

## Chromophore-Protein Coupling beyond Nonpolarizable Models: Understanding Absorption in Green Fluorescent Protein

This is the peer reviewed version of the following article:

*Original:*

Daday, C., Curutchet, C., Sinicropi, A., Mennucci, B., Filippi, C. (2015). Chromophore-Protein Coupling beyond Nonpolarizable Models: Understanding Absorption in Green Fluorescent Protein. JOURNAL OF CHEMICAL THEORY AND COMPUTATION, 11(10), 4825-4839 [10.1021/acs.jctc.5b00650].

*Availability:*

This version is available <http://hdl.handle.net/11365/982626> since 2015-12-09T13:46:07Z

*Published:*

DOI: <http://doi.org/10.1021/acs.jctc.5b00650>

*Terms of use:*

Open Access

The terms and conditions for the reuse of this version of the manuscript are specified in the publishing policy. Works made available under a Creative Commons license can be used according to the terms and conditions of said license.

For all terms of use and more information see the publisher's website.

(Article begins on next page)

# Chromophore-protein coupling beyond non-polarizable models: Understanding absorption in green fluorescent protein

Csaba Daday,<sup>†</sup> Carles Curutchet,<sup>\*,‡</sup> Adalgisa Sinicropi,<sup>¶</sup> Benedetta Mennucci,<sup>\*,§</sup>  
and Claudia Filippi<sup>\*,||</sup>

*MESA+ Institute for Nanotechnology, University of Twente, P.O. Box 217, 7500 AE Enschede,  
The Netherlands , Universitat de Barcelona, Av. Joan XXIII, s/n 08028, Barcelona, Spain,  
University of Siena, Via A. Moro, 2, 53100 Siena, Italy, University of Pisa, Via Giuseppe Moruzzi  
3, 56124 Pisa, Italy, and MESA+ Institute for Nanotechnology, University of Twente, P.O. Box  
217, 7500 AE Enschede, The Netherlands*

E-mail: carles.curutchet@ub.edu; benedetta.mennucci@unipi.it; c.filippi@utwente.nl

## Abstract

The nature of the coupling of the photoexcited chromophore with the environment in a prototypical system like green fluorescent protein (GFP) is to date not understood and its description still defies state-of-the-art multiscale approaches. To identify which theoretical framework of the chromophore-protein complex can realistically capture its essence, we employ here a variety of electronic-structure methods, namely, time-dependent density functional

---

\*To whom correspondence should be addressed

<sup>†</sup>MESA+ Institute for Nanotechnology, University of Twente, P.O. Box 217, 7500 AE Enschede, The Netherlands

<sup>‡</sup>Universitat de Barcelona, Av. Joan XXIII, s/n 08028, Barcelona, Spain

<sup>¶</sup>University of Siena, Via A. Moro, 2, 53100 Siena, Italy

<sup>§</sup>University of Pisa, Via Giuseppe Moruzzi 3, 56124 Pisa, Italy

<sup>||</sup>MESA+ Institute for Nanotechnology, University of Twente, P.O. Box 217, 7500 AE Enschede, The Netherlands

theory (TD-DFT), multireference perturbation theory (NEVPT2 and CASPT2), and quantum Monte Carlo (QMC) in combination with static point charges (QM/MM), DFT embedding (QM/DFT) and classical polarizable embedding through induced dipoles (QM/MMpol). Since structural modifications can significantly affect the photophysics of GFP, we also account for thermal fluctuations through extensive molecular dynamics simulations. We find that a treatment of the protein through static point charges leads to significantly blue-shifted excitation energies and that including thermal fluctuations does not cure the coarseness of the MM description. While TDDFT calculations on large cluster models indicate the need of a responsive protein, this response is not simply electrostatic: An improved description of the protein in the ground state or in response to the excitation of the chromophore via ground-state or state-specific DFT and MMpol embedding does not significantly modify the results obtained with static point charges. Through the use of QM/MMpol in a linear response formulation, a different picture in fact emerges in which the main environment response to the chromophore excitation is the one coupling the transition density and the corresponding induced dipoles. Such interaction leads to significant red-shifts and a satisfactory agreement with full QM cluster calculations at the same level of theory. Our findings demonstrate that, ultimately, faithfully capturing the effects of the environment in GFP requires a quantum treatment of large photo-excited regions but that a QM/classical model can be a useful approximation when extended beyond the electrostatic-only formulation.

## 1 Introduction

The development of multiscale methods to treat complex environments is at the forefront of research in computational chemistry.<sup>1-6</sup> An important application of these approaches is the description of the excited states of large photosensitive systems, where the excitation of a limited active region is modeled at a high-level electronic structure method and a lower level of theory describes the rest of the system. While several possible combinations of computational ingredients are possible in such hierarchical models, the simplest embedding approach of a static classical en-

vironments in so-called quantum mechanics in molecular mechanics (QM/MM) calculations has seen widespread use for more than a decade.<sup>1,3</sup> There is however mounting evidence that this approach is inadequate for the treatment of excitation energies: Several recent papers using classical point charges in combination with a variety of quantum approaches<sup>7–13</sup> have reported significant errors in the computed excitation energies with respect to experimental absorption maxima in different photoactive proteins.

There are essentially two main paths one may follow to alleviate the shortcomings stemming from the inadequacy of classical point charges. One may either employ a more sophisticated but still classical environment or attempt to treat parts of the environment at the quantum level. In the first class of models, the use of a polarizable embedding (MMpol)<sup>14–17</sup> is gaining popularity: Within this framework, the introduction of induced dipoles appears to be a quite effective strategy in describing electronic excitations<sup>13,18–21</sup> and offers the appealing feature that different polarization effects induced in the ground state or responding to the excitation of the embedded system can be systematically studied.<sup>22</sup> The second class of models, seeking to treat more atoms at a quantum level, has several possible members. In the so-called “mechanical embedding” schemes, the high-level quantum calculation is only performed on an isolated part of the total system and determines a correction to a low-level supermolecular quantum calculation.<sup>23</sup> A self-consistent quantum method is sub-system density functional theory (DFT) embedding, where the environment density creates a one-electron embedding potential to be used in a high-level calculation of the active sub-system. Similarly to the case of polarizable dipoles, ground-state<sup>24,25</sup> and state-specific<sup>26</sup> schemes are available to study the effect of the environment response. Finally, a more “brute-force” approach<sup>10,13,27</sup> is to simply increase the quantum region to include part of the protein environment and treat the whole cluster at the same level of theory. The clusters might however be much larger than what is viable with most quantum methods: A recent study<sup>8</sup> reported that more than 700 quantum atoms are necessary to converge the excitation energy of the photoactive yellow protein.

Here, we focus on wild-type *Aequorea* green fluorescent protein (GFP), as a particularly ideal case, which challenges our understanding of how the excitation of the photoreceptor (the chro-

mophore) couples to the protein environment, and of what the necessary ingredients in a computational method are to describe its photophysics. The importance of GFP as photobiological system stems from being the progenitor of the large family of intrinsically fluorescent proteins<sup>28,29</sup> which have launched a revolution in cell biology being compatible with non-invasive imaging in living cells. More recently, the development of fluorescent proteins with targeted photochemical properties has also enabled remarkable advances in super-resolution bio-imaging techniques capable of beating the diffraction limit.<sup>30,31</sup> Relevantly to the purpose of the current study, the environment surrounding the chromophore in these proteins is known to play a critical role in tuning its excited-state behavior: Through the mutagenesis and continuous discovery of GFP-like proteins in different sea organisms,<sup>28</sup> this class of proteins now covers almost the entire visible spectrum both in emission and absorption. On the other hand, predicting the relation between relatively fine structural changes in the chromophore pocket and the spectral properties of the chromophore-protein complex remains a very demanding task for multiscale approaches. Several recent studies<sup>12,19,32</sup> have attempted a computational characterization of these photoactive systems but reproduced trends in absorption between the various proteins with limited degree of success.

For wild-type GFP, in particular, many computational articles have appeared in the literature<sup>7,9,12,13,18–21,27,32–40</sup> due to the importance of the system and abundance of experiments characterizing its structural and spectral properties. The bulk of these studies produced however excitation energies blue-shifted with respect to absorption experiments when using a classical static treatment of the protein.<sup>7,9,12,32,34–40</sup> Consequently, the focus has recently shifted towards improving the description of the environment either through polarizable dipoles<sup>18–21</sup> or simply by increasing the quantum region and therefore treating all atoms at the same level of theory.<sup>10,27</sup> While both approaches appear to improve on experimental agreement, many fundamental questions remain unanswered that transcend the choice of GFP as system of study: Does the excitation of the chromophore polarize the environment and how long range is this effect? Is the coupling between the active site and the protein of electrostatic nature or do we ultimately need a fully quantum mechanical description? Are there clear signatures indicating which computational route one

must follow in modeling other photobiological systems?

Here, we seek to provide robust answers to these questions and, to this aim, employ a large battery of tools to describe the chromophore-protein complex. We will begin with extensive QM/MM molecular dynamics simulations and generate a large number of representative frames (about 100 in total) to investigate the interplay between structure and excited states of both protonation states of GFP. Analyzing numerous frames allows us to faithfully assess temperature effects, explore the main factors that determine the spread of absorption energies, and avoid the danger of drawing far-reaching conclusions on a single configuration of the system. We will consider the three main approaches to multiscale modeling of excited states, namely, static point charges, polarizable dipoles, and DFT embedding (the latter being here applied for the first time to fluorescent proteins) and, for many configurations sampled from the dynamics, either compare their results to supermolecular calculations on very large clusters or change the quantum method within the same hybrid scheme to understand how the response varies with the sophistication of the quantum treatment. For this purpose, we will perform calculations with several different electronic-structure methods: Time-dependent density functional theory (TDDFT) with two different functionals and three wave-function methods, that is, the complete active space perturbation theory (CASPT2), *n*-electron valence state perturbation theory (NEVPT2), and quantum Monte Carlo (QMC). In particular, we will explore the nature of polarization effects on the excitation energies of GFP using a frozen and a polarizable environment both at the level of DFT and MMpol embedding to identify the origin of the failure of a static classical description. Finally, we will establish the validity of representing the whole protein as a small cluster in several different ways; this analysis has been lacking in the bulk of previous literature.

The remainder of this paper is organized as follows: In Section 2, we describe the computational details and, in Section 3, we present our main results. We discuss them and draw our conclusions in Section 4.

## 2 Computational details

For the QM/MM simulations,<sup>41,42</sup> we use the CPMD 3.17.1<sup>43</sup> and Gromos96<sup>44</sup> codes with the Amber 03<sup>45</sup> force field and the TIP3P<sup>46</sup> water model for the MM atoms. We treat the QM region with density functional theory (DFT) and employ the PBE<sup>47,48</sup> exchange-correlation functional, the Martins-Toullier pseudopotentials<sup>49</sup> and a plane-wave cutoff of 70 Ry. The size of the quantum box is such that the distance between periodic replicas is at least 8 Å. We perform the Car-Parrinello molecular dynamics (MD) simulations with a time step of 4 a.u. (about 0.1 fs) and a fictitious electron mass of 400 a.u. For the first 0.25 ps of the room-temperature (300 K) runs, we use the Berendsen thermostat<sup>50</sup> with a time constant of 10 a.u. We then switch to a Nosé-Hoover thermostat<sup>51,52</sup> with a characteristic frequency of 2000 cm<sup>-1</sup> and default settings for the rest of the 27 and 25 ps MD simulations for B and A form, respectively. In the annealing runs, we rescale the velocities at each step by 0.999 until the temperature is below 1 K for both forms.

We use Molcas 7.4 and 7.8<sup>53</sup> for the CASPT2 calculations and a module adapted from the Molcas-Embed interface developed by Carter and co-workers<sup>54,55</sup> for the CASPT2/DFT embedding runs. We employ the ANO-S-VDZP<sup>56,57</sup> basis set and, in the convergence tests, the aug-cc-pVDZ.<sup>58</sup> We use the Cholesky decomposition of the two-electron integrals<sup>59</sup> with a threshold of 10<sup>-4</sup>. In the CASPT2 calculations, we always adopt the default IPEA zero-order Hamiltonian<sup>60</sup> with an additional constant imaginary shift<sup>61</sup> of 0.1 a.u. and, unless otherwise noted, report the excitation energies computed at the single-state level. We also freeze as many  $\sigma$  orbitals as there are heavy atoms. In the MM calculations, we describe the protein with the Amber 99<sup>62</sup> force field and set the charges of the charge group closest to the MM boundary to zero, redistributing them to the next charge group, weighted by nuclear charge. All CASSCF calculations are carried out in state average over two states. We list the CAS spaces employed in Section S4 of the SI together with additional tests on the inclusion of a third state in the state average.

