introduction

The design of selective receptors of anionic species is a very active area of research within supramolecular chemistry due to the potential applications to catalysis, separation processes, and biomolecular systems [87]. Common neutral receptors bind the anion by hydrogen bonding or coordinate the anion at the Lewis acidic center of an organometallic ligand. As compared to cationic hosts, neutral receptors avoid the presence of competing negative counterions and are characterized by higher selectivity due to the directionality of the interactions. In recent years, the alternative route of anion complexation by neutral hosts via anion-π interactions has received considerable interest. The favorable binding interactions between an anionic species and an electron-deficient π-electron compound has been demonstrated in several theoretical studies [88–104] and experimental evidence of these attractive interactions is now cumulating from both X-ray structures [105–113] and solution data [114, 115].

Aromatic systems which have been investigated as potential anion-host candidates are either substituted benzene or electron-poor heteroaromatics such as triazines. Even though π-systems are expected to interact repulsively with anions, the presence of electron-withdrawing substituting atoms in the aromatic compounds modulates the reactivity inverting their natural electron-donor character. Hexafluorobenzene is the extreme example with a permanent quadrupole moment of similar magnitude as benzene but opposite sign, leading to attractive electrostatic interactions with electron-donor species [116]. In general, the interactions between the anion and the π-system

are predominantly of electrostatic and polarization nature, but dispersion forces and charge transfer [93,94,102] also contribute to the stability of these complexes.

Recently, experimental evidence of anion-π-π interactions has emerged from crystallographic studies [106–109] on synthesized coordination compounds based on the electron-deficient 1,3,5-triazine moieties, suggesting the possibility to enhance anion-π binding by π-π stacking. Particularly intriguing are the structural features of the nitrate-triazine-triazine complex of Ref. [109], as the two aromatic rings are staggered and not perfectly faced, and the nitrate ion is not parallel to the closest ring. The unusual asymmetrical configuration of the closely stacked triazines could be induced by the particular coordination within the compound, or governed by a subtle interplay between anion-π and π-π interactions [101, 104]. In the original paper, the compound was also investigated theoretically but, due to the poor description of dispersive interactions by the standard density functional theory approach employed, no conclusions could be drawn on the stabilization effect of π-π interactions on the whole complex.

In the present theoretical study, we investigate and rationalize the structural features of this anion-π-π complex, and quantitatively address the issue of cooperativity of anion-π and π-π interactions using a combination of dispersion-corrected density functional theory (DFT) and quantum Monte Carlo (QMC) calculations. The calculated structure is remarkably close to the one observed experimentally even though the anion-π-π complex was not additionally coordinated as in the crystal structure. Therefore, the unusual stacking is an intrinsic feature which stabilizes the anion-π binding, indicating that the principle of anion-π-π cooperativity is regulating the self-assembly in this coordination compound. Energetically, the cooperative effect of anion-π and π-π interactions in the triazine-triazine-nitrate complex is not negligible but amounts to roughly 6% of the total binding energy.

We want to emphasize that the theoretical investigation of anion-π-π interactions is particularly demanding. In anion-π systems, correlation significantly contributes to the interaction energy [93, 94] and must therefore be accurately treated. In the presence of aromatic stacking, the need to also address π-π interactions further complicates matters. Finally, even though most previous studies of anion-π systems within the MP2 approach were limited to highly symmetrical configurations, it is important to be able to explore lower-symmetry complexes for a realistic representation of the anion-π-π systems observed experimentally. Therefore, we choose here to employ the efficient DFT approach in combination with the recently proposed dispersion-corrected atom-centered pseudopotentials (DCACPs) [117, 118], which we validate against accurate highly-correlated quantum Monte Carlo

calculations. The DFT-DCACP method is found to reliably predict equilibrium structures as well as the relative stability of different complexes, and is therefore a very promising tool for the investigation of even larger anion-π systems.

5.2 Computational approaches

We briefly review below the two theoretical methods employed in this work, that is, the recently developed semi-empirical DFT scheme augmented with dispersion-corrected atom-centered potentials (DCACPs) and the quantum Monte Carlo (QMC) approach. We also give all relevant computational details.

5.2.1 Semi-empirical dispersion corrected DFT

A simple semi-empirical approach has been recently proposed to correct the deficiency of approximate density functionals in describing London dispersion forces [117, 118]. The non-local electron-nucleus pseudopotentials used in the Kohn-Sham DFT scheme are augmented with DCACPs whose parameters are fitted against references data obtained in ab-initio highly-correlated approaches. By construction, these potentials do not affect valence electronic properties but appear to significantly improve the description within DFT of weakly bound systems at no additional computational cost [119, 120].

While the original DCACPs were calibrated against MP2 reference properties, we use here the latest library of potentials constructed from more accurate coupled-cluster singles and doubles with a perturbative treatment of the triples [CCSD(T)] and configuration interaction data [121]. We employ the generalized gradient approximation (GGA) functional of Becke, Lee, Yang and Parr (BLYP) [16] and the corresponding DCACPs non-local potentials given in the Troullier-Martins [122] form. We use the DCACPs for all the atomic species except for F where we use the original Troullier-Martins BLYP pseudopotential as the corresponding DCACP is not yet available. We expect that the absence of dispersion corrections for F is not important in the complexes studied in Section 5.3.3 as the F-F interaction is dominated by electrostatic repulsion. All the DFT calculations are performed using the plane-wave basis set program CPMD 3.11.1 [46] with a plane-wave cutoff of 80 Ry. We employ the isolated system module in CPMD which allows studying an isolated molecule or complex within periodic boundary conditions. The Poisson equations are solved with the Hockney method [123]. We use a box size of $15 \times 15 \times 15$ $\mathring{A}^3$ which is sufficiently large for all the complexes

considered in this work. All the geometry optimizations are performed without imposing any symmetry constraints and with a threshold for the residual force of 0.0005 a.u. The binding energies of complexes are computed by subtracting the energies of the optimized fragments from the total energy of the complex. We note that all computed binding energies do not include zero point energy corrections.

5.2.2 Quantum Monte Carlo methods

QMC methods [124,125] offer an efficient alternative to conventional highly-correlated ab-initio methods as they can be applied to sufficiently large systems and still provide an accurate description of both dynamical and static electronic correlation. The key ingredient which determines the quality of a QMC calculation is the many-body trial wave function which, in the present work, is chosen of the Jastrow-Slater type with the particular form,

$$\Psi = D^{\uparrow} D^{\downarrow} \prod_{A,i,j} \mathcal{J}(r_{ij}, r_{iA}, r_{jA}) \,, \tag{5.1}$$

where $D^{\uparrow}$ and $D^{\downarrow}$ are Slater determinants of single-particle orbitals for the up- and down-spin electrons, respectively, and the orbitals are represented using atomic Gaussian basis. The Jastrow correlation factor $\mathcal{J}$ depends on the distance r_{ij} between electrons i and j, and on the distance r_{iA} and r_{jA} of electrons i and j from nucleus A. The Jastrow factor is here expressed as the exponential of the sum of three fifth-order polynomials of electron-nuclear, of electron-electron, and of pure three-body mixed electron-electron and electron-nucleus distances, respectively [126]. Different Jastrow factors are used to describe the correlation with different atom types.

In variational Monte Carlo (VMC), the square of the wave function is sampled using the Metropolis algorithm and the expectation value of the Hamiltonian on the wave function is computed by statistically averaging over a large number of electronic configurations sampled from Ψ^2. The wave function is then used in diffusion Monte Carlo (DMC), which produces the best energy within the fixed-node approximation, i.e. the lowest-energy state with the same zeros (nodes) as the trial wave function Ψ. All QMC results presented below are from DMC calculations.

All QMC calculations are performed with the program package CHAMP [127]. We employ scalar-relativistic energy-consistent Hartree-Fock pseudopotentials [67] for all the elements, and the hydrogen potential is softened by removing the Coulomb divergence. To represent the orbitals in the determinantal component, we employ the Gaussian basis sets [67] constructed

for these pseudopotentials and augment them with diffuse functions. All calculations are performed with the cc-pVDZ basis augmented with two additional diffuse s and p functions with exponents 0.04690 and 0.04041 for carbon, 0.06124 and 0.05611 for nitrogen, 0.07896 and 0.06856 for oxygen, and 0.06080 and 0.04660 for chlorine [68]. Only in the computation of the binding energy of the far-triazine-NO_3^- fragment (see Table 5.3), the basis is further augmented with two diffuse s and p functions with exponents 0.0138 and 0.0108 for carbon, 0.0167 and 0.0144 for nitrogen, and 0.0206 and 0.0171 for oxygen [68]. The use of these additional diffuse functions allows a stable QMC simulation in this compound where the triazine and NO_3^- molecules are at very large distances (about 7 $\AA$). Further augmentation of the basis with two diffuse d functions for all heavier atoms does not change the binding energy of the compound.

The parameters in the Jastrow factor are always optimized within VMC by energy minimization [66] and, when stated, the coefficients of the orbital expansions over the atomic Gaussian basis are simultaneously optimized with the Jastrow component. Otherwise, orbitals from a B3LYP density functional theory [16, 19] calculation are employed, which are obtained using the same pseudopotentials and basis set with the program GAMESS(US) [57]. An imaginary time step of 0.075 a.u. is used in the DMC calculations.