We also perform CASPT2/MMpol calculations in the presence of a polarizable environment using the Amber pol12 parameters (AL model in Ref.<sup>63,64</sup>) to describe the protein and the water solvent. For water, we derived the point charges from a standard RESP fit using the electrostatic

potential computed at the MP2/aug-cc-pVTZ level on the TIP3P water geometry ( $q_O = -0.726$ ,  $q_H = 0.363$ ). These CASPT2/MMpol results are estimated using a two-step strategy recently presented,<sup>65</sup> where the MM induced dipoles and charges from a CASSCF/MMpol calculation performed using the Gaussian code are later used in Molcas as a static external potential to obtain the CASPT2 results. Such CASSCF/MMpol calculations, performed using a locally modified version of Gaussian09, revision A.02,<sup>66</sup> are obtained using the state-average procedure for the ground and first excited states and adapting the MM polarization either to the ground state (polGS), or to the ground and excited states in two separate calculations (polSS).

The time-dependent density functional theory (TDDFT) runs are also performed using a locally modified version of Gaussian09, revision A.02,<sup>66</sup> and with the CAM-B3LYP<sup>67</sup> and LC-BLYP<sup>68</sup> functionals. We use the AL Amber pol12 parameters<sup>63,64</sup> for the polarizable dipole calculations and Amber 99 for the non-polarizable runs for consistency with CASPT2/MM. We employ the 6-31G<sup>69</sup> basis set on the hydrogens and the 6-31+G(d)<sup>69-71</sup> for the other atoms. In all QM/MMpol calculations, the polarizability of the MM atoms located at distances greater than 20 Å to any QM atom is set to zero in order to reduce the computational cost, whereas all partial charges were included. This value of polarization cutoff has been shown to provide highly converged excitation energies.<sup>72</sup>

To generate the embedding potentials for the CASPT2/DFT and TDDFT/DFT calculations, we use the ADF 2013.2 code<sup>73-75</sup> with the Slater DZP<sup>76</sup> basis set. We employ the M06HF<sup>77</sup> exchange-correlation functional for the intramolecular calculations, and the PW91k<sup>78</sup> kinetic functional and the PW91<sup>79</sup> exchange-correlation potential for the non-additive part of the embedding potential. For frozen-density embedding TDDFT/DFT calculations, we use the ADF program with the same embedding potentials used in the CASPT2 runs and the CAMY-B3LYP functional (where the separation of  $1/r$  is based on a Yukawa potential).<sup>67,80,81</sup> The impact on the excitation energies of the use of CAMY-B3LYP/DZP instead of CAM-B3LYP/6-31+G(d) is discussed in Section S8 of the SI.

For the NEVPT2 calculations in the strongly contracted formulation,<sup>82-84</sup> we use ORCA 3.0.2<sup>85</sup>



with the ANO-L-VDZP<sup>56,57</sup> basis set and the aug-cc-pVTZ<sup>58</sup> as auxiliary basis set for the resolution of identity.<sup>86</sup> The RIJCOSX approximation<sup>87</sup> is employed in the CASSCF step. The orbital energies for the doubly occupied and virtual orbitals appearing in the Dyall Hamiltonian<sup>88</sup> (used for the definition of the zero order Hamiltonian in NEVPT2) were obtained by the diagonalization of a generalization of the Fock operator<sup>89</sup> (canonical orbital option in ORCA). In the construction of the third- and fourth-order density matrices, the CASSCF wave function is truncated so that configurations with lower weight than a threshold of  $10^{-8}$  are discarded. The NEVPT2/MM runs are performed in the presence of the same point charges as in the CASPT2 calculations.

For the QMC calculations, we use CHAMP<sup>90</sup> with scalar-relativistic energy-consistent Hartree-Fock pseudopotentials and the corresponding cc-pVDZ basis sets<sup>91,92</sup> augmented with diffuse s and p functions on the heavy atoms.<sup>93</sup> We employ a two-body Jastrow factor to account for electron-electron and electron-nucleus correlations and use different Jastrow factors to describe different atom types.<sup>94</sup> The starting orbitals are obtained in a state-average CASSCF calculations with Molcas 7.8 (in the presence of the MM charges, if appropriate) and are not optimized in the QMC runs. We truncate the CAS expansion in state-average natural orbitals with a cutoff on the CI coefficients as detailed in Section S4 of the SI. The CI coefficients and Jastrow parameters are optimized with the linear method<sup>95</sup> in a state-averaged fashion<sup>96</sup> within variational Monte Carlo (VMC). The pseudopotentials are treated beyond the locality approximation<sup>97</sup> and an imaginary time step of 0.05 or 0.075 a.u. is used in the diffusion Monte Carlo (DMC) calculations. The potential generated by the external MM charges is put on a grid which is padded to have at least 5.0 Å distance from any atom in the chromophore and has a step size of 0.2 a.u. A finer grid with a step size of 0.1 a.u. yields VMC excitation energies compatible to better than 0.02 eV. The electron loss ratio (attempted evaluations outside the cubic grid) is less than  $10^{-6}$  in all runs. All QMC excitation energies presented below are computed within DMC and the VMC results are reported Section S11 in the SI.

## 2.1 Models

To prepare the structures of the A and B forms, we start from the protein models obtained in our previous study.<sup>9</sup> For the A form, we remove one GFP barrel from the previously used dimeric system as well as all water molecules which are more than 5 Å from the remaining monomer. We then solvate the system and bring the added water in equilibrium by performing a 5 ns NPT classical MD simulation with NAMD<sup>98</sup> while keeping all other atoms fixed. The solvation box is approximately  $72 \times 84 \times 82 \text{ Å}^3$  for the A form and  $70 \times 65 \times 70 \text{ Å}^3$  for the B form.

In the QM/MM simulations, we set the boundary between the QM and MM regions at the single bonds  $\text{C}_{\text{OOH}}-\text{C}_\alpha$  of Phe64 and  $\text{N}-\text{C}_\alpha$  of Val68 as shown in Fig. S1 of the SI. With this QM part of 40 (A form) or 39 (B form) atoms, we run long QM/MM room-temperature simulations of 27 and 25 ps for the A and B form, respectively. We then perform a cluster analysis<sup>99</sup> with VMD<sup>100</sup> on 4000 frames sampled every 5 fs from the last 20 ps of the QM/MM trajectories. Each cluster is populated by frames that are within a certain cutoff distance of the central frame of the same cluster. As distance between two frames, we employ the mass-weighted fitted RMSD where the frames are first aligned to obtain the lowest possible RMSD. In the computation of the distance between two frames, we include all the atoms in the protein but not the water molecules, and use a cutoff distance of 0.5 Å for the A form and 0.6 Å for the B form to define a cluster. The central frames of the most populated 10 clusters from the cluster analysis are then used in the calculation of the excitation energies. For further analysis, we isolate 50 equidistant frames sampled every 400 fs from the QM/MM MD runs and also determine a final annealed structure for both forms of the protein.

For the B form, we also consider a larger QM region, which includes the chromophore and the residues Gln94 and Arg96, and the boundary between the QM and MM regions also includes the single bonds  $\text{C}_{\text{OOH}}-\text{C}_\alpha$  of Val93,  $\text{N}-\text{C}_\alpha$  of Thr97, and  $\text{C}_\alpha-\text{C}_\beta$  of Glu95 (the backbone of Glu95 is modeled at the quantum level but its sidechain classically). With the resulting QM region, containing 93 atoms, 5 of which are capping hydrogens, we perform a QM/MM simulation of 1.5 ps and determine then an annealed structure.

Finally, we construct QM cluster models for the TDDFT, CASPT2/DFT, and TDDFT/DFT calculations. The smallest cluster contains 168 atoms from 10 residues (with deep cuts at  $C_\beta$  or even deeper) and 8 water molecules, and is similar to the models used in Ref.<sup>12</sup> For the three larger clusters of 279, 345, and 529 atoms (4 waters and 8 residues, 7 waters and 10 residues, and 7 waters and 19 residues, respectively), we use capping methyl groups and cut the residues at the neighboring  $C_\alpha$  atoms in the backbone. The list of residues included in the clusters is given in Table S1 of the SI.

### 3 Results

In equilibrium, the chromophore of wild-type GFP exists in either a neutral A form or an anionic B form. Upon excitation in the blue, the A form of the chromophore loses the phenolic proton converting to an intermediate anionic I form, which in turn infrequently transforms into the thermodynamically stable B form.<sup>101</sup> The neutral chromophore is known to transfer its proton to Glu222 through the wire CRO-Wat-Ser205-Glu222<sup>102,103</sup> but other differences in the surroundings are difficult to assess experimentally: Since the predominant form of wild-type GFP at room temperature is the neutral one, the structure of the anionic form cannot be established through X-ray diffraction. As workaround, one can adopt as starting model a suitably modified X-ray structure of the mutant S65T which stabilizes the anionic form as also done in earlier computational work.<sup>9,104</sup> The available X-ray structures of this mutant (PDB entries 1EMA, 1EMG, and 1Q4A<sup>105–107</sup>) suggest that the hydrogen-bond network of the B form is similar to that of the A form except for a different orientation of Thr203. A recent theoretical study<sup>40</sup> challenges the accepted orientation of Thr203 and also suggests a new conformation of the anionic chromophore. We will explore this possibility and its consequences in Subsection 3.5.

The solvated chromophore of GFP does not exhibit fluorescence at room temperature (quantum yield of less than 0.001<sup>108</sup>) but fluoresces in solution at 77 K<sup>109,110</sup> and also in the protein, being rigidly held by the environment which inhibits its numerous available modes of relaxation.<sup>110–112</sup>

Temperature effects on the chromophore and its surroundings have nevertheless an impact on the excited-state properties of GFP: When temperature is reduced, the absorption maximum moves from 2.59 to 2.63 eV for the B form, and from 3.12 to 3.05 eV for the A form and, consequently, the distance between the peaks of two forms decreases from 0.5 to 0.4 eV. Furthermore, the width of the peaks is much smaller in the low-temperature than in the room-temperature spectrum.<sup>113</sup> We begin our study by investigating the effects of temperature on the chromophore-protein complex of both forms of GFP and explore the range of conformations visited by the system in our QM/MM MD trajectories.

### **3.1 Thermal stability of hydrogen-bond network**

For the A form, the hydrogen-bond network around the chromophore is quite stable during the 27 ps of QM/MM dynamics. In particular, the CRO-WAT-Ser205-Gln222-CRO wire is always connected except for a very short instance in the equilibration part where the water loses the H-bond with Ser205 (see Figure 1 for the position of the residues with respect to the chromophore and the labeling). His148 is always bonded to the water close to the phenolic oxygen and also the imidazolinone ring has a stable environment. In particular, O12 is bonded to Arg96, Gln94, and a water molecule, while N2 is weakly bonded to a water molecule. Only the O12-water interaction is sporadically broken. The annealed frame has the same hydrogen-bond network as the one in the dynamics.

For the B form, we observe instead that the hydrogen-bond network adopts several configurations even during the relatively short 25 ps dynamics as can be seen in Figure 2. The CRO-WAT-Ser205-Glu222-CRO wire is occasionally broken as the water molecule loses the bond to the phenolic O1. There are in fact four residues around O1 that compete to form hydrogen bonds with it: Tyr145, His148, Thr203, and a water molecule as illustrated in Figure 1. Through the trajectory, either two or three of these four residues are bonded to the phenolic oxygen. In particular, in most of the trajectory, exclusively one of the two residues Tyr145 and His148 (see Figure 2) is bonded to the oxygen, while Thr203 is bonded most of the time only rarely losing the bond due to the