As side test, we compute the DMC binding energy and equilibrium distance of the prototypical triazine-chloride complex with the ion along the C_3 axes of the ring, which has been the subject of several MP2 studies [88, 92, 93, 95, 96, 99]. We perform a correlated sampling run [128] using as reference the MP2/aug-cc-pVDZ geometry with a chloride-centroid distance of 3.13 $\AA$ [93], and the corresponding fully optimized QMC wave function. We find a DMC equilibrium distance of 3.24 $\AA$, which is in the range of the MP2 values obtained with different basis sets [93]. The DMC binding energy of 6.0(3) kcal/mol is slightly smaller than the MP2/aug-cc-pVDZ value of 6.93 kcal/mol [93]. We note that using optimized or B3LYP orbitals in the determinantal component of the wave functions yields statistically equivalent results even though the B3LYP approach underestimates the binding energy by about 2 kcal/mol [95].

5.3 Results

To investigate cooperative effects of anion-π and π-π interactions in the unusual triazine-triazine-nitrate complex observed experimentally, we first need to characterize how the anion-π and π-π fragments are separately stabilized. Studying these smaller components also allows us to access the performance

of QMC and in particular of the semi-empirical DCACP approach, by comparing to MP2 or CCSD(T) calculations when available.

5.3.1 Triazine and NO_3^-

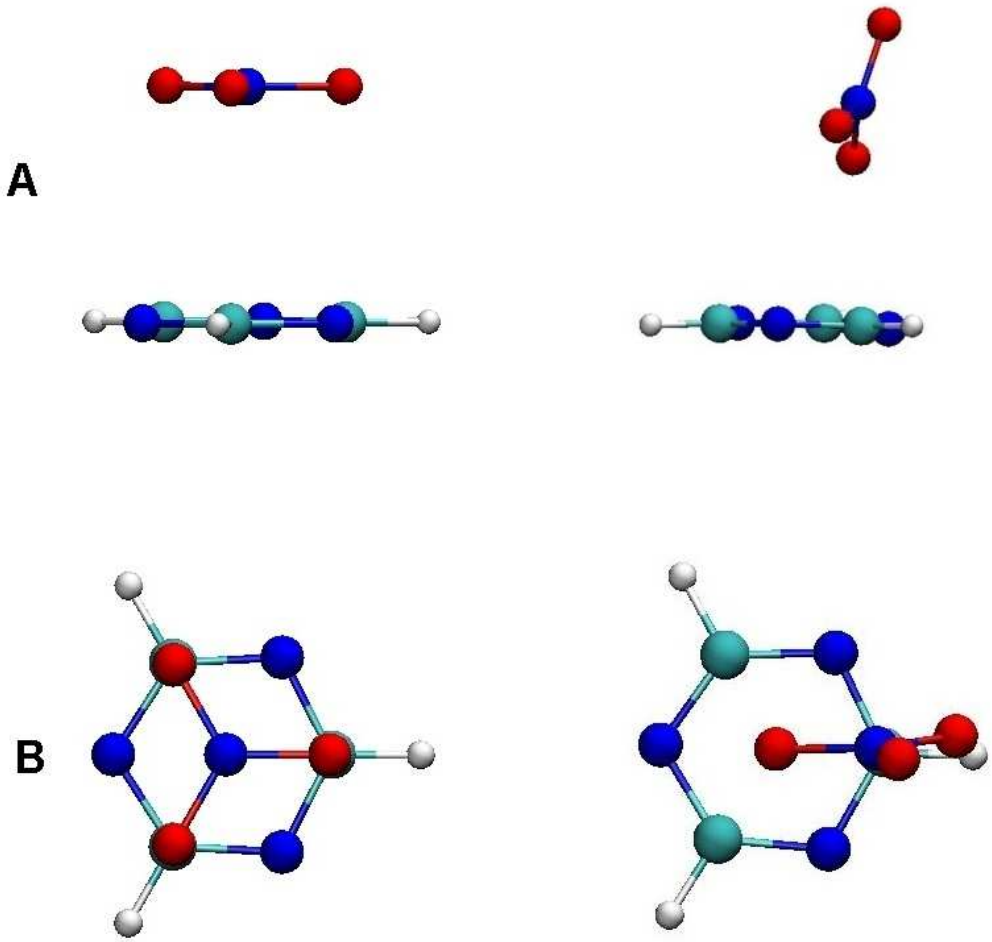

Figure 5.1: Side (A) and top (B) view of the triazine-nitrate complex in the parallel geometry (left) and in the T-like form (right).

The triazine-NO_3^- complex represents one of the first examples of anion-π interactions observed experimentally [109, 110], and has already been the subject of several computational studies both at the DFT and MP2 level [93, 109,110]. Thus, it is an ideal system to assess the accuracy and transferability

of the DCACPs in describing this novel weak interactions as all the atomic elements in the complex are available in the current library of dispersion corrected pseudopotentials [121].

Table 5.1: DFT/BLYP-DCACP and DMC binding energies in kcal/mol of the triazine-NO_3^- complex. The parallel and T-like geometries corresponding to the MP2 and DFT/BLYP-DCACP equilibrium distances are shown. The MP2 binding energies are also given when available. R_0 and $R_0^{'}$ are the equilibrium distances in $\mathring{A}$ between the ring centroid and the nitrogen atom of NO_3^-, and between the centroid and the closest oxygen atom of NO_3^-, respectively. The statistical error on the last figure of the DMC binding energy is given in parenthesis.

Geometry		R_0	$R_0^{'}$	DFT	DMC	MP2
DFT	parallel	3.18	–	-6.7	-5.1(3)	–
MP2	parallel	2.90[a]	–	–	-4.9(3)[c]	-6.8[a]
DFT	T-like	3.69	3.07	-8.8	-6.7(3)	–
MP2	T-like	–	2.75[b]	–	–	-8.4[b]

[a] Ref. [93] MP2/aug-cc-pVDZ; [b] Ref. [110] MP2/6-311++G(3df,p).
[c] Wave function fully optimized with VMC [129].

In Table 5.1, we show the binding energy and the equilibrium distance between the ring centroid and the nitrogen of NO_3^-, calculated with different approaches. We focus first on the symmetrical geometry where the plane of the nitrate is parallel to the triazine plane with the nitrogen of the anion located on top of the ring centroid and the oxygens facing the carbon atoms (Fig. 5.1). We find that DFT/BLYP-DCACP gives a binding energy which is very close to the one obtained within the MP2 approach [93] but with a slightly larger (about 10%) equilibrium distance. DMC gives a binding energy $\approx$ 1.7 kcal/mol smaller than the DFT and MP2 values. A DMC equilibrium distance of about 2.9 $\mathring{A}$ is estimated in a correlated sampling run using as reference geometry the MP2 geometry and the corresponding fully optimized QMC wave function [128]. In Fig. 5.2, we show the DFT binding energies obtained with the standard BLYP and the BLYP-DCACP pseudopotentials for the parallel geometry at different distances between the planes. The structure of the fragments is kept fixed at the optimal geometry of the complex while changing the triazine-nitrate distance. The computed BLYP binding energy of 2.4 kcal/mol is in line with the value previously obtained with a Slater-type orbital ET-pVQZ basis [109]. It clearly appears that the dispersion corrections have a very large effect on the description of the anion-π interactions. With the inclusion of the DCACP pseudopoten-

tials, the binding energy increases from 2.4 kcal/mol to 6.7 kcal/mol and the equilibrium distance decreases by about 6%.

The triazine-nitrate complexes observed experimentally [109, 110] show however a very different structure from the more intuitive face-to-face configuration. This observation prompted further calculations both at the DFT and MP2 level [109, 110], which yield an energetically lower T-like arrangement of the triazine-nitrate complex. As shown in Table 5.1, the structure and the binding energy of the T-like complex computed within DFT/BLYP-DCACP compare well with the MP2 values [110]. In particular, the dispersion corrected potentials predict this configuration to be more stable than the parallel arrangement by about 2 kcal/mol. Similarly to the MP2 results, we find that the two oxygens of the nitrate point towards the triazine plane, one facing the ring centroid and the other facing one hydrogen atom at a distance $d_{O...H} = 2.52\mathring{A}$ (see also Fig. 5.1). The interaction with the hydrogen appears to further stabilize this geometrical arrangement. The shortest oxygen-centroid distance $R_0^{'}$ predicted by DFT/BLYP-DCACP is again about 10% larger than the MP2 value. The DMC calculations performed using the DFT optimized complexes give the same trend for the binding energy with the T-like geometry being the most stable. DMC results however indicate that both DFT and MP2 tend to overestimate the binding energy by ≈2 kcal/mol.

5.3.2 The triazine dimer

The DCACPs have been developed with the main aim of improving the generally poor description of π-π interactions within DFT. The triazine dimer makes no exception in this respect, with the BLYP functional giving an unbound state. At the contrary, ab-initio correlated methods such as MP2 and CCSD(T) find a bound state with the optimal conformation being a stacked complex with a 60° relative orientation and a vertical distance of 3.4 $\mathring{A}$ [130, 131]. As shown in Table 5.2, the MP2 method tends to overestimates the binding energy in comparison with the more accurate CCSD(T) approach, a feature which appears to be general in the MP2 description of van der Waals complexes [130–132].

In Table 5.2, we show the results for the triazine dimer obtained with the DFT/BLYP-DCACP approach. We consider the face-to-face conformation with relative orientations, 0°, 30°, and 60°, between the two triazine molecules (Fig. 5.3). All structures are bound with the binding energy increasing as we move from 0° to 60°. The DFT-DCACP global minimum at 60° is consistent with the MP2 finding, but has a slightly smaller binding energy and a centroid-to-centroid distance 6% larger. DMC calculations per-

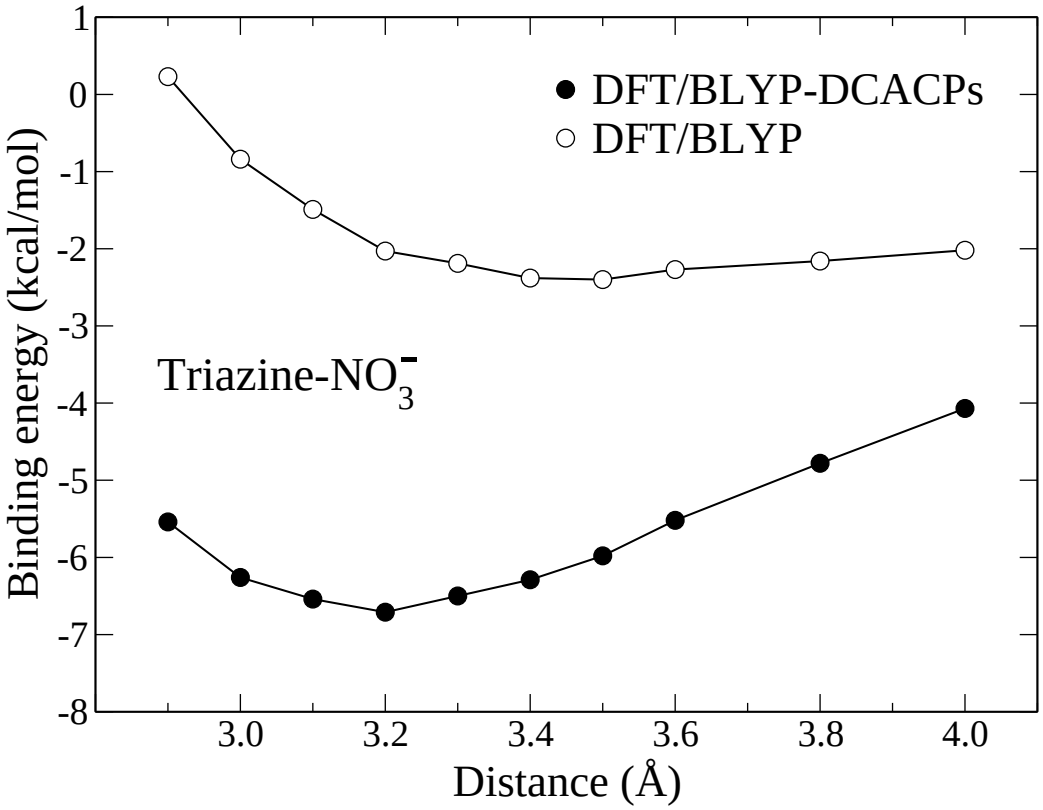

Figure 5.2: Binding energy of the triazine-NO_3^- system in the parallel geometry as a function of the distance between the centroid of the triazine ring and the nitrogen atom of NO_3^-. The curves are computed within DFT/BLYP with and without the DCACPs.

formed on the DFT optimized geometries display the same trend as DFT but give a systematically smaller binding energy by about 1 kcal/mol. We note that, at the MP2 geometry, the DMC binding energy is equal to the CCSD(T) prediction. Once more, these results indicate that the BLYP-DCACPs are able to predict the proper trend and global minimum of the weakly-bound triazine dimer although they slightly overestimate the binding energy similarly to the MP2 approach.

In the crystalline structure [109] containing the triazine-triazine-nitrate complex, an unusual triazine-triazine moiety is observed: The planes of the triazine molecules are slightly off-centered with a centroid-to-centroid distance of 3.45 $\AA$, and have a relative orientation of about 30° with a dihedral angle of about 15°. If we optimize the triazine dimer within DFT-BLYP-DCACP starting from this experimental conformation, we find a local mini-

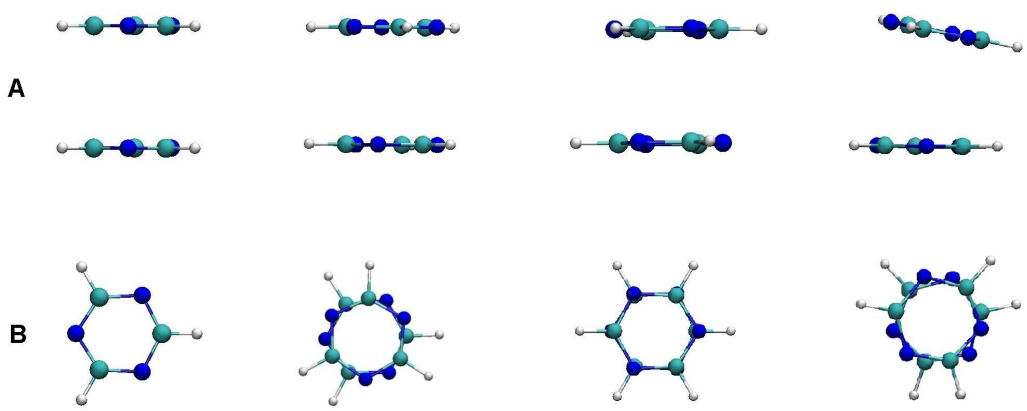

Figure 5.3: Side (A) and top (B) view of four different conformations of the triazine-triazine dimer. From left to right, the two molecules are rotated with respect to each other by 0°, 30°, and 60°. The bended geometry (right) is rotated by roughly 30°.

mum very similar to the experiment, that is, a distance between the centroids of 3.79 $\mathring{A}$ and a dihedral angle between the two rings of 11° (see Fig. 5.3). As shown in Table 5.2, the predicted binding energy for this local minimum is only 0.6 kcal/mol lower than the global minimum. However, the bending slightly stabilizes the dimer with respect to the parallel conformation at the same angle of 30°. For the bended conformation, DMC gives the same energy within statistical error as for the DFT global minimum geometry.

5.3.3 Cooperativity of anion-π-π interactions

We now focus on the experimental triazine-triazine-nitrate complex observed in Ref. [109] to understand the unusual structural features and quantify the stabilization induced on the anion-π system by π-π stacking. The relevant anion-π-π unit of the crystalline structure represents the starting point for our calculations and is shown in Fig. 5.4 together with the DFT/BLYP-DCACP optimal geometry. The similarity between the experimental and theoretical complexes is remarkable: The two triazine are staggered by about 30° and bended, and slip with respect to each other with the nitrate bound in a T-like configuration. Moreover, the nitrogen of the anion is displaced with respect to the normal through the centroid of the ring, consistently with the experiment. More specifically, we find a centroid-to-centroid distance of 3.74 $\mathring{A}$ and a distance from the centroid to the N of NO_3^- of 3.50 $\mathring{A}$ as compared to the experimental values of 3.45 $\mathring{A}$ and 3.71 $\mathring{A}$, respectively. The two aro-

Table 5.2: DFT/BLYP-DCACP and DMC binding energies in kcal/mol of different conformations of the triazine-triazine dimer. The equilibrium distance R_0 in $\AA$ is between the centroids of the two rings and the angle indicates their relative rotation.

Geometry		R_0	DFT	DMC	MP2	CCSD(T)
DFT	0°	3.80	-1.7	-0.6(3)	–	–
DFT	30°	3.70	-2.5	-1.6(3)	–	–
DFT	60°	3.60	-3.5	-2.2(3)	–	–
MP2[a]	60°	3.40	–	-2.8(3)	-3.8[a]	-2.8[a]
DFT	30° bended	3.79	-2.9	-2.1(3)	–	–
DFT	60° bended	3.61	-3.4	–	–	–

[a] Ref. [130, 131] with a diffuse cc-pVDZ' basis set.

matic rings form a dihedral angle of 22°, close to the experimental value of 18°. Finally, two oxygen atoms of the nitrate are interacting with the triazine ring, similarly to the observed structure. A more detailed comparison with the experimental structure is not appropriate as the experimental complex is embedded in the crystalline environment which may induce further distortions. We note that, with respect to the geometries of the subunits separately optimized, the presence of the nitrate does not significantly change the centroid-centroid distance but yields a larger tilt angle between the rings. As for triazine-nitrate system, the main geometrical effect is that the nitrate is more parallel and closer to the ring with a centroid-oxygen distance of 2.98 $\AA$.

The total computed binding energy of this anion-π-π complex is -12.4 kcal/mol. To establish the stabilization effect induced by π-π stacking on this supramolecular complex, we separately compute the π-π and anion-π contributions to the binding energy. We extract from the optimized structure the three fragments corresponding to the triazine dimer and the two possible triazine-nitrate units, and compute their binding energies which are listed in Table 5.3. By subtracting these contributions from the total binding energy of the anion-π-π system, we derive a cooperative contribution to the binding energy of 0.8 kcal/mol, which corresponds to a stabilization enhancement of about 6%. The cooperative energy predicted by DMC using the DFT geometries is compatible with the DFT value within two standard deviations, while the DMC binding energies of the individual fragments are always lower as already pointed out in the previous sections. In particular, the compound given by the nitrate and the distant triazine ring is unbound within DMC due to geometrical deformations of the molecules within the triazine-triazine-nitrate complex.