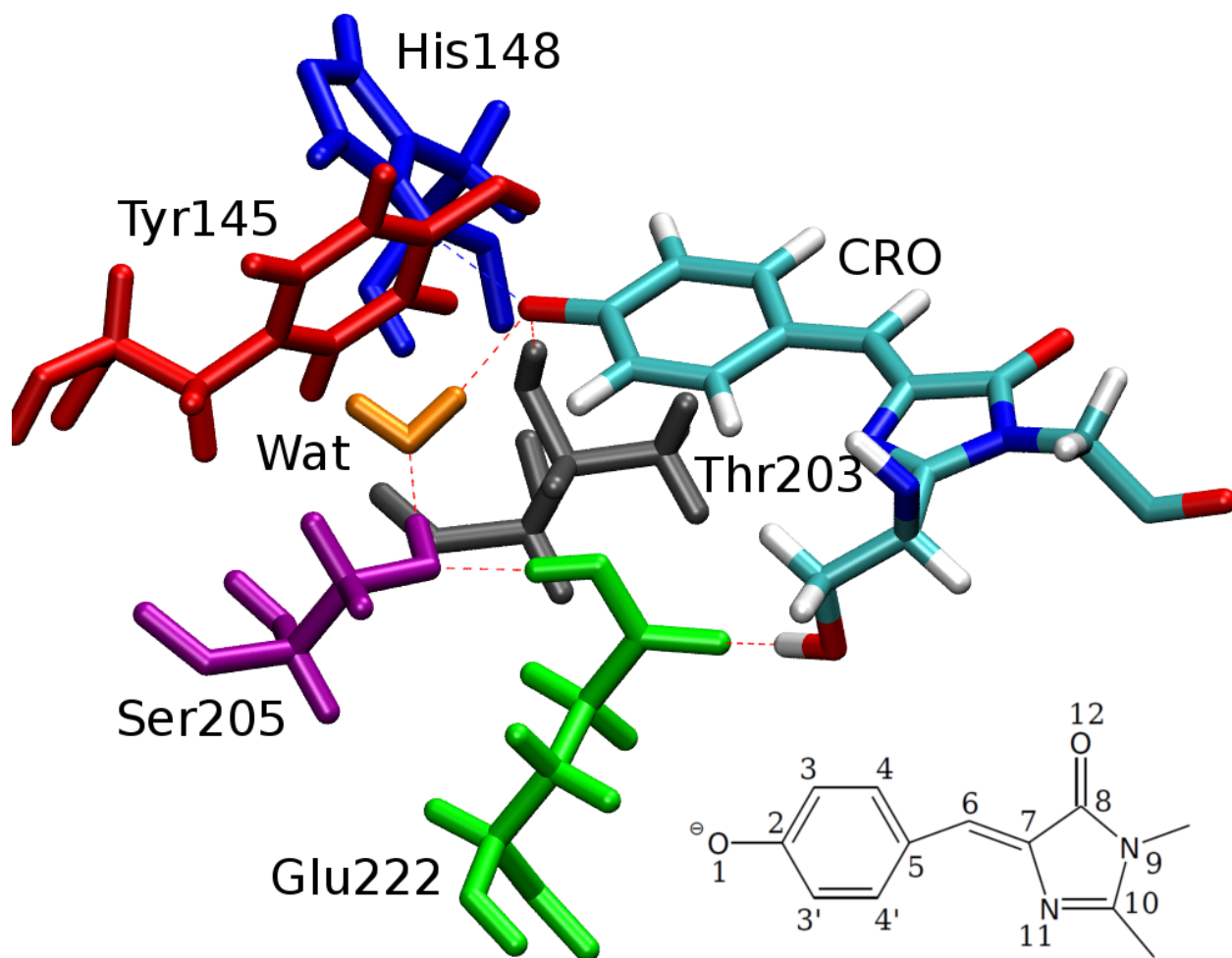


Figure 1: B form: The hydrogen-bond network around the chromophore (CRO). The phenolic oxygen O1 is bonded to the water molecule (orange), His148 (red), Thr203 (gray), while Tyr148 (red) does not participate in this particular snapshot. The hydrogen-bond network is completed by Ser205 (purple) and Glu222 (green). The hydrogen-bond network of the A form is similar but Glu222 loses its proton to O1 and Thr203 is no longer bonded to CRO owing to a change in conformation.

movement of the H atom (the heavy atoms of the residue are always at a distance consistent with a hydrogen bond). Finally, we note that the water molecule close to the phenolic oxygen O1 forms a bond with Thr203 for a short time interval, losing the one with O1 but keeping the other hydrogen bonded with Ser205. The annealed frame of the B form has Tyr145 bonded to the phenolic O1 even though, for most of the dynamics, it had His148 instead. This simply indicates that multiple close minima are available to the system. In fact, the annealing run with the larger quantum region results instead in Tyr145, His148, and the water bonded to the phenolic ring, while Thr203 forms a bond with the carbonyl group in His148, mainly due to the movement of the gamma hydrogen atom in the threonine residue. Consequently, the commonly used approach of using a single frame obtained through annealing or optimization can lead to a configuration which is not representative of the average behavior.

### 3.2 Relationship between structure and excitation energies

To analyze the excitation energies over a wide range of representative conformations of the protein, we choose 50 equidistant frames from the last 20 ps of the QM/MM trajectories of both the A and B forms (i.e. one frame every 500 fs). Given the significant number of frames, we perform a first screening of the behavior of the excitation energies over these frames using the relatively cheap TDDFT/MM method. We choose CAM-B3LYP as exchange-correlation functional since it has been shown<sup>114,115</sup> to ameliorate problems with spurious charge-transfer effects that may plague conventional hybrid methods.<sup>116,117</sup> Later in this paper, we will also employ more sophisticated methods to treat either the chromophore or the protein environment and focus then on 10 representative frames obtained from cluster analysis for each protonation state. Our findings on the 50 frames for the A and B forms are summarized in Figure 3. The A form exhibits a spread in the TDDFT excitation energies of more than 0.5 eV and a standard deviation of 0.1 eV, while the B form has a smaller but still considerable spread of almost 0.4 eV and a slightly smaller standard deviation of 0.07 eV. The extent of values for the A form is particularly surprising since the protein environment is very stable during the dynamics as noted above.

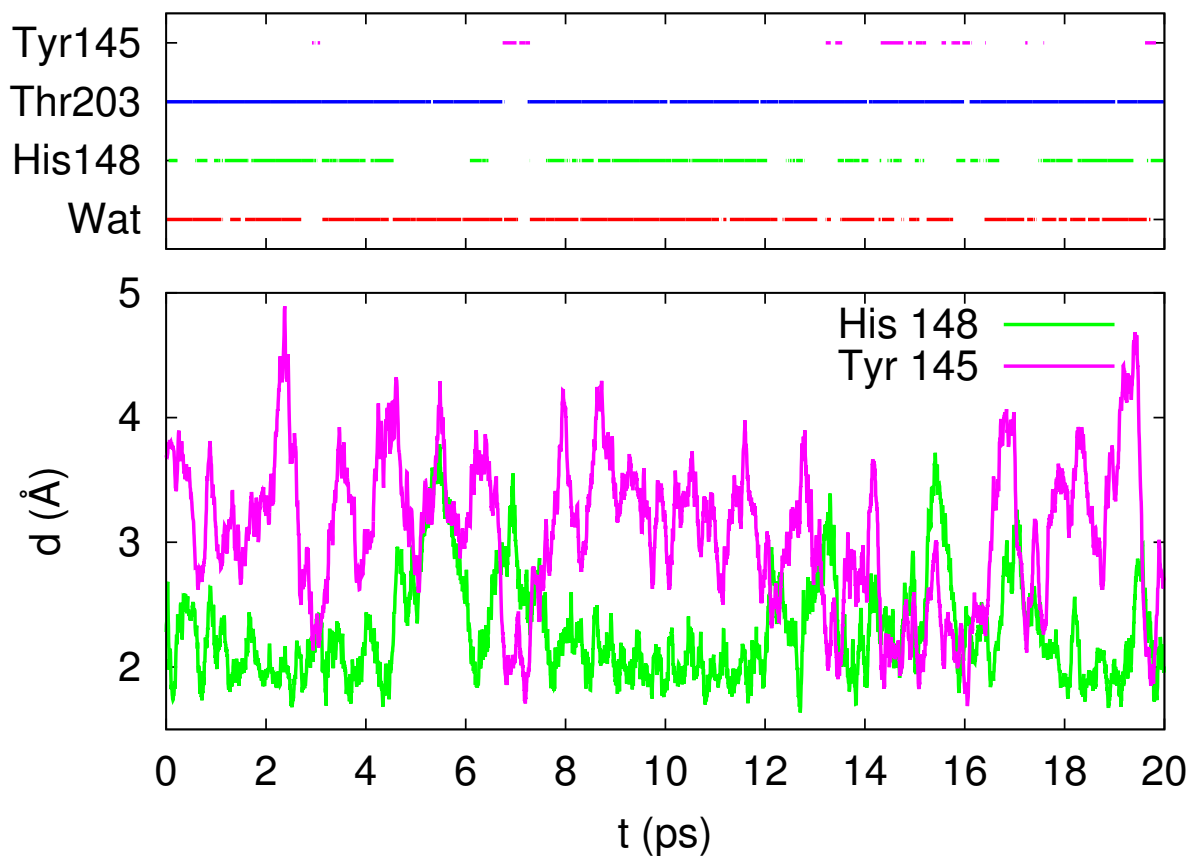


Figure 2: B form: The hydrogen bonds formed with the phenolic oxygen of the chromophore during the QM/MM trajectory as a function of time. Top: The bonding of the four possible residues (when the distance between their hydrogen atom and O1 is less than 2.3 Å). Bottom: The distances between O1 and the bonded hydrogen atoms of Tyr145 and His148 as a function of time.

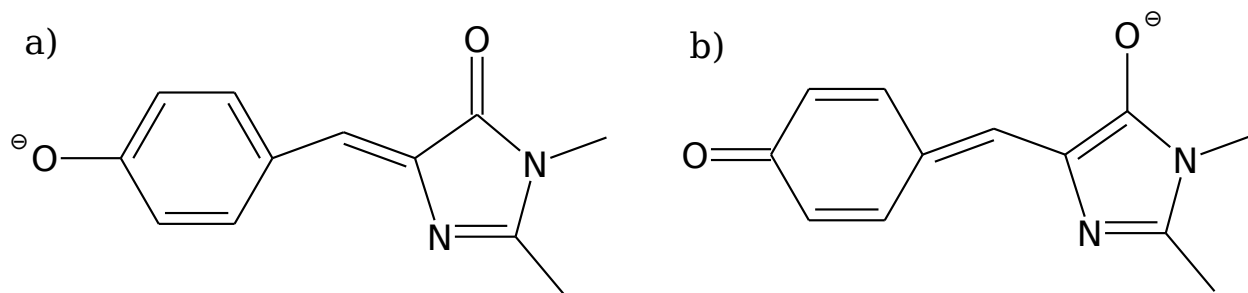
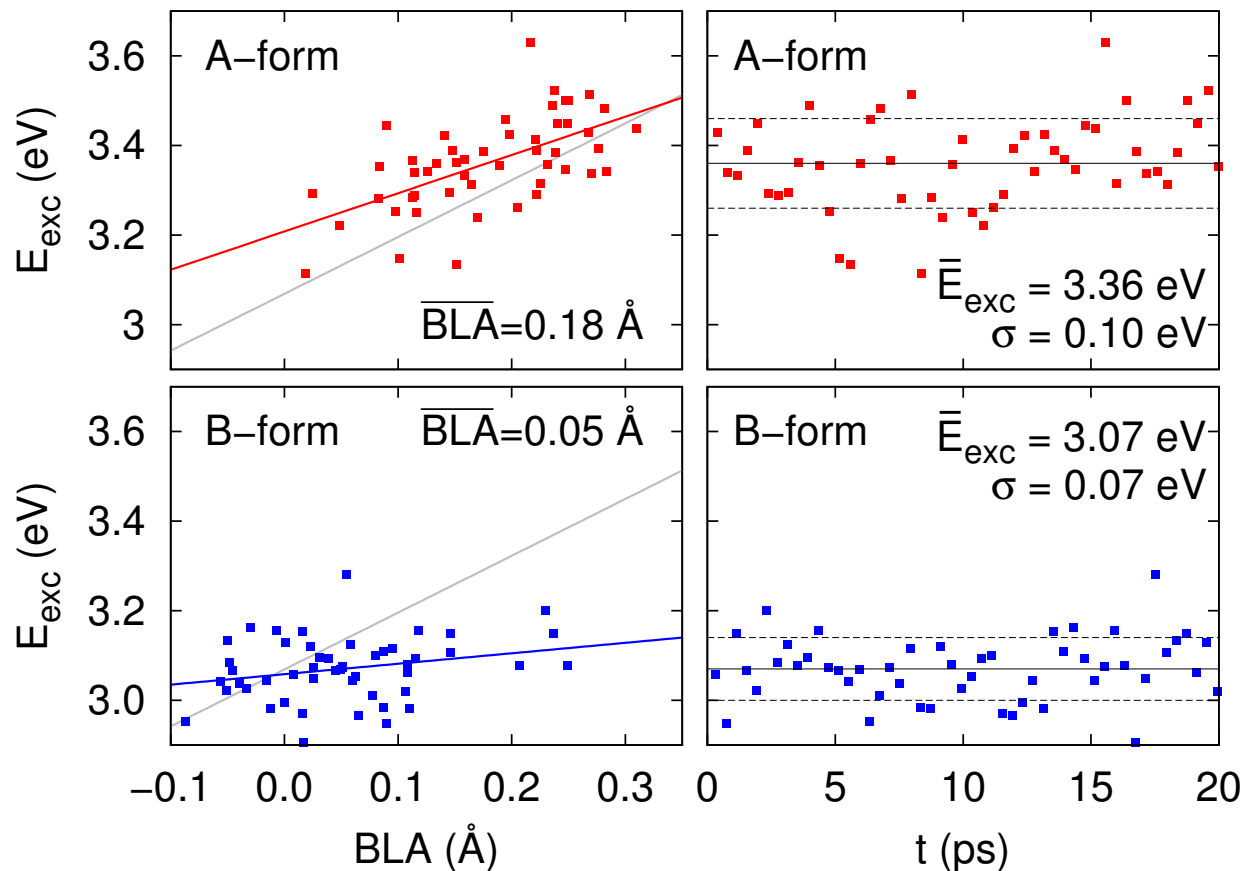


Figure 4: The a) benzenoid and b) quinoid resonance structures of the anionic chromophore.