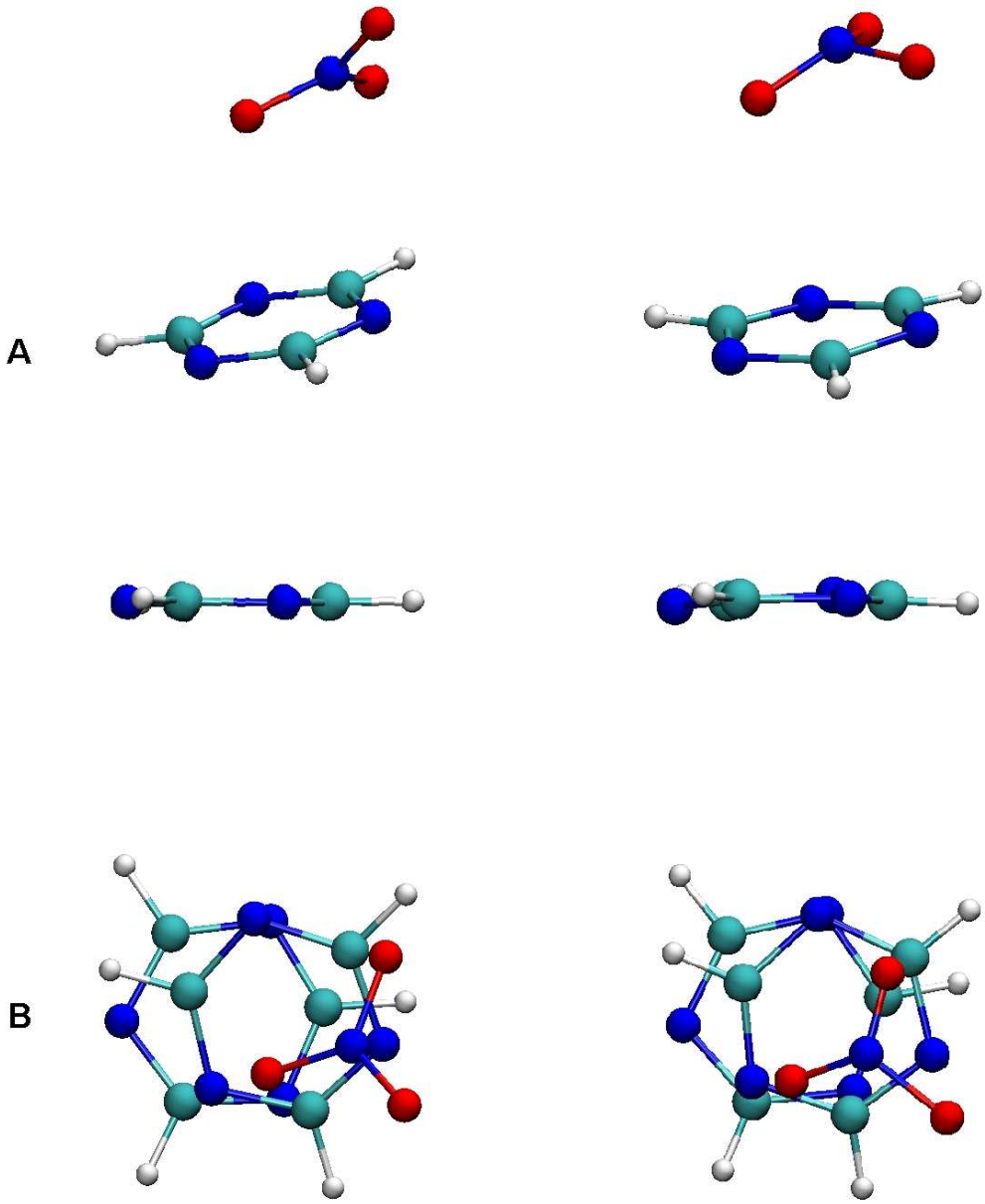

Figure 5.4: Side (A) and top (B) view of the triazine-triazine-NO_3^- complex as obtained within DFT/BLYP-DCACP (left) and experimentally [109] (right). In the top view, the bottom ring is parallel to the page.

We note that the triazine rings are not in the optimal staggered configuration of 60° (see Table 5.2) but have a relative orientation of about 30°. This is certainly due to the geometrical constraints within the crystalline frame-

work as the triazine rings are strongly coordinated via a complex network to the metal centers of the compound. To generate a triazine-triazine-nitrate complex arrangement consistent with the optimal geometries for the separate fragments, we thus start from a structure in which the triazine rings are parallel in the optimal 60° orientation, with the nitrate in the T-like form and two oxygens interacting with the ring (Fig. 5.6). The geometry optimization shows that the relative orientation remains at 60°, and the centroid-centroid distance is unchanged with respect to the isolated dimer. However, the presence of the anion induces a tilt of about 14° of the triazine close to it, which is likely induced by the interaction between one oxygen of the nitrate and one hydrogen of the ring. As expected, the binding energy of this complex is slightly larger than in the structure derived from the experiment (see Table 5.3) and the cooperative effect is of comparable magnitude.

Table 5.3: DFT/BLYP-DCACP and DMC binding energies in kcal/mol of five π-π complexes interacting with NO_3^-, i.e. two triazine (TAZ) rings at 30°, two TAZ rings at 60°, two trifluorotriazine (TFT) at 60°, and the two mixed systems with one TAZ and one TFT ring at 60°. The mixed systems denoted TAZ-TFT and the TFT-TAZ have the TFT and TAZ ring closer to the NO_3^-, respectively. The geometry of the ternary compound (ring-ring-NO_3^-) is optimized within DFT/BLYP-DCACP and the geometries of the ring-ring, close-ring-NO_3^-, and far-ring-NO_3^- fragments are here kept fixed when computing their binding energies. The binding energies are obtained as the difference between the total energy and the energies of the isolated ring and NO_3^- optimized separately. For example, the binding energy of the ternary compound is given by E_b(ring-ring-NO_3^-)=E(ring-ring-NO_3^-)−2×E(ring)−E(NO_3^-). The cooperative energy is defined as E_b(ring-ring-NO_3^-)−E_b(close-ring-NO_3^-)−E_b(far-ring-NO_3^-)−E_b(ring-ring).

Compound	TAZ 30°		TAZ 60°	TFT	TAZ-TFT	TFT-TAZ
	DFT	DMC	DFT	DFT	DFT	DFT
Ring-ring-NO_3^-	-12.4	-8.5(2)	-13.5	-21.5	-20.3	-17.3
Ring-ring	-2.8	-1.7(2)	-3.0	-0.2	-1.6	-2.6
Close-ring-NO_3^-	-8.3	-6.0(2)	-8.8	-16.5	-16.5	-8.3
Far-ring-NO_3^-	-0.5	0.7(2)	-0.7	-4.3	-0.7	-4.6
Cooperativity	-0.8	-1.6(4)	-0.9	-0.6	-1.5	-1.8

Finally, we check the effect of enhancing the π-anion interaction by substituting the hydrogens in the triazine rings at 60° with the strongly electronegative fluorine [88, 92, 96]. With this procedure, we generate three complexes, i.e. (1) two trifluorotriazine rings with a nitrate, (2) one triazine and one tri-

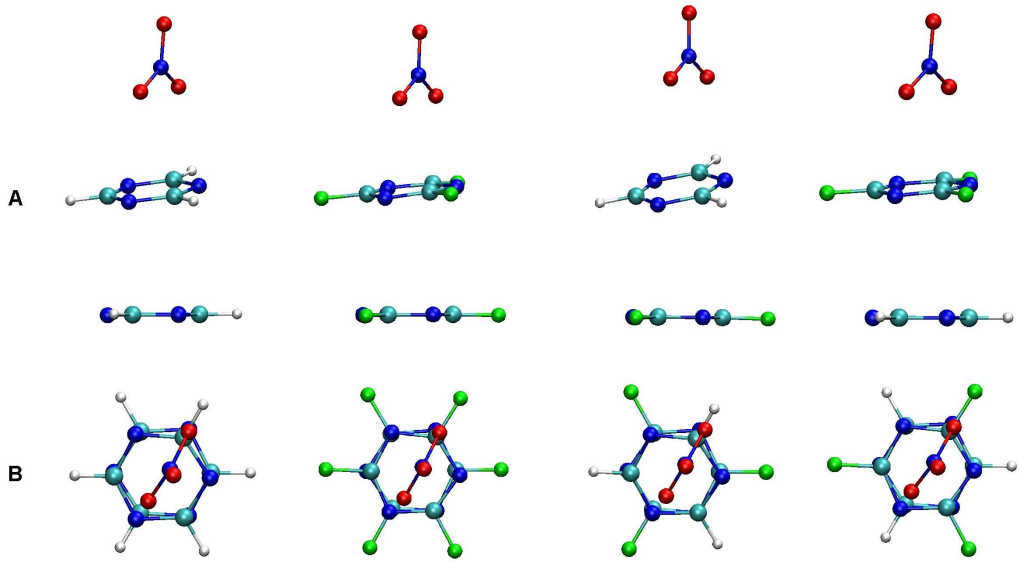

Figure 5.5: Side (A) and top (B) view of four different π-π complexes interacting with NO_3^-. From left to right, we show two triazine (TAZ) rings at 60°, two trifluorotriazine (TFT) at 60°, and the two mixed systems, TFT-TAZ and the TAZ-TFT. The TFT-TAZ and TAZ-TFT complexes have the TFT and TAZ ring closer to the NO_3^-, respectively.

fluorotriazine coordinated with the nitrate, and (3) one trifluorotriazine and one triazine coordinated with the nitrate. The three models are also shown in Fig. 5.5.

As seen in Table 5.3, the total binding is increased by the fluorine substitution due to the stronger attractive interaction between the nitrate and the trifluorotriazine ring(s). However, this increase does not always correlate with a cooperative enhancement of the π-anion interaction by π-π stacking. In particular, in the complex with two trifluorotriazine rings, the cooperative effect is smaller than in the original system with two triazine rings as the π-π system is now very weakly bound: The binding energy of the ring-ring fragment is only 0.2 kcal/mol, and significantly reduced from the 2.0 kcal/mol of the trifluorotriazine dimer optimized in the absence of the nitrate. This small binding can be explained with the strong deformation of the trifluorotriazine ring close to the nitrate, with one of the fluorine's bending out of the ring plane away from one of the nitrate oxygens.

In the mixed triazine-trifluorotriazine compounds, we observe instead a more favorable balance of π-π and anion-π interactions, leading to an increased cooperativity. In both complexes, the ring-ring fragment is still significantly bound even though its binding energy is smaller than the optimal

value of 3.2 kcal/mol for the isolated dimer. The ring-ring binding energy of the complex with the triazine coordinated to the nitrate is larger by about 1 kcal/mol than the energy obtained in the case of the trifluorotriazine coordinated to the nitrate. The reduced binding of the latter is due to a strong deformation of the trifluorotriazine ring vicinal to the nitrate, similarly to what observed in the compound with two trifluorotriazine rings. Correspondingly, the largest cooperative effect is observed in the mixed complex with the triazine coordinated to the nitrate where the non-additive contribution amounts to roughly 10% of the total binding energy. Finally, we note that the cooperative effect is always present in all the studied complexes while this does not appear to be a general feature [104] of anion-π-π complexes, and points to the very versatile nature of the triazine moiety in supramolecular chemistry [133].

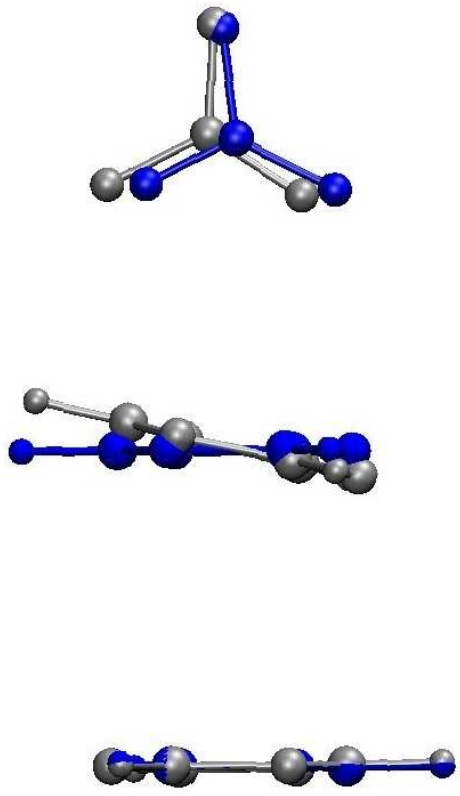

Figure 5.6: Triazine-triazine-NO_3^- complex with relative orientation of the rings of 60° optimized within DFT/BLYP-DCACP (grey). In blue, we show the starting symmetrical geometry.

5.4 Conclusions

Using state-of-the-art dispersion corrected DFT and QMC calculations, we have investigated the geometrical and energetic effects induced by π-π stacking on the anion-π system of the unusual triazine-triazine-nitrate complex recently observed experimentally. We have reproduced and rationalized the highly asymmetrical features of the structure, which are not imposed by the coordination of the anion-π-π subunit within the particular synthesized compound. We show that the two triazines are staggered and bended, and slip with respect to each other with the nitrate bound off-center in a T-like configuration. The stabilization induced by π-π stacking amounts energetically to about 6% of the total binding energy. An increased cooperative effect of 10% is obtained if the hydrogens in one of the triazine rings are substituted with the strongly electronegative fluorine atoms and the dimer is further staggered in a 60° orientation. We want to emphasize that the theoretical investigation of a realistic anion-π-π system as the one treated in this paper is particularly demanding as correlation plays an important role, while the system is not small and must be treated without symmetry constraints. We find that the use of the recently proposed dispersion corrected DFT approach represents a good compromise between accuracy and the ability to study complex and realistic systems involving weak interactions.

Bibliography

[1] R. Y. Tsien, Annu. Rev. Biochem. **67**, 509 (1998).

[2] M. Zimmer, Chem. Rev. **102**, 759 (2002).

[3] V. Tozzini, V. Pellegrini, and F. Beltram, in *Handbook of organic photochemistry and photobiology*, Edited by W. M. Horsphool and F. Lenci, Chapter 139 (CRC press, Washington DC, 2004).

[4] O. Shimomura, F. H. Johnson, and Y. Saiga, J. Cell. Comp. Physiol. **59**, 223 (1962).

[5] Mills, C.E. 1999-present. Bioluminescence of Aequorea, a hydromedusa. Electronic internet document available at http://faculty.washington.edu/cemills/Aequorea.html. Published by the author, web page established June 1999, last updated 15 February 2007.

[6] A. K. Hadjantonakis, M. Gertsenstein, M. Ikawa, M. Okabe, A. Nagy, Mech. Dev. **76**, 79 (1998).

[7] T. M. H. Creemers, A. J. Lock, V. Subramaniam, T. M. Jovin, and S. Volker, Nat. Struct. Biol. **6**, 557 (1999).

[8] W. Weber, V. Helms, J. A. McCammon, and P. W. Langhoff, Proc. Natl. Acad. Sci. **96**, 6177 (1999).

[9] A. A. Voityuk, M-E. Michel-Beyerle, N. Rösch, Chem. Phys. Lett. **272**, 162 (1997); *ibid.* **296**, 269 (1998); Chem. Phys. **231**, 13 (1998).

[10] J. El Yazal, F. G. Prendergast, D. E. Shaw, and Y-P. Pang, J. Am. Chem. Soc. **122**, 11411 (2000).

[11] V. Helms, C. Winstead, and P. W. Langhoff, J. Mol. Struct. **506**, 179 (2000).

[12] M. A. L.Marques, X. Lopez, D. Varsano, A. Castro, and A. Rubio, Phys. Rev. Lett. **90**, 258101 (2003).

[13] A. Sinicropi, T. Andruniow, N. Ferre, R. Basosi, and M. Olivucci, J. Am. Chem. Soc. **127**, 11534 (2005).

[14] P. Hohenberg and W. Kohn, Phys. Rev. **136**, B864 (1964).

[15] W. Kohn and L.J. Sham, Phys. Rev. **140**, A1133 (1965).

[16] A. D. Becke, Phys. Rev. A **38**, 3098 (1988); C. T. Lee, W. T. Yang, and R. G. Parr, Phys. Rev. B **37**, 785 (1988).

[17] J. P. Perdew, K. Burke, and M. Ernzerhof, Phys. Rev. Lett. **77**, 3865 (1996).

[18] J. P. Perdew, M. Ernzerhof, and K. Burke, J. Chem. Phys. **105**, 9982 (1996).

[19] A. D. Becke, J. Chem. Phys. **98**, 5648 (1993).

[20] R. G. Parr and W. Yang, *Density-Functional Theory of Atoms and Molecules* (New York, Oxford University Press, 1989).

[21] R. M. Dreizler and E. K. U. Gross, *Density Functional Theory,An Approach to the Quantum Many-Body Problem* (New York, Springer-Verlag, 1990).

[22] W. Koch and M. C. Holthausen, *A Chemist's Guide to Density Functional Theory* (New York, Wiley-VCH, 2000).

[23] E. Runge and E. Gross, Phys. Rev. Lett. **52**, 997 (1984).

[24] R. van Leeuwen, Phys. Rev. Lett. **82**, 3863 (1998).

[25] M. E. Casida, Recent Advances in Density Functionals methods (World Scientific:Singapore).

[26] M. E. Casida, J. Chem. Phys. **122**, 054111 (2005).

[27] M. E. Casida, K. C. Casida, and D. R. Salahub, Int. J. Quantum Chem. **70**, 933 (1998).

[28] D. J. Tozer, R. D. Amos, N. C. Handy, B. O. Roos, and L. S. Andres, J. Mol. Physics **97**, Issue 7, 859 (1999).

[29] M. E. Casida, F. Gutierrez, J. Guan, F. X. Gadea, and D. Salahub, J. Chem. Phys. **113**, 7062 (2000).

[30] A. Dreuw, J. L. Weisman, and M. Head-Gordon, J. Chem. Phys. **119**, 2943 (2003).

[31] N. T. Maitra, F. Zhang, R. J. Cave, and K. Burke, J. Chem. Phys. **120**, 5932 (2004).

[32] N. Metropolis, A. Rosenbluth, M. N. Rosenbluth, A. H. Teller, and E. Teller, J. Chem. Phys. **21**, 1087 (1953).

[33] C. J. Umrigar, Phys. Rev. Lett. **71**, 408 (1993).

[34] M. H. Kalos, Phys. Rev. **128**, 1791 (1962); J. Comp. Phys.**2**, 257 (1967); M. H. Kalos, D. Levesque, and L. Verlet, Phys. Rev. A**9**, 2178 (1974).

[35] D. M. Ceperley and B. J. Alder, Phys. Rev. Lett.**45**, 566 (1980).

[36] C. J. Umrigar, M. P. Nightingale and K. J. Runge, J. Chem. Phys.**99**, 2865 (1993).

[37] C. J. Umrigar, J. Toulouse, C. Filippi, S. Sorella, and R. G. Hennig, Phys. Rev. Lett. **98**, 110201 (2007).

[38] M. P. Nightingale and V. Melik-Alaverdian, Phys. Rev. Lett. **87**, 043041 (2001).

[39] S. Fahy, X. W. Wang and S. G. Louie, Phys. Rev. B **42**, 3503 (1990).

[40] B. L. Hammond, P. J. Reynolds, and W. A. Lester Jr., J. Chem. Phys. **87**, 1130 (1987); M. M. Hurley and P. A. Christiansen, J. Chem. Phys. **86**, 1069 (1986); P. A. Christiansen, *ibid.* **88**, 4867 (1988); L. Mitas, E. L. Shirley and D. M. Ceperley, J. Chem. Phys. **95**, 3467 (1991).