The origin of these large variations is worth exploring. Previous studies<sup>12,34,118</sup> have shown that the excitation energies of fluorescent proteins depend on the bond-length alternation (BLA) within the chromophore. The BLA was then defined as a linear combination of the 15 heavy-atom bonds in the chromophore, whose weights were fitted to the excitation energies in various ways. We also perform such an analysis but use a much simpler definition of the BLA than in previous works, namely,

$$\begin{aligned}
 \text{BLA} = & \left[ \text{O1C2} + \frac{1}{2} * (\text{C3C4} + \text{C3'C4'}) \right. \\
 & + \left. \text{C5C6} + \text{C7C8} \right] \\
 & - \left[ \text{O1C8} + \frac{1}{2} * (\text{C2C3} + \text{C2C3'}) \right] \\
 & + \frac{1}{2} * (\text{C4C5} + \text{C4'C5}) + \text{C6C7} \quad (1)
 \end{aligned}$$

where the atom labeling is given in Figure 2. The significance of this parameter can be understood in terms of the two resonance structures of the anionic chromophore shown in Figure 4: The BLA defined above is positive when the chromophore is in the dominant benzenoid structure and negative in the quinoid one. Such a simple definition avoids the danger of over-fitting our data and, as shown in Figure 3, suffices to describe a sizable part of the tuning in the A form. In this case, since the chromophore adopts exclusively the benzenoid resonance structure, when the BLA approaches zero through thermal fluctuations, the electronic structure of the ground state is severely perturbed, thereby closing the gap. On the other hand, for the anionic B form, small values of the BLA are more typical since the quinoid structure acquires importance and the ground state is more capable of adjusting also to conformations with a negative BLA and a dominant quinoid resonance structure. The dependence on the BLA of the excitation energy is therefore much less steep as shown in Figure 3.

We also note that the quality of the fit for the B form is lower, indicating that other factors than the BLA affect its excitation energy. We can in fact improve the fit if we account for environmental effects by adding for instance fitting parameters which measure the presence or lack of a bond

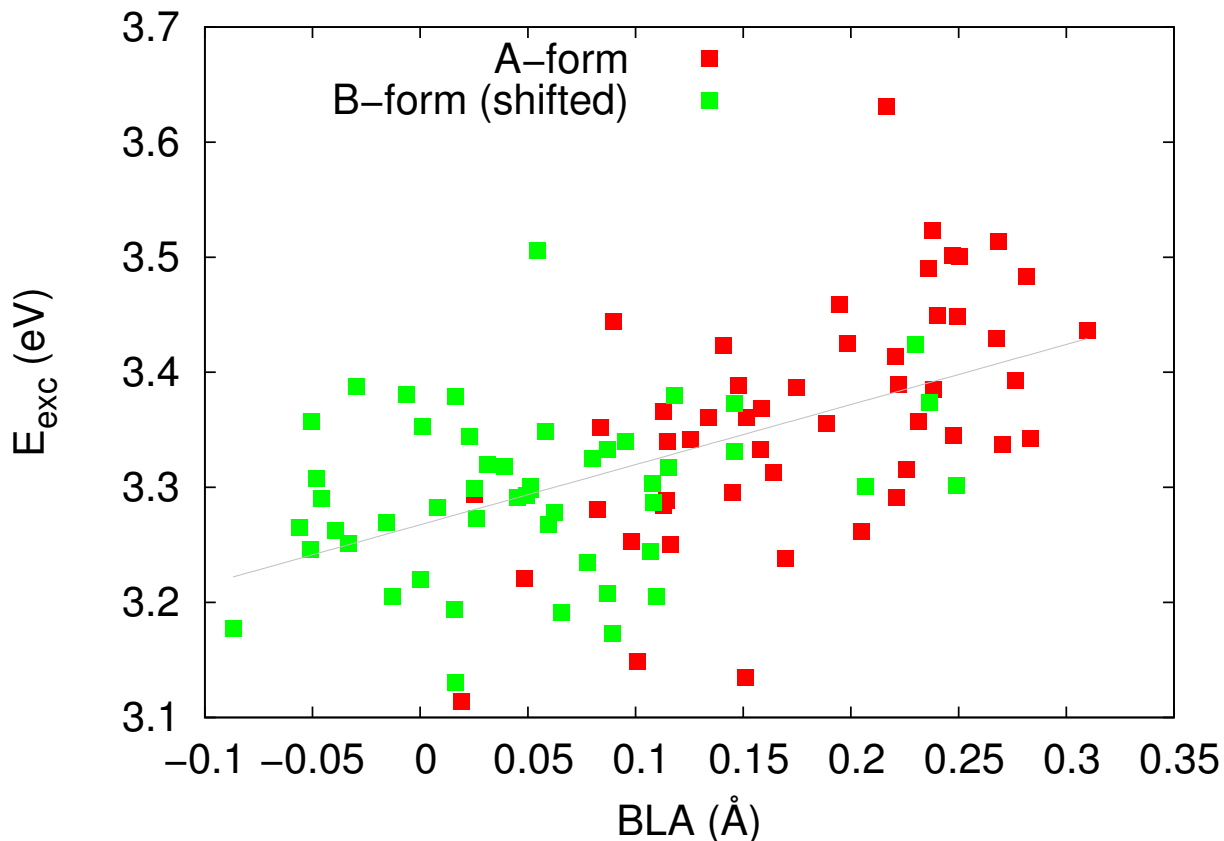


Figure 5: A and B forms: TDDFT/MM excitation energies as a function of the BLA. A shift between the two forms is introduced as parameter to yield the best linear fit, the value (A-B) being 0.22 eV.

between O1 and His148 and/or Tyr145 (see section S3 of the SI). Even though more data would be required to elucidate the precise role of the different amino acids in the surrounding of the chromophore, our analysis suggests that the spread of the absorption energies has different causes in the two forms: For the A form, it is mostly determined by the internal coordinates (which in turn are coupled to the external conformations) and, for the B form, the flexible hydrogen-bond network around the chromophore seems to be a more directly relevant factor.

While the average BLA is smaller for the B form than the A form, this fact alone is not enough to explain why the A form has a higher excitation energy. If we attempt to fit the dependence of the excitation energies on the BLA for both forms together as done for instance in Ref.,<sup>34</sup> we observe that the excitation energies of the A form are systematically higher than those of the B form as

is evident from Figure 3 and that a single fit is inadequate (we reach the same conclusion if we allow arbitrary weights for the bonds in the fitting procedure). We therefore add an energy shift between the two forms in the linear model to account for this offset, which is found to be 0.22 eV and improves the quality of the fit as shown in Figure 5. Consequently, of the total shift of 0.29 eV between the average excitation energies of the A and B forms, only 0.07 eV is accounted for by a lower average BLA in the B form than in the A form and the remaining 0.22 eV is caused by the different electronic structure of the two protonation states. We note in passing that, for both forms, the excitation energy displays the strongest dependence on the bridging distance C6-C7 (double bond in the A form and predominantly double bond in the B form) when we search for correlations between the excitation and any single bond length, while the other bridging distance, C5-C6, has a less significant influence. This is in line with the Hückel model description<sup>37</sup> for the excitation energy of the A form, which is approximated as a simple ethylene-like HOMO-LUMO transition on the C6-C7 bond. However, the same model assumes the bonds C5-C6 and C6-C7 to be equivalent in the B form, which does not fit our findings. Further insight on the electronic structure of the chromophore and relation to its geometry can be obtained through the development of alternative simplified models as in Refs.<sup>119–121</sup>

In agreement with previous studies,<sup>12,19,39,122</sup> we find that CAM-B3LYP gives blue-shifted values for the excitation energies of these photoactive systems. For the B form, the average excitation energy over 50 frames is  $3.07 \pm 0.01$  eV, which is almost 0.5 eV blue-shifted with respect to the experimental absorption maximum of 2.59 eV. Furthermore, the dependence of the excitation energies on the BLA is not as clear as for the A form and attempts to correlate the spread in the energies with environmental changes gave inconclusive results. Given these relatively unclear findings, one can ask whether the observed trends are physically meaningful or just noise due to the use of approximate TDDFT. To answer this question, we carry out CASPT2/MM calculations on all 50 frames of the B form and compare the resulting excitation energies with those obtained with CAM-B3LYP/MM. As illustrated in Figure 6, the conclusion is unequivocal: The two methods are in excellent agreement as regards general trends. While these calculations reinforce our preceding

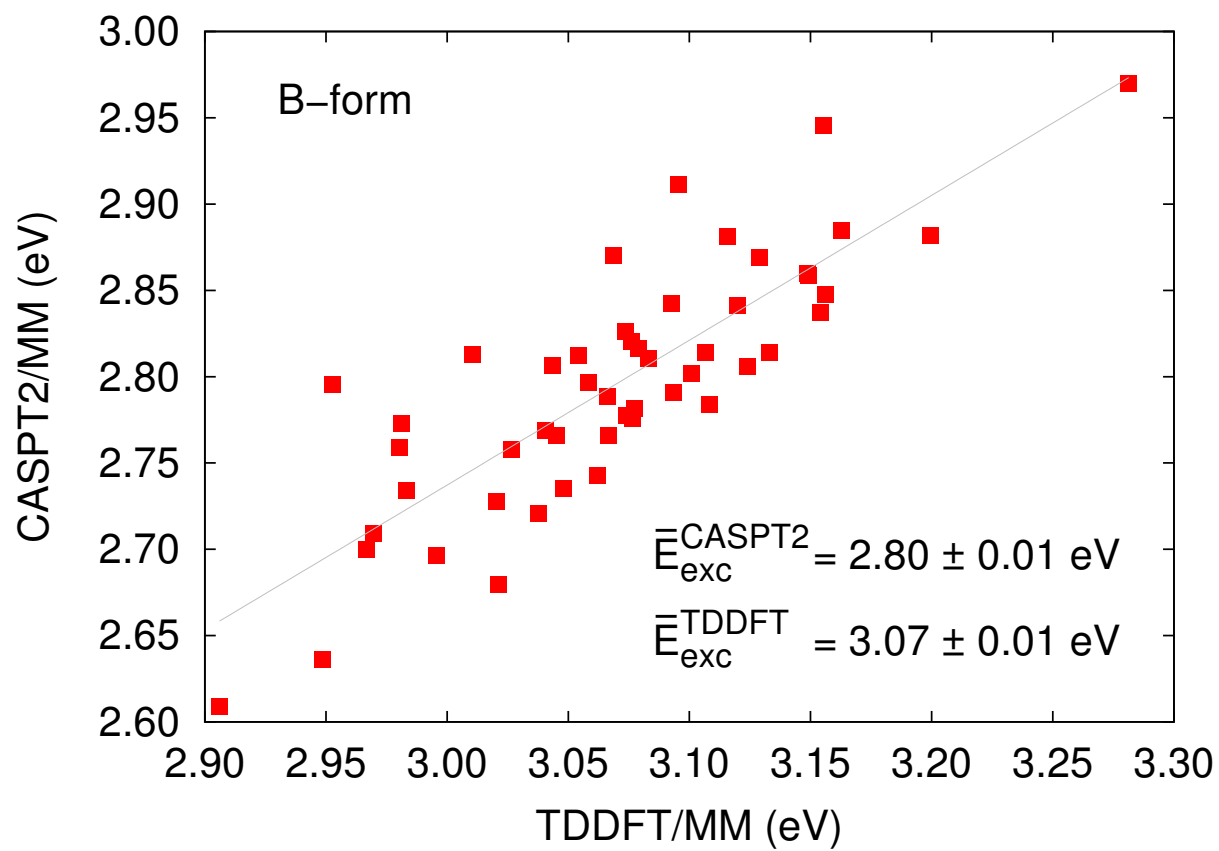


Figure 6: B form: Comparison of the TDDFT/MM and CASPT2/MM excitation energies computed on the 50 equidistant frames of the QM/MM MD simulation.

analysis (the BLA dependence is qualitatively identical for CASPT2 energies), the CASPT2/MM average excitation energy of  $2.80 \pm 0.01$  eV is still clearly blue-shifted compared to experiment. This means that thermal sampling by itself is not sufficient to overturn the strong indications coming from previous studies<sup>9,11,39,123</sup> that describing the protein environment as only classical static point charges produces an unphysical blue shift in the excitation energies of related photoactive systems.

To verify that the choice of quantum method is not responsible for the blue shift, we extend our investigation to other computational approaches for 10 representative frames obtained through cluster analysis. We employ the LC-BLYP functional in the TDDFT calculations and also use quantum Monte Carlo and NEVPT2 as alternative wave function-based methods. To pinpoint the effect of the MM environment, we consider the difference between the QM/MM description and the excitation energies computed on the isolated chromophore at the geometry obtained within the protein. As shown in Figure 7, the trends across the frames are qualitatively similar at all levels of theory but the shifts induced by the protein are larger for the wave function approaches. As regards the absolute excitation energies with the other WF methods, we find that NEVPT2 and QMC are rather close and about 0.2 eV blue-shifted compared to CASPT2 (see Table S6 in the SI for the excitation energies on all frames), thereby further worsening the agreement with experiments. This finding renders even more pressing the search for an alternative, better description of the environment, which will be the focus of the following subsection.

### 3.3 Improved description of the environment

To further investigate whether the MM description of the protein is responsible for the significantly blue-shifted excitation energies with respect to experiment, we explore here the use of other embedding methods to improve on the simple MM scheme with static point charges. A possible strategy is that of adding a polarizable embedding (MMpol) in terms of induced dipoles.<sup>15</sup> From the QM/MM trajectories of the A and the B forms, we choose 10 representative snapshots using cluster analysis and repeat the TDDFT investigation for these frames with induced dipoles in

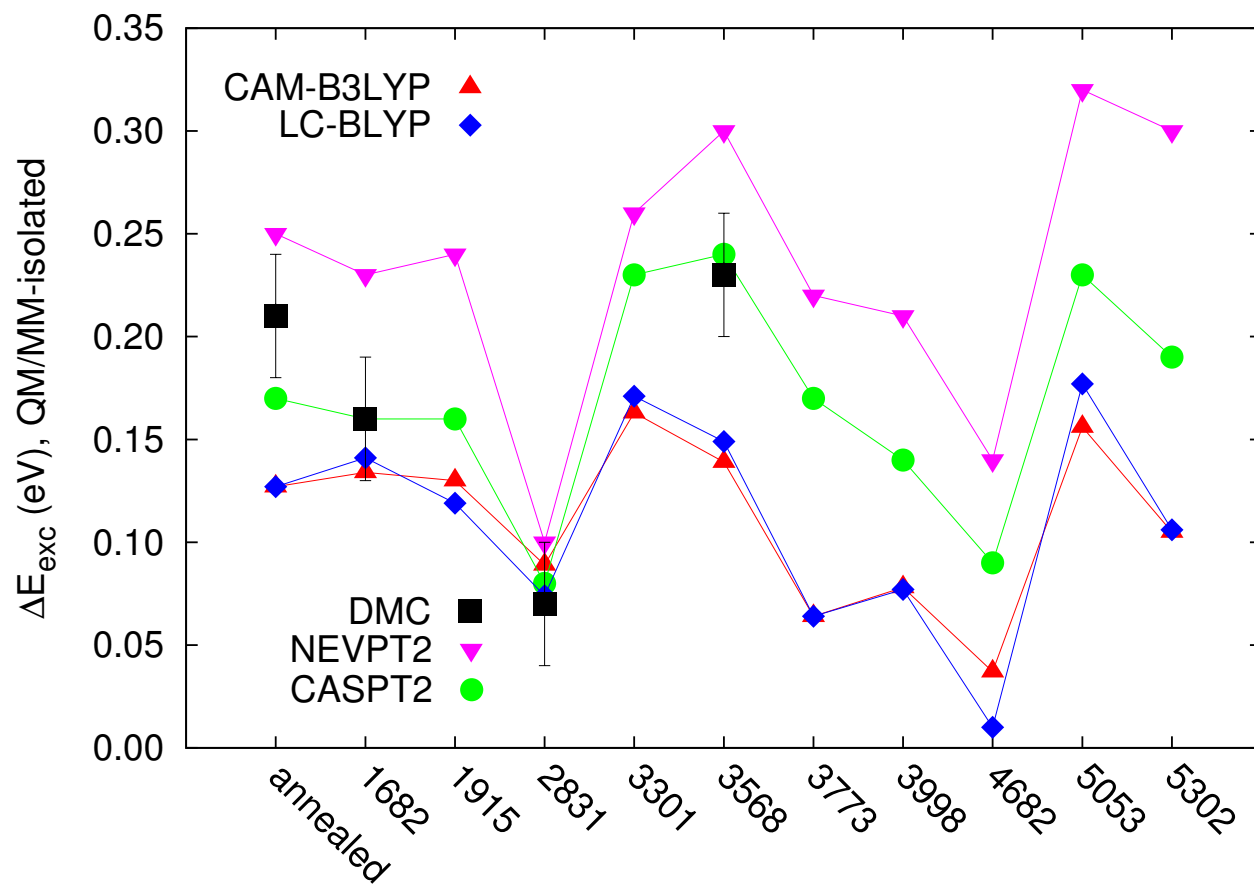


Figure 7: B form: Difference between the QM/MM and isolated excitation energies computed with TDDFT (CAM-B3LYP and LC-BLYP), CASPT2, NEVPT2, and QMC.

addition to the MM charges to describe the protein.

On these 10 frames we apply three different MMpol approaches which represent three different approximations of the environment response to the electronic excitation: (i) In the first one, the polarization of the environment is limited to the ground state (polGS), (ii) in the second the polarization is extended to the excited state using a state-specific (polSS) model where upon excitation the induced dipoles are allowed to relax according to the new density corresponding to the chromophore excited state, and (iii) in the third a linear response (polLR) model is used where polarization is again extended to the excited state but this time the relaxation of the induced dipoles is determined by the transition density. We note that the present polSS formulation corresponds to what is called corrected LR (cLR) method within polarizable continuum approaches.<sup>124</sup>

While the polGS completely lacks of any response of the environment to the excitation, both polSS and polLR account for relaxation effects but in two different ways: polSS is the standard way to introduce polarization effects in WF methods as it explicitly calculates the induced dipoles according to the densities of the ground and excited states while polLR does not require the determination of the excited state density but only that of the ground state and the transition density. This specificity has made polLR the standard one in combination with TDDFT methods where transition densities are directly available. In previous analyses<sup>125,126</sup> of the different physical nature of the responses introduced in the two models in combination with polarizable continuum models, it has been shown that while polSS recovers the "exact" electrostatic response of the environment, polLR introduces an effect that in the old literature of solvatochromism was described as part of the dispersion interactions. In the same analyses it was also shown that a full response should include both terms in addition to a purely dispersion term.

As summarized in Figure 8, we find that for the A form, polarizing the environment around the ground state (polGS) of the chromophore leaves the excitation energy essentially unchanged with respect to the non-polarizable MM approach. On the other hand, for the B form, polGS leads to a blue shift (0.06 eV). That the ground-state polarization has a larger effect on the B form is understandable as the chromophore is charged. Moving to the two descriptions which

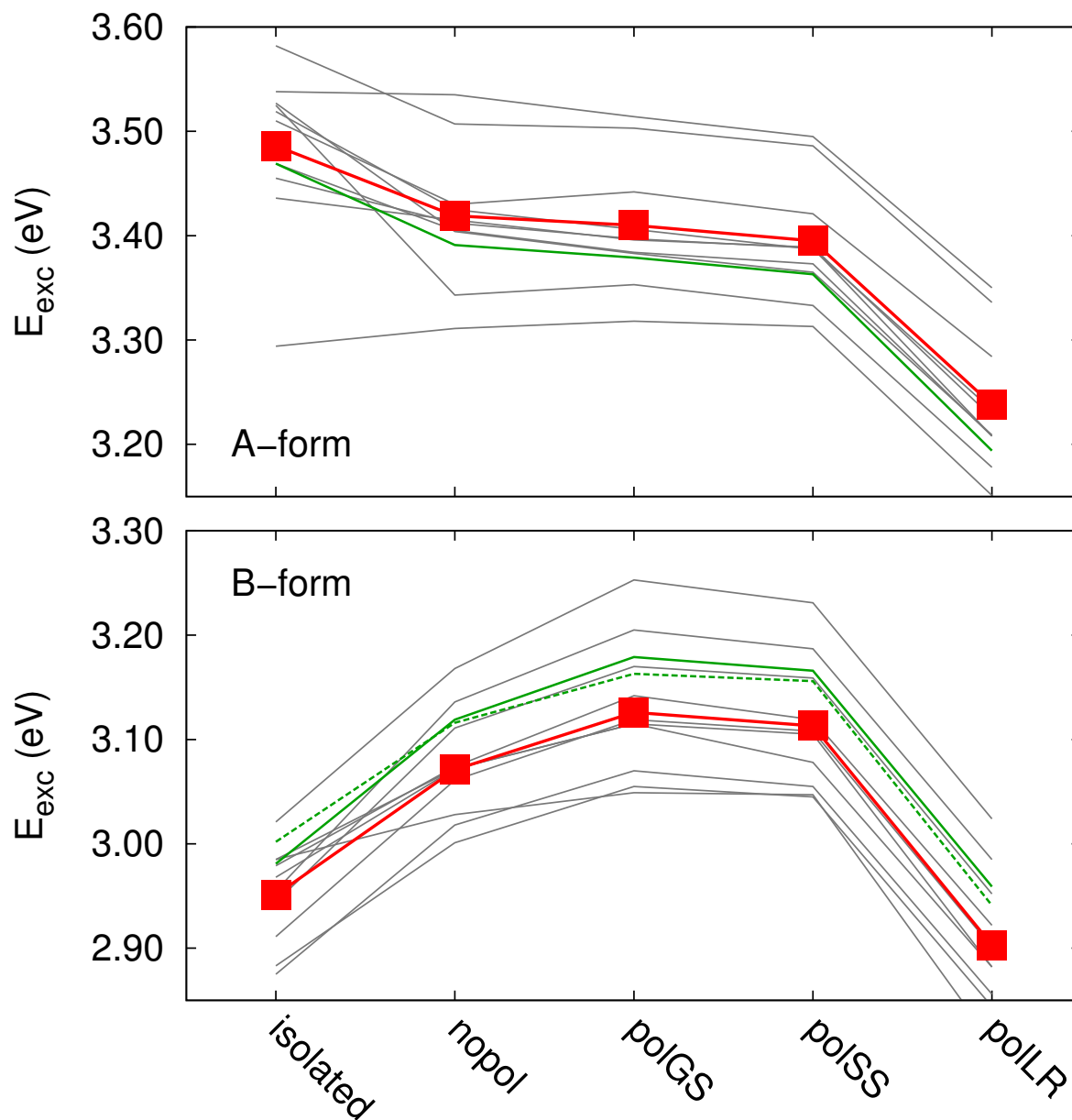


Figure 8: A and B forms: Excitation energies computed with TDDFT on the chromophore only (isolated) or surrounded by the environment described with static point charges (nopol), with point charges and dipoles polarized in the ground state (polGS), in the excited state (polSS) or using a linear-response (polLR) formulation. The thin, gray lines represent the results on a specific frame, while the thick, red line shows average values across the frames. The thin, green lines represent the annealed frame, while the dashed, green line for the B form represents the annealed frame obtained with a larger QM region in the QM/MM dynamics.



introduce an environment response to the excitation, we obtain that, for both forms, the polSS description does not significantly change the excitation energy with respect to a completely frozen solvent (polGS). This seems to indicate that the variation of the electron density upon excitation is not very significant either in the A or the B form (as a matter of fact, the TDDFT/CAM-B3LYP absolute variation of the dipole moment between ground and excited state is around 3 Debye for both molecules). On the contrary, the polLR causes a substantial red shift with respect to polGS for both the A ( $-0.17$  eV) and the B ( $-0.22$  eV) forms. Such a finding can be explained in terms of the large transition dipole which characterizes the excitation in both forms and which can strongly interact with the polarizable environment. Perturbatively to linear order, the shifts in the polSS and polLR excitation energies with respect to polGS can in fact be shown<sup>125</sup> to be proportional to the square of the change in dipole moment between the ground and excited states and to the square of the transition dipole moment, respectively; a relation which is very convincingly demonstrated in Figure 9. From this analysis, we conclude that the chromophore-environment interactions in the ground state are almost equally described by the non-polarizable and the polarizable models. To account for full response to the excitation, we need instead to move beyond a non-polarizable model and include a direct coupling between the transition density of the chromophore and the corresponding polarization of the protein environment.