[41] M. Casula, C. Filippi and S. Sorella, Phys. Rev. Lett. **95**, 100201 (2005).

[42] A. Laio, J. VandeVondele, and U. Rothlisberger, J. Chem Phys. **116**, 6941 (2002).

[43] D. A. Case, T. E. Cheatham, T. Darden, H. Gohlke, R. Luo, K. M. Merz, A. Onufriev, C. Simmerling, B. Wang, and R.J. Woods, J. Comput. Chem. **26**, 1668(2005).

[44] J. W. Ponder and D. A. Case, Advances in Protein Chemistry **66**, 27 (2003).

[45] T. E. Cheatham and M. A. Young , Biopolymers **56**, 232 (2001).

[46] J. Hutter *et al.*, `CPMD`, Version 3.11.1, Copyright IBM Corp. 1990-2006, Copyright MPI-FKF Stuttgart 1997-2004; `http://www.cpmd.org`

[47] W. F. van Gunsteren, S. R. Billeter, A. A. Eising, P. H. Hnenberger, P. Krger, A. E. Mark, W. R. P. Scott, I. G. Tironi, Biomolecular Simulation: The GROMOS96 Manual and User Guide; Vdf Hochschulverlag AG an der ETH Zurich: Zurich, 1996.

[48] A. Laio, J. VandeVondele, and U. Rothlisberger, J. Chem. Phys. **116**, 6941 (2002).

[49] M. C. Colombo, L. Guidoni, A. Laio, A. Magistrato, P. Maurer, S. Piana, U. Rohrig, K. Spiegel, and U. Rothlisberger, CHIMIA **56**, 11 (2002).

[50] N. Troullier and J. L. Martins, Phys. Rev. B **43**, 1993 (1991).

[51] G. J. Martyna and M. E. Tuckerman, J. Chem. Phys. **110**, 2810 (1999).

[52] J. P. Perdew, K. Burke, and M. Ernzerhof, Phys. Rev. Lett. **77**, 3865 (1996).

[53] Frisch, M. J., G. W. Trucks, H. B. Schlegel, G. E. Scuseria, M. A. Robb, J. R. Cheeseman, J. A. Montgomery, Jr., T. Vreven, K. N. Kudin, J. C. Burant, J. M. Millam, S. S. Iyengar, J. Tomasi, V. Barone, B. Mennucci, M. Cossi, G. Scalmani, N. Rega, G. A. Petersson,H. Nakatsuji, M. Hada, M. Ehara, K. Toyota, R. Fukuda, J. Hasegawa,M. Ishida, T. Nakajima, Y. Honda, O. Kitao, H. Nakai, M. Klene, X. Li, J. E. Knox, H. P. Hratchian, J. B. Cross, V. Bakken, C. Adamo,J. Jaramillo, R. Gomperts, R. E. Stratmann, O. Yazyev, A. J. Austin, R. Cammi, C. Pomelli, J. W. Ochterski, P. Y. Ayala, K. Morokuma, G. A. Voth, P. Salvador, J. J. Dannenberg, V. G. Zakrzewski, S. Dapprich,A. D. Daniels, M. C. Strain, O. Farkas, D. K. Malick, A. D. Rabuck,K. Raghavachari, J. B. Foresman, J. V. Ortiz, Q. Cui, A. G. Baboul,S. Clifford, J. Cioslowski, B. B. Stefanov, G. Liu, A. Liashenko,P. Piskorz, I. Komaromi, R. L. Martin, D. J. Fox, T. Keith, M. A.Al-Laham, C. Y. Peng, A. Nanayakkara, M. Challacombe, P. M. W.Gill, B. Johnson, W. Chen, M. W. Wong, C. Gonzalez, and J. A. Pople.2004. Gaussian 03, Revision C.02. Gaussian, Inc., Wallingford, CT.

[54] E.J. Baerends, J. Autschbach, A. Brces, J.A. Berger, F.M. Bickelhaupt, C. Bo, P.L. de Boeij, P.M. Boerrigter, L. Cavallo, D.P. Chong, L. Deng, R.M. Dickson, D.E. Ellis, M. van Faassen, L. Fan, T.H. Fischer, C.

Fonseca Guerra, S.J.A. van Gisbergen, J.A. Groeneveld, O.V. Gritsenko, M. Grning, F.E. Harris, P. van den Hoek, C.R. Jacob, H. Jacobsen, L. Jensen, E.S. Kadantsev, G. van Kessel, R. Klooster, F. Kootstra, E. van Lenthe, D.A. McCormack, A. Michalak, J. Neugebauer, V.P. Nicu, V.P. Osinga, S. Patchkovskii, P.H.T. Philipsen, D. Post, C.C. Pye, W. Ravenek, P. Romaniello, P. Ros, P.R.T. Schipper, G. Schreckenbach, J.G. Snijders, M. Sol, M. Swart, D. Swerhone, G. te Velde, P. Vernooijs, L. Versluis, L. Visscher, O. Visser, F. Wang, T.A. Wesolowski, E.M. van Wezenbeek, G. Wiesenekker, S.K. Wolff, T.K. Woo, A.L. Yakovlev, and T. Ziegler, ADF package

[55] P. R. T. Schipper, O. V. Gritsenko, S. J. A. van Gisbergen, and E. J. Baerends, J. Chem. Phys. **112**, 1344 (1999).

[56] R. van Leeuwen and E. J. Baerends, Phys. Rev. A **49**, 2421 (1994).

[57] M.W.Schmidt, K.K.Baldridge, J.A.Boatz, S.T.Elbert, M.S.Gordon,J.H.Jensen, S.Koseki, N.Matsunaga, K.A.Nguyen, S.J.Su, T.L.Windus, M.Dupuis, J.A.Montgomery J. Comput. Chem. 14, 1347-1363 (1993).

[58] S. B. Nielsen, A. Lapierre, J. U. Andersen, U. V. Pedersen, S. Tomita, and L. H. Andersen, Phys. Rev. Lett. **87**, 228102 (2001).

[59] L. Lammich, M. A. Pertersen, M. B. Nielsen, and L. H. Andersen, Biophys. J. **92**, 201 (2007).

[60] M. E. Martin, F. Negri, and M. Olivucci, J. Am. Chem. Soc. **126** ,5452 (2004).

[61] F. Schautz, F. Buda, and C. Filippi, J. Chem. Phys. **121**, 5836 (2004).

[62] F. Schautz and C. Filippi, J. Chem. Phys. **120**, 10931 (2004).

[63] A. Scemama and C. Filippi, Phys. Rev. B **73**, 241101(R) (2006).

[64] F. Cordova, L. J. Doriol, A. Ipatov, M. E. Casida, C. Filippi, and A. Vela, J. Chem. Phys. **127**, 164111 (2007).

[65] C. Filippi (unpublished).

[66] C. J. Umrigar, J. Toulouse, C. Filippi, S. Sorella, and R. Henning, Phys. Rev. Lett. **98**, 110201 (2007).

[67] M. Burkatzki, C. Filippi, and M. Dolg, J. Chem. Phys. **126**, 234105 (2007); `http://www.tc.uni-koeln.de/data/psdb/intro.html`

[68] The exponents of the diffuse functions are taken from the aug-cc-pVDZ basis sets `http://www.emsl.pnl.gov/forms/basisform.html`

[69] R. Nifosi' and V. Tozzini, Proteins:Structure, Functions and Genetics **51**, 378 (2003).

[70] V. Tozzini, private communication.

[71] F. Yang, L. G. Moss, and G. N. Phillips, Nat. Biotechnol **14**, 1246 (1996).

[72] Abola et al., *Crystallographic Database-Information Content, Software Systems, Scientific Applications* ; Data Commission of the Internationa Union of Crystallography: Bonn 1987; pp107-132

[73] W. L. Jorgensen, J. Chandrasekhar, J. D. Madura, R. W. Impey, and M. L. Klein, J. Chem. Phys. **79**, 926 (1983).

[74] N. Reuter, H. Lin, and W. Thiel, J. Phys. Chem. B **106**, 6310 (2002).

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

[76] M. Ormo, A. B. Cubitt, K. Kallio, L. A. Gross, R. Y. Tsien, and S. J. Remington, Science **273**, 1392 (1996).

[77] H. Lousseau, A. Krummer, R. Heinecke, F. Pöllinger-Dammer, C. Kompa, G. Bieser, T. Jonsson, C. M. Silva, M. M. Yang, D. C. Youvan, and M. E. Michel-Beyerle, Chem. Phys. **213**, 1 (1996).

[78] M. Chattoraj, B. A. King, G. U. Bublitz, and S. G. Boxer, Proc. Natl. Acad. Sci. **93**, 8362 (1996).

[79] J. J. van Thor, A. J. Pierik, I. Nugteren-Roodzant, A. Xie, and K. J. Hellingwerf, Biochemistry **37**, 16915 (1998).

[80] H. Yoo, J. A. Boatz, V. Helms, J. A. McCammon, and P. W. Langhoff, J. Phys. Chem. B **105**, 2850 (2001).

[81] T. Laino, R. Nifosi, and V. Tozzini, Chem. Phys. **298**, 17 (2004).

[82] T. Laino , private communication.

[83] A. Castro, private communication.

[84] M. E. Casida, C. Jamorski, K. C. Casida, and D. R. Salahub, J. Chem. Phys. **108**, 4439 (1998).

[85] W. W. Ward, H. J. Prentice, A. F. Roth, C. W. Cody, and S. C. Reeves, Photochem. Photobiol. **35**, 803 (1982).

[86] A. F. Bell, X. He, R. M. Wachter, and P. J. Tonge, Biochemistry **39**, 4423 (2000).