An alternative method for treating the environment is DFT embedding,<sup>127</sup> which one would expect to be superior to a classical treatment. To generate the embedding potential for the chromophore, we use here either the density of the isolated environment, relaxing the chromophore density,<sup>128</sup> or bring the environmental and chromophore densities at self-consistency after an appropriate number of so-called freeze&thaw cycles.<sup>129</sup> For the A form, we find that DFT embedding with freeze&thaw cycles yields a small blue shift (about 0.02 eV) compared to the MM description, while using the density of the isolated environment leaves the excitation energies essentially unchanged. However, for the B form, DFT embedding with freeze&thaw cycles causes a blue shift with respect to the MM charges as large as 0.15 eV at the level of TDDFT/DFT. (As discussed in section S8 of the SI, the use of a different basis and the CAMY-B3LYP functional within

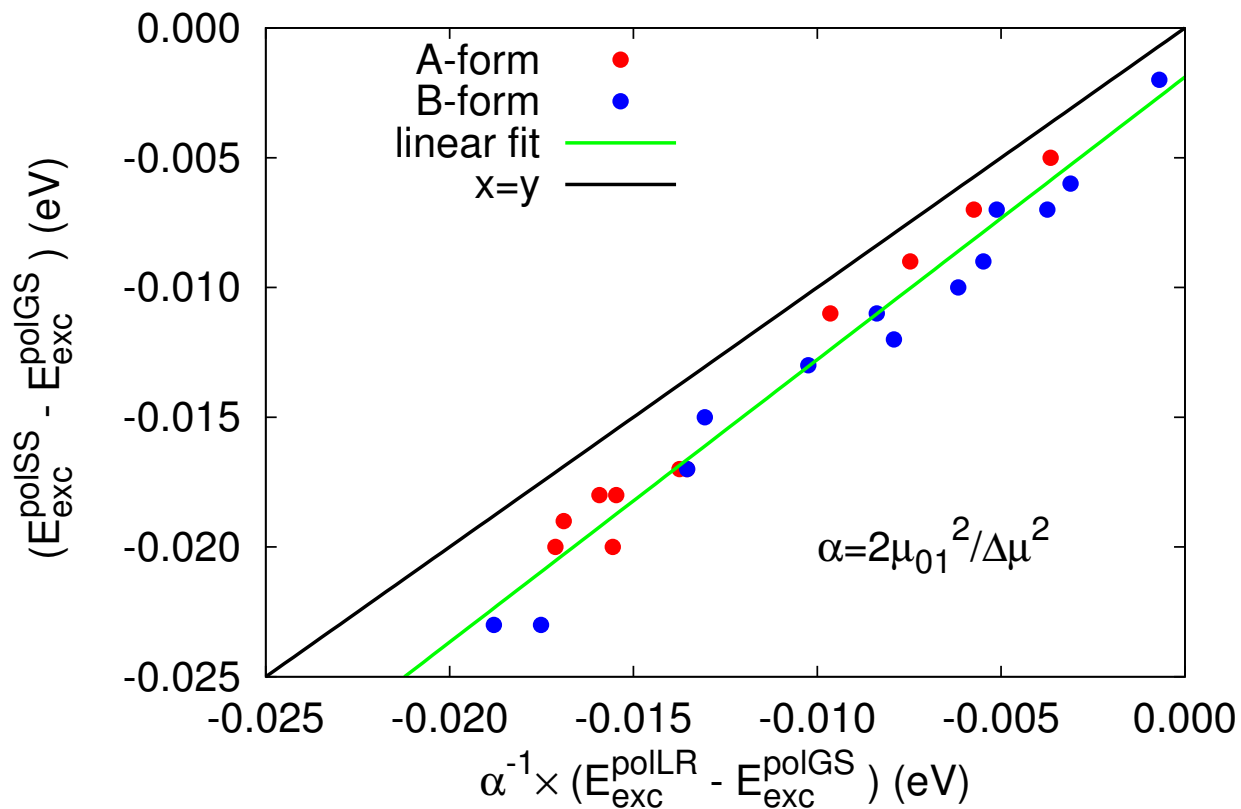


Figure 9: A and B forms: Shifts between the TDDFT/MMpol state-specific (polSS) and ground-state polarized (polGS) excitation energies versus the corresponding linear-response (polLR) results rescaled with twice the ratio squared of the transition dipole moment and the change in dipole moment between the ground and excited states.

TDDFT/DFT accounts for about 0.05 eV of this shift.) Since DFT embedding is conceptually a ground-state embedding method, this finding is reminiscent of the blue shift caused by the polGS model, although the shift is larger for DFT embedding. If one uses the density of the isolated environment, there is almost no shift with respect to the MM values. Therefore, DFT embedding gives at best excitation energies equivalent to the MM ones but only with an isolated environmental density, while one would expect a relaxed environment to provide a better description of the system especially in the anionic B form. Nevertheless, in the following, we present the DFT embedding results computed with the freeze&thaw scheme which is internally consistent, conceptually related to polGS, and not dictated by the performance of the method for this particular system.

To investigate the effect of polarizing the environment at the DFT level, we also apply the recently introduced state-specific formulation of DFT embedding<sup>26,130</sup> and find that the response to state-specific potentials at the TDDFT/DFT level is on average very small for both forms (see Table S7 of the SI). While the use of an excited-state functional in the state-specific scheme would in principle ensure the inclusion of all relevant environmental effects on the excitation, the use of ground-state functionals appears in practice to only capture the electrostatic response of the protein. It is therefore not surprising that the state-specific TDDFT/DFT results are in line with the polSS ones.

While the electrostatic contribution to the environment response is negligible for both forms, induced dipoles in linear response (polLR) appear to offer a possible remedy for the blue shift in the excitation energies caused by the classical point charges. To validate the behavior of polLR in describing the environment effects on the excitations, we compare the polLR results with supermolecular calculations at the same level of theory. Given the size of the protein model and the number of frames, we limit these test calculations to clusters of 279 atoms comprising the chromophore and close-by residues, and compare the TDDFT/MMpol calculations on the same clusters (MMpol<sub>Cl</sub>) with the supermolecular TDDFT results. As shown in Figure 10, we find a remarkably good correlation between the shifts in the polLR and supermolecular excitation energies with respect to the MM description for each frame. In fact, the parameters of a linear fit for the A and B

forms are compatible and imposing a intercept of zero in the fit leads to a slope of about one and not a significantly poorer quality of the fit for both forms. Furthermore, the average excitation energies computed at the polLR level on the cluster differ from the average supermolecular reference by less than 0.01 eV for both forms as detailed in Table 1.

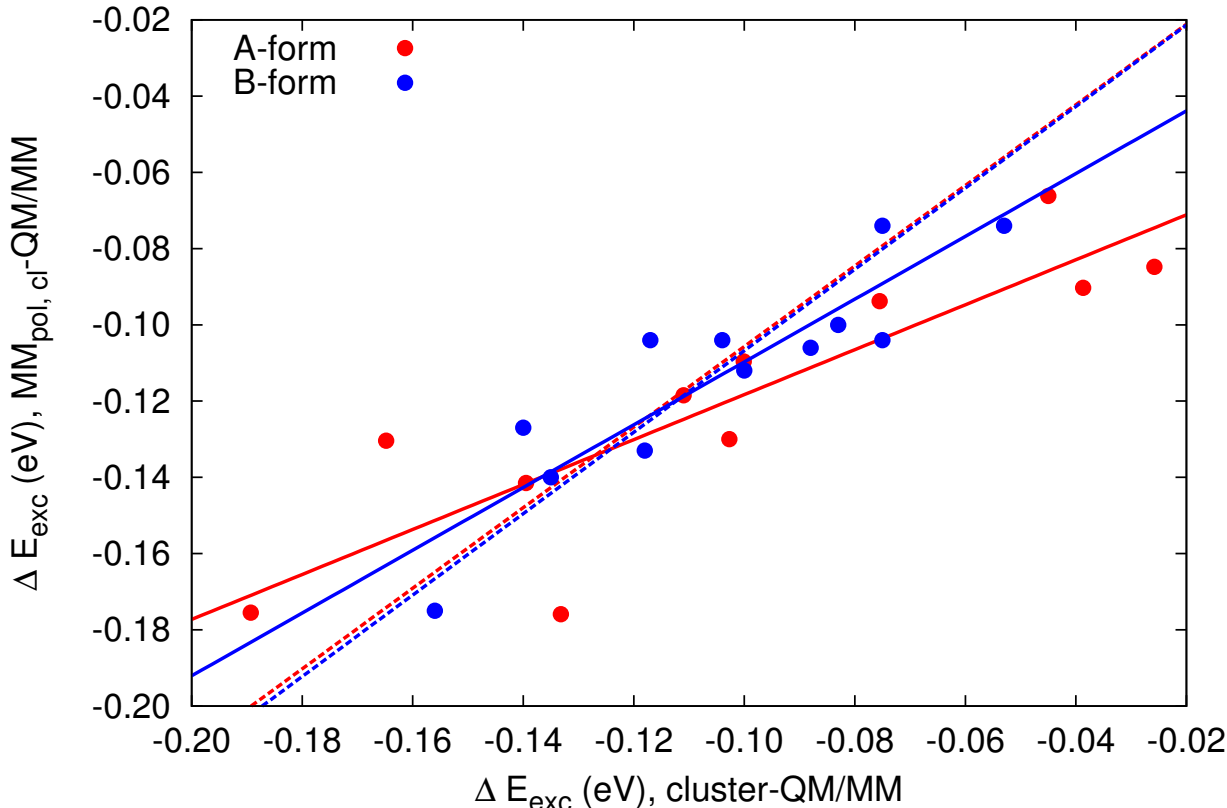


Figure 10: A (red) and B (blue) forms: Correlation between the excitation-energy differences cluster-MM and MMpol<sub>CI</sub>-MM (eV). Full lines: Fit function  $f(x)=a*x+b$  with coefficients  $a=0.59(11)$ ,  $b=-0.059(13)$  for the A form, and  $a=0.82(13)$ ,  $b=-0.027(14)$  for the B form. Dashed lines:  $b=0$  with  $a=1.06(9)$  for the A form and  $1.07(4)$  for the B form.

The recent study of Ref.<sup>13</sup> with the CC2 method reports a somewhat larger discrepancy of about 0.04 eV between linear-response MMpol and supermolecular calculations both done on 160-atom clusters of the A and B forms. This is in line with our results: When considering individual frames in Figure 10, one can find some frames with similar TDDFT/MMpol excitation energies but a spread of up to 0.05 eV in the cluster values, and vice versa. Averaging over many frames seems however to lead to a high degree of cancellation in the MMpol errors.

A summary of the results obtained for all 10 frames by applying the different ways of treating the environment in combination with a TDDFT description of the excitations is reported in Figure 11 and in Table 1, where we also list the excitation energies obtained with the LC-BLYP functional. We find that LC-BLYP responds to the changes in the description of the environment very similarly to CAM-B3LYP and that the agreement between the polLR and cluster excitation energies is equally good (see also Figures S6-S8 of the SI). We finally note in Table 1 and Figure 8 that increasing the QM region in the QM/MM dynamics and annealing has no significant effect on the excitation energies of the B form at variance with the suggestion of Ref.,<sup>10</sup> where only a partial and constrained optimization of the structure was carried out.

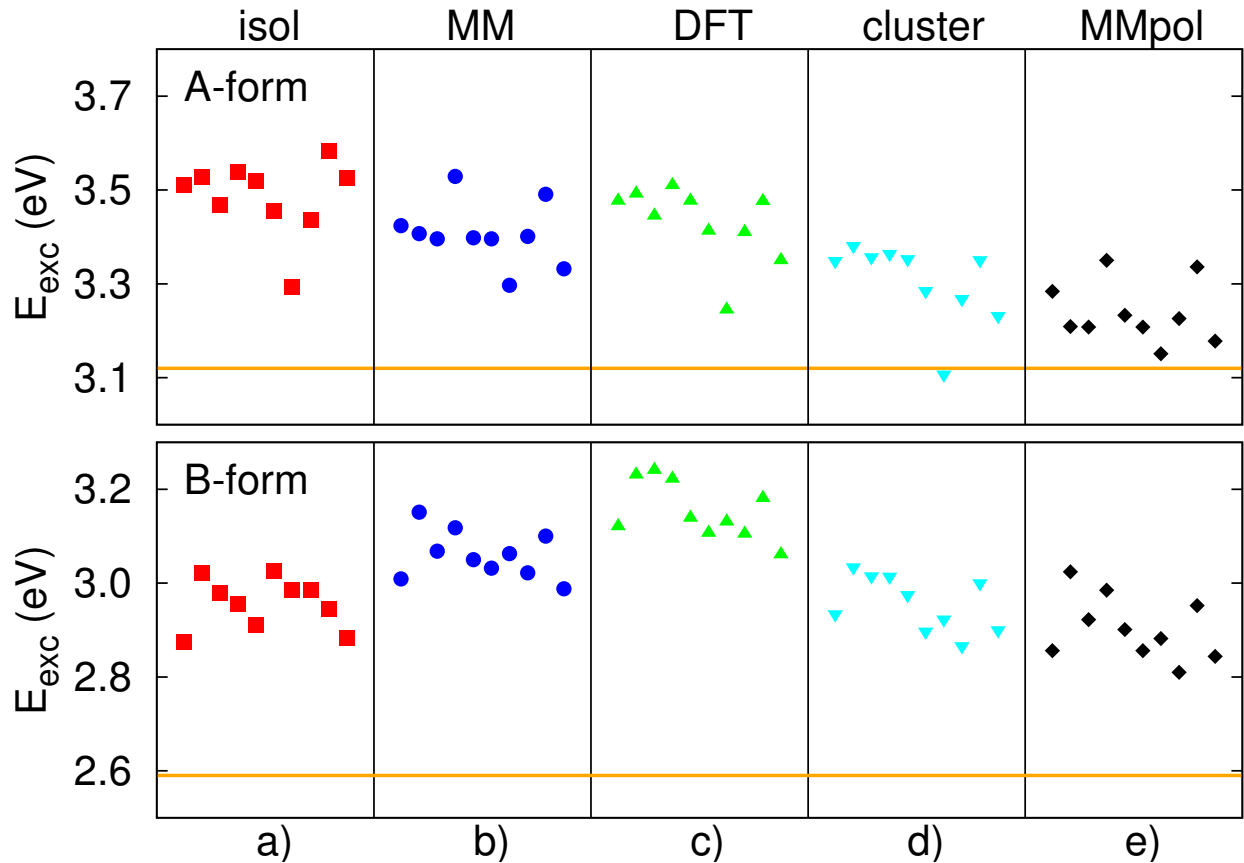


Figure 11: A (top) and B (bottom) forms: Excitation energies computed with TDDFT for a) the isolated QM chromophore, b) the QM chromophore/MM protein, c) DFT environment, d) a large QM cluster, e) QM chromophore/MMpol (polLR) protein. The orange lines indicate the room-temperature absorption maxima of 3.12 and 2.59 eV for the A and B forms, respectively.

As already discussed in the previous section, all WF methods considered here agree with

Table 1: TDDFT excitation energies (eV) of A and B forms averaged over 10 frames (300 K) and for the annealed structure (0 K). The polLR formulation is used in QM/MMpol. Experimental absorption maxima at 295 K and 1.6 K are also given. The error on the averages is in brackets.

CAM-B3LYP					
	A form		B form		
	300 K	0 K	300 K	0 K	0 K <sup>a</sup>
isolated	3.49(3)	3.47	2.96(2)	2.98	3.00
QM/MM	3.41(2)	3.38	3.06(2)	3.11	3.11
QM/MMpol	3.24(2)	3.19	2.90(2)	2.96	2.94
QM/MMpol <sub>cl</sub>	3.30(3)	3.26	2.95(2)	3.01	2.98
QM/DFT <sub>cl</sub>	3.43(3)	3.43	3.15(2)	3.23	3.19
Cluster	3.30(3)	3.28	2.96(2)	3.03	2.99
Exp. <sup>113</sup>	3.12	3.05	2.59	2.63	2.63
LC-BLYP					
	A form		B form		
	300 K	0 K	300 K	0 K	0 K <sup>a</sup>
isolated	3.70(3)	3.67	3.03(2)	3.05	3.07
QM/MM	3.58(3)	3.53	3.13(2)	3.18	3.16
QM/MMpol	3.41(3)	3.35	2.99(3)	3.04	3.00
QM/MMpol <sub>cl</sub>	3.49(4)	3.42	3.03(3)	3.09	3.02
QM/DFT <sub>cl</sub>	—	—	—	—	—
Cluster	3.51(4)	3.48	3.06(3)	3.14	3.06
Exp. <sup>113</sup>	3.12	3.05	2.59	2.63	2.63

<sup>a</sup> Large QM region in the QM/MM annealing.

TDDFT in predicting that the MM description of the protein leads to significantly blue-shifted excitation energies of the B form with respect to the experimental absorption maximum: The average excitation energy over the 10 frames is  $2.79 \pm 0.02$  and  $3.02 \pm 0.02$  eV for CASPT2/MM and NEVPT2/MM, respectively. The QMC/MM excitation energies on a subset of frames are very close to the NEVPT2 values as illustrated in Figure 12 and detailed for a subset of frames in Table 2 (see Table S7 of the SI for the NEVPT2 and CASPT2 results on all frames). Also the anomalous blue shift observed with TDDFT/DFT for the B form with respect to the MM embedding is confirmed at the wave function level, with CASPT2/DFT predicting an even larger blue shift of  $0.25 \pm 0.02$  eV and QMC/DFT a similar correction of  $0.11 \pm 0.04$  eV. The further localization of the excitation energy induced by DFT embedding is possibly due to a too high/steep barrier in the embedding potential induced at the edge of the chromophore by the approximate kinetic functional.<sup>130</sup> Increasing the cluster size worsens the situation, causing further blue shifts at the CASPT2 level as shown in Figure 13. As in the case of TDDFT/DFT, we find that the use of state-specific DFT potentials has on average almost no effect on the excitation energies at the CASPT2 level (less than 0.01 eV as reported in Table S7 in the SI). Additional tests on the use of different basis set or functionals in the construction of the potentials in the DFT embedding scheme are provided in Table S8 of the SI.

Table 2: A and B forms: Excitation energies (eV) computed with MM embedding at the TDDFT, SS/MS CASPT2, NEVPT2, and QMC levels. The statistical error on the QMC values is given in brackets.

frame	TDDFT	SS/MS CASPT2	NEVPT2	QMC
A-form				
ann.	3.38	3.24/3.32	—	3.55(2)
1189	3.42	3.36/3.43	—	3.69(2)
2103	3.40	3.31/3.40	—	3.58(1)
4519	3.33	3.23/3.28	—	3.51(2)
B-form				
ann.	3.11	2.82/2.82	3.06	3.10(2)
1682	3.01	2.72/2.73	2.98	3.03(2)
2831	3.07	2.71/2.72	2.92	2.95(2)
3568	3.05	2.85/2.86	3.03	3.13(2)

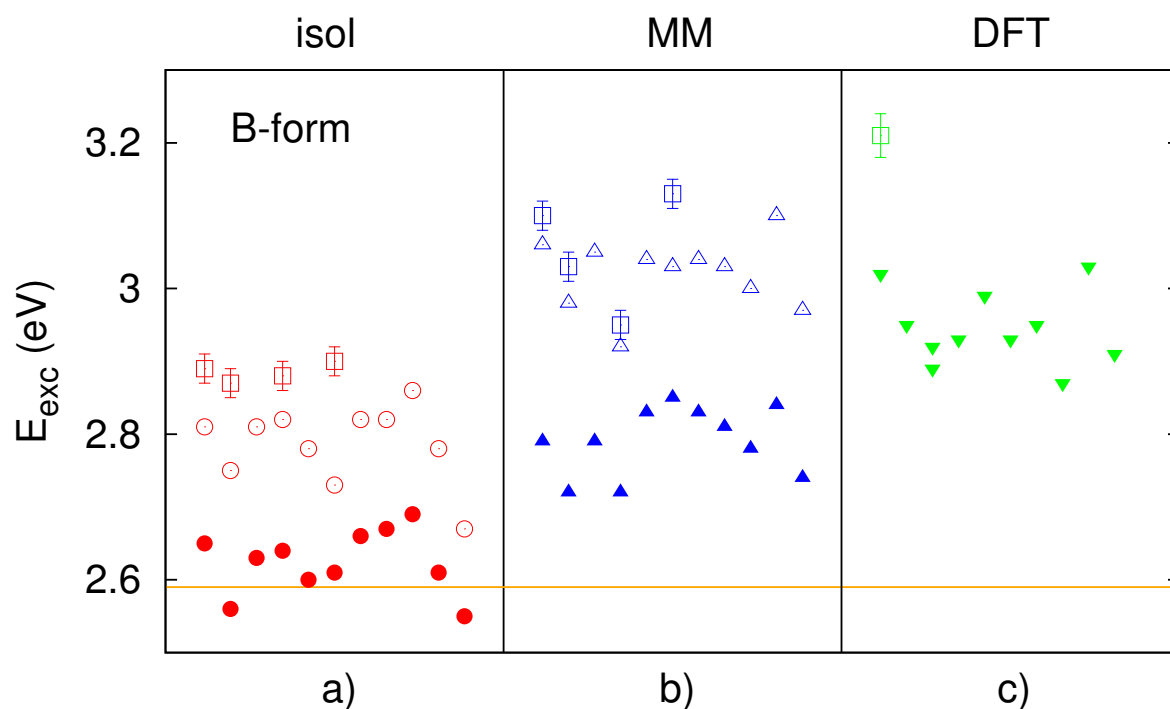


Figure 12: B form: Excitation energies computed with WF methods for a) the isolated QM chromophore, b) the QM chromophore/MM protein, c) DFT environment. Full symbols refer to CASPT2, empty squares to QMC, and other empty symbols to NEVPT2. The leftmost symbols in each panel represent the annealed structure and the others the room-temperature frames, while the orange line indicates the experimental room-temperature absorption maximum of 2.59 eV.

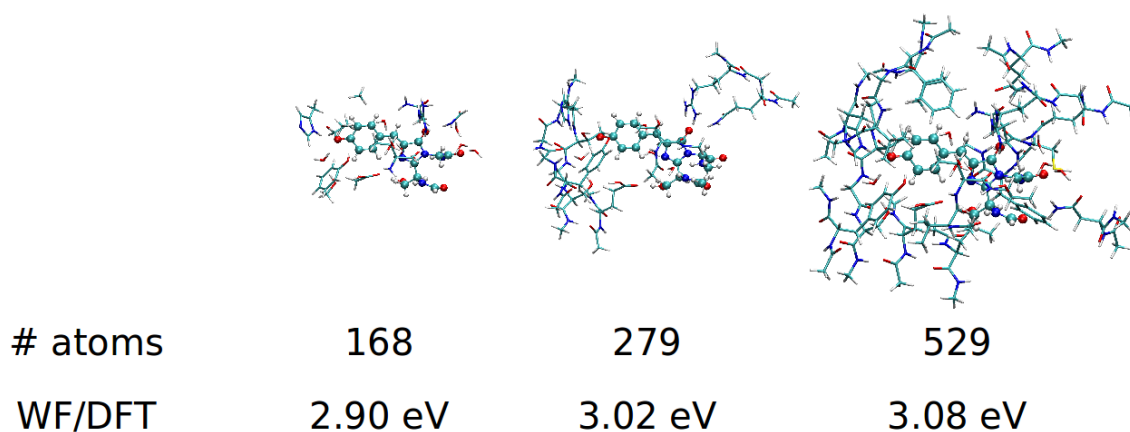


Figure 13: B form: Excitation energies computed with CASPT2/DFT for increasing cluster size for the annealed structure.



All findings with WF methods for the B form are summarized in Figure 12, which clearly illustrates how the trends for the different embedding methods mirror the TDDFT ones of Figure 11, with the difference that the shifts are generally slightly larger for the WF methods. One might therefore hope that the large red-shift between TDDFT/MM and TDDFT/cluster or TDDFT/MMpol would be even larger between WF/MM and WF/MMpol or a possible WF/cluster run. Unfortunately, WF/MMpol in the polLR formulation is not yet available for the chosen WF methods (an estimate of such effects at the CASPT2/MMpol level will be given in the Conclusions), while a calculation on a cluster of almost 300 atoms is out of reach for all but the simplest WF approaches due to prohibitive costs. For the A form, we encounter severe difficulties in the WF calculations due to the presence of multiple minima in the CASSCF reference calculations and a strong sensitivity of the CASPT2 excitation energies on the size of the active space, the minimal but yet large expansion not being sufficient. Therefore, in Table 2, we only present CASPT2/MM and QMC/MM results for four frames (the annealed and three room-temperature ones) where the CASPT2 calculations are robust at both the single- and multi-state level as discussed in detail in Section S4 of the SI. The available WF excitation energies are in line with the TDDFT results also for the A form with QMC/MM being a bit higher and CASPT2/MM a bit lower than the corresponding TDDFT/MM excitation energies on the same frames, and both being blue-shifted with respect to the experimental absorption maximum of 3.12 eV also of this form.

If we focus on the polarizable (MMpol) embedding in the polLR formulation which all tests indicate as the most reliable for this system, we can only attempt a comparison with experiments at the TDDFT level. Even though a direct comparison of the CAM-B3LYP results to experimental absorption maxima is tenuous as previously noted, we can use the data for each form to validate our account of thermal effects. Since we have performed MMpol calculations on both room-temperature and annealed frames as listed in Table 1, comparing shifts between the excitation energies of the two data sets to experimental shifts in the absorption maxima is possible. We find that for the A form, the annealed (frozen) frame has a lower excitation energy than the average of the room-temperature frames by  $0.05 \pm 0.02$  eV, while for the B form, the annealed frame has

a higher excitation energy than the room-temperature average ( $0.04 \pm 0.02$  eV if we focus on the annealing performed with the same QM region). These two shifts are rather close to the experimental values<sup>113</sup> of 0.07 eV red-shift between room- and low-temperature absorption maxima for the A form and 0.04 eV blue-shift for the B form. The precise values of the shifts are difficult to ascertain, as especially the A form has a wide peak, but reassuringly, the direction and order of magnitude is well reproduced by combining an MMpol embedding with our thermal sampling in the QM/MM MD simulations.

### 3.4 On the use of a cluster representation of the protein

In the previous subsection, we employed cluster calculations to validate the MMpol approach and to perform the DFT embedding calculations. Quite often in the literature, a cluster representation of the protein is instead used to compute the excitation energies for a direct comparison with experiments and the question therefore arises on how large a cluster must be for a converged calculation of the excitation energies. Distressingly, a recent article<sup>8</sup> suggested that clusters as large as 723 atoms are necessary for reliable results on the photoactive yellow protein.