[87] P. D. Beer and P. A. Gale, Angew. Chem. Int. Ed. **40**, 486 (2001).

[88] M. Mascal, A. Armstrong, and M. B. Bartberger,J. Am. Chem. Soc. **124**, 6274 (2002).

[89] I. Akkorta, I. Rozas, and J. Elguero, J. Am. Chem. Soc. **124**, 8593 (2002).

[90] D. Quiñonero, C. Garau, A. Frontera, P. Ballester, A. Costa, andP. M. Deyà, Chem. Phys. Lett. **359**, 486 (2002).

[91] C. Garau, D. Quiñonero, A. Frontera, A. Costa, P. Ballester, and P. M. Deyà, Chem. Phys. Lett. **370**, 7 (2003).

[92] C. Garau, D. Quiñonero, A. Frontera, P. Ballester, A. Costa, andP. M. Deyà, ChemPhysChem **4**, 1344 (2003).

[93] D. Kim, P. Tarakeshwar, and K. S. Kim, J. Phys. Chem. A **108**, 1250 (2004).

[94] C. Garau, A. Frontera, D. Quiñonero, P. Ballester, A. Costa, andP. M. Deyà, J. Phys. Chem. A **108**, 9423 (2004).

[95] D. Quiñonero, C. Garau, A. Frontera, P. Ballester, A. Costa, and P. M. Deyà, J. Phys. Chem. A **109**, 4632 (2005).

[96] C. Garau, D. Quiñonero, A. Frontera, P. Ballester, A. Costa, and P. M. Deyà, J. Phys. Chem. A **109**, 9341 (2005).

[97] C. Garau, A. Frontera, P. Ballester, D. Quiñonero, A. Costa, and P. M. Deyà, Eur. J. Org. Chem. **XXX**, 179 (2005).

[98] A. Frontera, F. Saczewski, M. Gdaniec, E. Dziemidowicz-Borys, A. Kurland, P. M. Deyà, D. Quiñonero, and C. Garau, Chem. Eur. J **11**, 6590 (2005).

[99] M. Mascal, Angew. Chem. Int. Ed. **45**, 2890 (2006).

[100] A. Clements and M. Lewis, J. Phys. Chem. A **110**, 12705 (2006).

[101] D. Quiñonero, A. Frontera, C. Garau, P. Ballester, A. Costa, and P. M. Deyà, ChemPhysChem **7**, 2487 (2006).

[102] O. B. Berryman, V. S. Bryantsev, D. P. Stay, D. W. Johnson, and B. P. Hay, J. Am. Chem. Soc. **129**, 48 (2007).

[103] D. Quiñonero, A. Frontera, D. Escusero, P. Ballester, A. Costa, and P. M. Deyà, ChemPhysChem **8**, 1182 (2007).

[104] A. Frontera, D. Quiñonero, A. Costa, and P. M. Deyà, New J. Chem. **31**, 556 (2007).

[105] Y. S. Rosokha, S. V. Lindeman, S. V. Rosokha, and J. K. Kochi, Angew. Chem. Int. Ed. **43**, 4650 (2004).

[106] S. Demeshko, S. Dechert, and F. Meyer, J. Am. Chem. Soc. **126**, 4508 (2004).

[107] P. de Hoog, P. Gamez, I. Mutikainen, U. Turpeinen, and J. Reedijk, Angew. Chem. Int. Ed. **43**, 5815 (2004).

[108] B. L. Schottel, H. T. Chifotides, M. Shatruk, A. Chouai, L. M. Pérez, J. Bacsa, andK. R. Dunbar, J. Am. Chem. Soc. **128**, 5859 (2006).

[109] H. Casellas, C. Massera, F. Buda, P. Gamez, and J. Reedijk, New J. Chem. **30**, 1561 (2006).

[110] P.U. Maheswari, B. Modec, A. Pevec, B. Kozlevcar, C. Massera, P. Gamez, J. Reedijk, Inorg. Chem. **45**, 6637 (2006).

[111] P. Gamez, T. J. Mooibroek, S. J. Teat, and J. Reedijk, Acc. Chem. Res. **40**, 435 (2007).

[112] P. S. Lakshminarayanan, I. ravikumar, E. Suresh, and P. Ghosh, Inorg. Chem. **46**, 4769 (2007).

[113] H. Casellas, O. Roubeau, S. J. Teat, N. Mascocchi, S. Galli, A. Sironi, P. Gamez, and J. Reedijk, Inorg. Chem. **46**, 4583 (2007).

[114] F. Hettche, and R. W. Hoffman, New J. Chem. **27**, 172 (2003).

[115] O. B. Berryman, F. Hof, M. J. Hynes, and D. W. Johnson, Chem. Commun. 506 (2006).

[116] I. Akkorta, I. Rozas, and J. Elguero, J. Org. Chem. **62**, 4687 (1997).

[117] O. A. von Lilienfeld, I. Tavernelli, U. Rothlisberger, and D. Sebastiani, Phys. Rev. Lett. **93**, 153004 (2004).

[118] O. A. von Lilienfeld, I. Tavernelli, U. Rothlisberger, and D. Sebastiani, Phys. Rev. B **71**, 195119 (2005).

[119] O. A. von Lilienfeld and D. Andrienko, J. Chem. Phys. **124**, 054307(2006).

[120] A. Tkatchenko and O. A. von Lilienfeld, Phys. Rev. B **73**, 153406 (2006).

[121] I-C. Lin, M. D. Coutinho-Neto, C. Felsenheimer, O. A. von Lilienfeld, I. Tavernelli, and U. Rothlisberger, Phys. Rev. B **75**, 205131 (2007).

[122] N. Troullier and J. L. Martins, Phys. Rev. B **43**, 1993 (1991).

[123] R. W. Hockney, Methods Comput. Phys. **9**, 136 (1970).

[124] W.M.C. Foulkes, L. Mitas, R.J. Needs, and G. Rajagopal, Rev. Mod. Phys. **73**, 33 (2001).

[125] Hammond, B. L.; Lester, Jr. W. A.; Reynolds, P. J. *MonteCarlo Methods in Ab Initio Quantum Chemistry* (World Scientific, Singapore, 1994).

[126] C. Filippi and C. J. Umrigar, J. Chem. Phys. **105**, 213 (1996). The Jastrow factor is adapted to deal with pseudo-atoms and the scaling factor κ is set to 0.5 for all atoms.

[127] `CHAMP` is a quantum Monte Carlo program package written by C. J. Umrigar and C. Filippi and collaborators; `http://www.ilorentz.org/~filippi/champ.html`.

[128] C. Filippi and C. J. Umrigar, Phys. Rev. B **61**, R16291 (2000). As secondary wave functions, we employ the reference wave function recentered at the secondary geometries, both for the triazine-Cl^- and the triazine-NO_3^- complexes.

[129] The wave function is fully optimized in VMC within energy minimization. The use of B3LYP orbitals yields a statistically compatible binding energy of 4.4(3) kcal/mol.

[130] P. Hobza, Ann. Rep. Prog. Chem. Sec. C **93**, 257 (1996).

[131] J.Sponer, and P. Hobza, Chem. Phys. Lett. **263**, 267, 1997.

[132] S. Sorella, M. Casula, and D. Rocca, J. Chem. Phys. **127**, 014105 (2007).

[133] P. Gamez and J. Reedijk, Eur. J. Inorg. Chem, 29 (2006).

Samenvatting

Absorptie van licht en het omzetten daarvan in andere vormen van energie ligt aan de basis van een aantal van de meest belangrijke cellulaire processen in levende organismen. Over het algemeen is een biologisch systeem gevoelig voor licht door een eiwit dat een chromofoor bevat (dat is een molecuul, gebonden aan het eiwit, dat verantwoordelijk is voor absorptie en emissie van licht). Deze chromofoor staat centraal in een fotochemische reactie. Het verdiepen van onze kennis van de primaire excitatie processen en van de daaropvolgende energieoverdrachtmechanismen in fotobiologische systemen is zowel fundamenteel van belang als in bestaande en potentile toepassingen in de biotechnologie, waar deze kennis bijvoorbeeld kan worden toegepast bij het ontwikkelen van autofluorescerende eiwitten met nieuwe spectroscopische eigenschappen, die verkregen worden door selectieve mutaties.

Theoretische berekeningen van de optische eigenschappen van fotoactieve systemen complementeren spectroscopische data door het geven van een beschrijving op atomair niveau van de reactie van het eiwit op licht. De theoretische benadering van dergelijke problemen moet een accurate kwantummechanische beschrijving van de grondtoestand en de gexciteerde toestanden van de fotoactieve component van het eiwit bevatten. Ook moet men een voldoende groot model van het biosysteem kunnen beschrijven, aangezien de omgeving van het eiwit een rol kan spelen bij het vastleggen van de optische respons van het chromofoor. Het is verre van triviaal om aan deze voorwaarden te voldoen en ondanks significante vooruitgang van methoden om elektronische structuur te berekenen, blijft het een moeilijke opgave om zelfs van relatief kleine organische moleculen de excitatie-energien te berekenen.

In deze dissertatie gebruiken we een aantal state-of-the-art methoden om het probleem op verschillende niveaus van nauwkeurigheid te behandelen. Wij menen dat conventionele methoden toereikend zijn voor het beschrijven van de grondtoestand van deze fotoactieve biomoleculen. Voor de gexciteerde toestanden zullen we de prestaties van een andere benadering onderzoeken. In het bijzonder worden eigenschappen van de grondtoestand beschreven met

density functional theory in combinatie met ab-initio moleculaire dynamica om de gegenereerde structuur naar een evenwicht te brengen en thermische fluctuaties van de chromofoor en zijn onmiddelijke omgeving te bestuderen. De interacties tussen het eiwit en de chromofoor, die door een lange afstand worden gekarakteriseerd, worden kwantummechanisch beschreven; de interacties met de rest van het macromolecuul worden beschreven door een klassiek atomair krachtenveld. Voor de berekening van de gexciteerde toestanden gebruiken we een ander theoretisch kader, gebaseerd op veel-deeltjes kwantum Monte Carlo technieken en, voor de lange afstandsinteracties tussen eiwit en chromofoor, combineren we voor het eerst kwantum Monte Carlo met klassieke moleculaire mechanica.

Met deze hirarchische combinatie van methoden bestuderen we het complexe gedrag van Green Fluorescent Proteins (GFPs, het prototype van de klasse van autofluorescente eiwitten en n van de meest gebruikte moleculen in de celbiologie als fluorescent label) onder invloed van licht. In het bijzonder bekijken we de wisselwerking tussen de spectrale eigenschappen en de microscopische structuur van het chromofoor-eiwit complex in zijn verschillende verschijningsvormen. Naast zijn enorme belang in de biotechnologie, is GFP interessant omdat het experimenteel zeer goed is gekarakteriseerd en ook vaak het onderwerp is van zowel semi-empirisch als first principles theoretisch onderzoek. Desalniettemin is er nog geen volledige theorie. Daarom is GFP de ideale arena om de door ons voorgestelde methoden te testen en zo mogelijk te verbeteren.

In hoofdstuk 1 beschouwen we de algemene relevantie van autofluorescente eiwitten en beschrijven we het fluorescentie mechanisme van wild-type GFP. In zijn neutrale vorm absorbeert de chromofoor van GFP blauw licht, wat hem in een anionische toestand brengt na het doneren van een proton aan de rest van het eiwit. De chromofoor fluoresceert daarna groen licht en keert terug naar zijn neutrale toestand door weer een proton op te nemen. Als het chromofoor met groen licht wordt gexciteerd, komt het eiwit in een andere, aangeslagen, anionische toestand terecht.

De berekeningsmethoden die we gebruiken in dit proefschrift worden beschreven in hoofdstuk 2. We beschouwen kwantumchemische methoden om sterk gecorreleerde systemen te modelleren en density functional theory (DFT), ook in zijn tijdsafhankelijke vorm (TDDFT). We behandelen kwantum Monte Carlo (QMC) methoden in detail en beschrijven kort enkele moleculaire mechanica technieken en de hybride kwantum mechanica in moleculaire mechanica (QM/MM) aanpak.

Hoofdstuk 3 construeren we een paar modellen van neutrale en anionische chromoforen van GFP in de gas fase om de prestaties van adiabatische TDDFT en QMC methoden te onderzoeken. De resultaten zijn nogal vreemd.

TDDFT blijkt de excitatie-energie van een klein anionisch model van het chromofoor te overschatten vergeleken met fotodestructie spectroscopie experimenten, terwijl het experimentele absorptiemaximum met dezelfde techniek voor een cationisch model wel redelijk gereproduceerd kan worden. We kunnen de redenen voor het klaarblijkelijk falen van TDDFT voor dit kleine, anionische model niet vaststellen. Door gebruik te maken van kwantum Monte Carlo technieken en geavanceerde golffuncties verkrijgen we de energien van de excitaties voor het kleine, anionische model die redelijk overeenstemmen met TDDFT. Een significant verschil met TDDFT is dat QMC een grote verschuiving oplevert van de excitatie als we van het neutrale naar het anionische model van het GFP chromofoor in de gas fase gaan.

In hoofdstuk 4 berekenen we de eiwitmodellen van de neutrale vorm en de twee anionische vormen van wild-type GFP, gebruik makende van density functional theory QM/MM. De uitkomst van deze berekening van de grondtoestand is verrassend en laat zien hoe moeilijk het is om een complex biosysteem correct te beschrijven en hoe makkelijk het is om misleid te worden door te geloven in de correctheid van een gegeven model als dat vergeleken wordt met relatief weinig experimentele data. Door zorgvuldige structuuranalyse laten we zien hoe vorige DFT QM/MM berekeningen zoals men die aantreft in de literatuur incorrect zijn. Dit komt volgens ons door een onjuiste beschrijving van de bindingslocatie van de chromofoor; de perfecte overeenkomst van TDDFT met experimenten is daardoor puur toeval. Onze TDDFT/MM berekeningen op onze chromofoor-eiwit complexen geven een absorptiemaximum dat correspondeert met experimenten voor de neutrale, maar niet voor de anionische vormen van GFP. De roodverschuiving van de excitatie, veroorzaakt doordat de chromofoor een proton verliest, wordt enorm onderschat door adiabatische TDDFT dat bijna geen onderscheid aangeeft tussen de neutrale en anionische excitaties. Vervolgens onderzoeken we voor het eerst het gebruik van QMC voor het beschrijven van de excitaties van een chromofoor in zijn eiwit-omgeving en doen we QMC/MM berekeningen van de excitatie energien van de drie vormen van wild-type GFP, op dit moment slechts gebruik makende van een simpele Ansatz voor de veel-deeltjes golffunctie. De experimenteel gevonden verschuiving tussen de verschillende geladen toestanden van GFP wordt goed gereproduceerd door QMC, maar de absolute excitatie-energien worden overschat in vergelijking met experimenten. We laten de eerste stappen zien die we genomen hebben om de mogelijke tekortkomingen van de QM/MM beschrijving van de chromofoor-eiwit wisselwerking te onderzoeken, waarvan we verwachten dat ze de problemen verhelpen als ze gecombineerd worden met meer geavanceerde golffuncties.

Hoofdstuk 5 staat op zichzelf en wijkt af van de hoofdlijn van dit proef-

schrift: het behandelt de coperatieve effecten van π-π en π-anion interacties, een relevant thema binnen de supramoleculaire chemie voor het ontwikkelen van receptoren van anionische typen. In het bijzonder onderzoeken we de geometrische en energetische effecten van π-π stapeling op het anion-π systeem van het bijzondere triazine-triazine-nitraatcomplex, dat recentelijk experimenteel is waargenomen, met behulp van met semi-empirische dispersie gecorrigeerde density functional theory en QMC methoden. We kwantificeren de stabilisatie van de energetische en structurele eigenschappen die genduceerd worden door π-π stapeling en bespreken manieren om dit coperatieve effect verder te versterken, wat nuttig is voor het ontwikkelen van anion-gast structuren.

List of publications

- M. Zaccheddu, C. Filippi, and F. Buda, *Anion-π and π-π interactions regulating the self-assembly of nitrate-triazine-triazine complexes*, J. Phys. Chem A, **112**, 1627 (2008).
- M. Zaccheddu, L. Guidoni, and C. Filippi, *Gas-phase optical properties of the Green Fluorescent Protein chromophore: A theoretical study*, (in preparation).
- M. Zaccheddu, L. Guidoni, and C. Filippi, *Impact of the protein environment on the spectra of Green Fluorescent Protein*, (in preparation).

Curriculum Vitae

I was born on the 14th of June 1978 in Cagliari, Italy. I did my high school studies at the Liceo Scientifico L. B. Alberti in Cagliari, and received my diploma in July 1997. In the same year, I enrolled as an undergraduate student at Cagliari University where I graduated in Physics in April 2003. I conducted my master thesis in the field of dielectric properties of amorphous materials under the supervision of Prof. Vincenzo Fiorentini. From September 2003 to December 2003, I was employed as researcher in the computational solid state physics group in the Physics Department at Cagliari University.

In December 2003, I started my PhD at the Instituut-Lorentz under the supervision of Prof. Claudia Filippi. During my PhD, I presented my work at several international conferences as oral or poster contributions. Particularly relevant were the workshops *Progress in ab initio modeling of biomolecules: Methods and applications* (Leiden, 2006), *Progress in ab initio modeling of biomolecules: Towards computational spectroscopy* (Rome, 2007), and *Workshop on computational material science: Total energy and forces methods* (Trieste, 2007). I also attended several schools as the joint DEMOCRITOS-ICTP School on *Continuum Quantum Monte Carlo Methods* (Trieste, 2004), and the two schools organized by the Dutch Research School of Theoretical Physics (DRSTP) in Nijmegen in 2004 and 2005.

Acknowledgments

A special thank goes to Dr. Leonardo Guidoni who hosted me in his group in Roma for four months at the beginning of of my PhD and taught me all secrets of molecular dynamics and hybrid quantum mechanics in molecular mechanics simulations. I am grateful to Dr. Francesco Buda for involving us in the investigation of novel anion-π-π cooperative interactions. I also want to thank Eneritz Muguruza from King's College London for the long discussions about spectral properties of fluorescent proteins we had during her stay in Leiden. Last, I would like to thank the technical support staff of SARA Computing and Networking Services where I ran most of my calculations: Thanks for the technical help which was always fast and reliable.

It was a pleasure to be part of the Instituut-Lorentz and problems of space make it difficult to thank all the people with whom I shared my days there. I will just mention few of them: Jonathan, Kepa, Chiara, XiaoFeng, Babak, Gianluca, Rachel, and Petja. A special thank goes to Luuk for helping me with the translation of the Samenvatting. I am also very grateful to our secretaries, Fran and Marianne, for the help they gave me to settle in this country and to handle Dutch bureaucracy.

During these four years in Leiden, I met many new people and with some of them I spent some beautiful moments; therefore, I would like to thank all my friends.

I would like to deeply thank my family for the constant support and encouragement.

Finally I want to thank Kiki for all the times she stayed close to me even when I was tired and stressed.