In Table 3, we answer this question for wild-type GFP by enhancing the 279-atom cluster calculations in several ways, namely, by increasing the cluster size to 345 atoms, including the rest of the protein outside the cluster as point charges, or comparing the TDDFT/MMpol results for the cluster only and for the full protein. We present also an estimate based on CASPT2/DFT calculations on two different cluster sizes for the B form but consider the observed unphysical increase in the blue shift an artifact of the use of approximate kinetic functionals. In the case of TDDFT-based estimates, the error introduced by the use of a 279-atom cluster is small and comparable for both protonation states, between 0 and 0.07 eV for the A form and between  $-0.03$  and 0.06 eV for the B form. Therefore, it appears that a cluster of about 300 atoms is sufficient to estimate the excitation energies and, especially, the trends between the two protonation forms of wild-type GFP. On the other hand, smaller clusters as for instance employed in previous coupled-cluster studies of GFP<sup>10</sup> probably suffer from much larger cluster errors and carry therefore a large

uncertainty as regards comparison with experiments. For example, adding the rest of the protein in Ref.<sup>10</sup> as standard classical force-field point charges yielded a 0.2 eV blue shift, which is about an order of magnitude larger than the errors we obtain here due to the addition of the MM protein (0.01 eV for the A and 0.03 eV for the B form).

Table 3: Cluster convergence analysis for the A and B forms. Results on the chromophore (CRO) plus 12 residues (279-atom cluster) are compared with calculations on a larger cluster (345 and 529 atoms) or the whole protein treated at different levels of theory.

T (K)	model			E <sub>exc</sub> (eV)		
	279-atom cluster			cluster	cluster+rest	error
	CRO	12 res. <sup>a</sup>	rest			
A form						
0	TDDFT		TDDFT <sup>b</sup>	3.28	3.25	0.03
0	TDDFT		MM	3.28	3.28	0.00
0	TDDFT	MMpol	MMpol	3.26	3.19	0.07
300	TDDFT	MMpol	MMpol	3.30(3)	3.24(2)	0.06(2)
B form						
0	TDDFT		TDDFT <sup>b</sup>	3.02	3.02	0.01
0	TDDFT		MM	3.02	3.05	−0.03
0	TDDFT	MMpol	MMpol	3.01	2.96	0.05
0 <sup>c</sup>	TDDFT	MMpol	MMpol	2.98	2.94	0.03
300	TDDFT	MMpol	MMpol	2.95(2)	2.90(2)	0.05(1)
0	CASPT2	DFT	DFT <sup>d</sup>	3.02	3.08	−0.06

<sup>a</sup> 8 amino acids and 4 water molecules.

<sup>b</sup> Enlarging cluster to 345 atoms (+2 amino acids and 3 waters).

<sup>c</sup> Large QM region in the QM/MM annealing.

<sup>d</sup> Enlarging cluster to 529 atoms (+11 amino acids and 3 waters).

### 3.5 Alternative H-bond network for the B form

Our results strongly indicate that a poor description of the environment via classical point charges is mainly responsible for the blue shift observed in the excitation energies with respect to experiments. Nevertheless, we will here also explore the possibility that a drastically different configuration of the chromophore might lead to the desired agreement with experiments. In particular, we will investigate the impact on the excitation energies of the formation of an internal hydrogen bond between O<sub>γ</sub> in the Ser65 side chain and N11 of the imidazolinone ring. The extra hydrogen

bond on the imidazolinone side of the chromophore is expected to stabilize the excited state as the excitation has a partial charge transfer character from the phenolic to the imidazolinone ring, and therefore red shift the excitation energy. (We note that the an analogous internal hydrogen bond network between the Thr65 side chain and the chromophore is observed in the X-ray structure 1Q4B<sup>107</sup> of the S65T mutant.) The distortion of Ser65 however significantly perturbs the surroundings of the chromophore since it is also coupled to a twist of Glu222 from the *anti* to the *syn* conformation and, consequently, to a breaking of the proton-transfer wire, so the combined effect on the excitation energy is in fact hard to predict.

Such a structure was recently put forward as the “true” conformation of the B form of wild-type GFP while the standard hydrogen-bond network we have adopted above was assigned to the I form of GFP.<sup>40</sup> This assignment was motivated by the the small blue shift of 0.02–0.04 (depending on the quantum method) computed between the excitation energies of their B and I forms and the fact that, experimentally, the 0→0 transition of the B form is also blue shifted (albeit by 0.1 eV) with respect to the I form.

To prepare this alternative B form, we appropriately twist Ser65 and Glu222, but keep the S65T-like orientation of Thr203 bonded to the phenolic oxygen for consistency with our standard B form and with the bulk of the literature (Thr203 is twisted away from the phenolic oxygen in Ref.<sup>40</sup>). We then optimize a small cluster within DFT with fixed surrounding residues and, starting from this structure, perform a QM/MM simulation of 15 ps followed by annealing. During the 15 ps run, the hydrogen-bond network around the phenolic oxygen behaves similarly to what observed for the standard setup. In particular, the water molecule and Thr203 are constantly bonded to the phenolic oxygen but His148 and Tyr145 intermittently bond and detach. The final structure resembles the one of Ref.<sup>40</sup> but displays a hydrogen bond between His148 and the phenolic oxygen of the chromophore instead of having the water molecule bridging the histidine to the chromophore (see Figure S5 of the SI).

Finally, we compute the excitation energies on this structure at the level of CASPT2/MM, TDDFT/MM, and TDDFT on a cluster model. Disappointingly and in agreement with Ref.,<sup>40</sup>

we obtain a small blue shift of 0.04-0.08 in the excitation energies with respect to the annealed structure of our standard B form (see Table S5 of the SI). Our results therefore suggest that the conformational changes of Ser65 and Glu222 do indeed have some effect on the excitation energy but in the opposite direction than what a simple analysis would have suggested. We stress however that we do not agree with Ref.<sup>40</sup> in assigning these structural changes to the difference between the I and B forms based on this small difference in excitation energies. The only available experimental data for the I form<sup>113</sup> is the 0 $\rightarrow$ 0 transition at room temperature and not the absorption maximum at low temperature, we should compare our vertical excitation energy for the annealed structure. Furthermore, we have shown above that temperature effects on the excitation energies are of this same order of magnitude. Finally, temperature effects were neglected altogether in Ref.,<sup>40</sup> so their relative absolute energies for the A, B, and I forms should be compared to low-temperature populations. Consequently, since they estimated the A form to be 1 kcal/mol below the I and B forms, this would lead to the A form being exclusively populated at low temperature in contradiction with experiments at 1.6 K,<sup>113</sup> which indicate that the B form is energetically favorable to the A form and the I form entirely unpopulated.

## 4 Discussion and conclusions

In this paper, we have thoroughly investigated the multi-scale computation of the excitation energies of wild-type GFP, examining the causes of the spectral spread and identifying the most suitable protocol for an accurate and affordable calculation of the absorption properties of this prototypical fluorescent protein.

Through extensive QM/MM molecular dynamics simulations, we explored the possible conformations of the protein in both the neutral A and anionic B forms, and found that the A form displays a very stable hydrogen-bond network, while the B form exhibits more significant deviations from the average structures even in the 20 ps time interval we examined. Analyzing 50 equidistant frames for each protonation form at the TDDFT/MM level, we found remarkably large

spreads in the excitation energies of up to 0.5 eV in both cases and, for the B form, confirmed the correctness of the TDDFT excitation-structure relation through CASPT2/MM calculations on the same 50 frames. Surprisingly, the structurally more stable A form has a larger standard deviation in the energies, which we find to be more sensitive to the internal coordinates of the chromophore, while the environment plays a bigger influence on the excitations of the B form. Including several frames, as done here, ensures a faithful sampling of the possible conformations of the chromophore and its surroundings.

A first screening of the excitation energies of these numerous frames with the cheap MM static charge embedding yielded excitation energies blue shifted with respect to experiments. Depending on the quantum method, the excitation energies are in the range of 3.2-3.6 for the A form and 2.8-3.1 eV for the B form compared to an experimental absorption maximum of 3.1 and 2.6 eV, respectively. For both forms, CASPT2 gives the lowest excitation energies, QMC the highest ones, and CAM-B3LYP lies between the two, while NEVPT2, only employed on the B form, agrees reasonably well with QMC. Our findings are in line with other studies in literature on wild-type GFP<sup>7,9,10,13</sup> and other photosensitive bio-systems<sup>8,11,12,123</sup> which found blue-shifted excitation energies compared to experiments using different quantum methods in combination with an MM environment. Including temperature effects, as done here and in previous work,<sup>11,12,19</sup> fails to improve on experimental agreement.

While there is clearly a spread in the values stemming from the choice of excited-state quantum method, the coarseness of the MM description is responsible for a systematic blue shift. To demonstrate this, we used 10 frames from cluster analysis for each protonation form and compared the MM results at the TDDFT level to calculations done on clusters containing almost 300 atoms, which we showed to be sufficiently large to account for most of the tuning by the protein environment. Within TDDFT, which is the only method allowing the treatment of such large clusters, the MM description causes a blue shift of about 0.03–0.19 eV when compared to the reference cluster calculations of both the A and the B forms. Even though the use of large clusters does not greatly ameliorate the agreement of TDDFT with experiments, the measure of success/failure of a

particular description of the environment is its ability/inability to reproduce the excitation energies on large clusters treated at the same level of theory. Given the evident inadequacy of static point charges, we explored two alternative approaches to improve the description of the environment. Remaining within classical methods, we investigated the inclusion of induced dipoles (MMpol) and performed TDDFT/MMpol in linear response (polLR) on ten frames for each protonation form. The cluster TDDFT excitation energies and the MMpol calculations on exactly the same cluster agree remarkably well for both the A and B forms.

We note that, when comparing with previous studies with a similar MMpol methodology,<sup>19,21</sup> we find a good agreement for the B form but not for the A form, which displays a higher sensitivity in the excitation to the internal coordinates of the chromophore. As discussed in Section S9 of the SI, if one accounts for the use of a different DFT functional in optimizing the quantum chromophore and the differences in the structural relaxation, the discrepancy reduces to about 0.1 eV also for the A form. Importantly for our discussion, the shifts with respect to the MM values induced by their multipoles obtained in a LoProp construction<sup>131</sup> are in very good agreement with our findings. This shows that the more cost-effective choice of using semi-empirical force-field parameters and neglecting the static quadrupoles included in Refs.<sup>19,21</sup> is adequate for the description of these systems as also apparent from the agreement of our TDDFT/MMpol results with the cluster calculations.

The good performance shown by the TDDFT/MMpol approach allows us to better understand the role of the environment on determining the excitation energies of the two forms of the GFP chromophore. By comparing three different formulations of the embedding model, namely, the ground-state (polGS), state-specific (polSS) and linear-response (polLR) flavors, it is apparent that, in GFP, the response of the environment to the excitation cannot be represented in terms of purely electrostatic effects induced by the change in the electronic density upon excitation: In fact, the polSS result which accounts for the electrostatic response of the environment to the excitation is almost equivalent to that obtained by assuming that the polarization of the environment is frozen in the configuration corresponding to the chromophore in its ground state (polGS). On the contrary,

if we switch on the polLR interaction, a quite different picture is obtained as regards the nature of the chromophore-protein coupling. This interaction comes from the dynamical response of the environment (here represented by the induced dipoles) to the transition density of the chromophore and it can be described as an environment polarization oscillating at the frequency of the chromophore excitation. Such a "resonance" term, which has been classified as a part of dispersion, leads to a significant red-shift with respect to polGS (or the non-polarizable MM scheme) and a good agreement with the reference supermolecular calculations is finally recovered.

In view of these findings, it is perhaps not so surprising that the other route we followed to improve over static MM charges, namely, sub-system DFT embedding was not successful: This scheme turned out to be at best equivalent to the non-polarizable MM description and, in the case of the anionic B form, to produce further unphysical blue shifts which persist across several quantum methods (CASPT2, QMC, and TDDFT) if the environment is relaxed through freeze&thaw cycles. The recently introduced state-specific formulation of DFT embedding<sup>26,130</sup> did not capture any response of the environment, and increasing the quantum DFT environment brought further blue shifts. While displaying a somewhat worse performance than the polGS and polSS formulations of MMpol probably due to the quality of the approximate kinetic functional, the behavior of DFT embedding confirms that the state-specific formulation of sub-system DFT can recover electrostatic effects but that these are not dominant in this system. Similarly, one expects that other alternative DFT-based embedding techniques that eliminate the need for kinetic-energy functionals either by reconstructing the embedding potentials<sup>132</sup> or imposing orthogonality of the orbitals of the sub-systems<sup>133</sup> will not possess the necessary ingredients to describe the excited states of GFP: They will improve the ground-state description of the environment and, even if they included state-specificity, will not easily capture the coupling with the environment beyond electrostatic effects.

While the MMpol embedding scheme has allowed us to identify the necessary ingredients in the description of the excited states of GFP, we have been able to include either the state-specific (electrostatic) or the linear-response (resonance) polarization of the environment but not both in a united formulation without resorting to a complete "brute-force" quantum calculation on a large



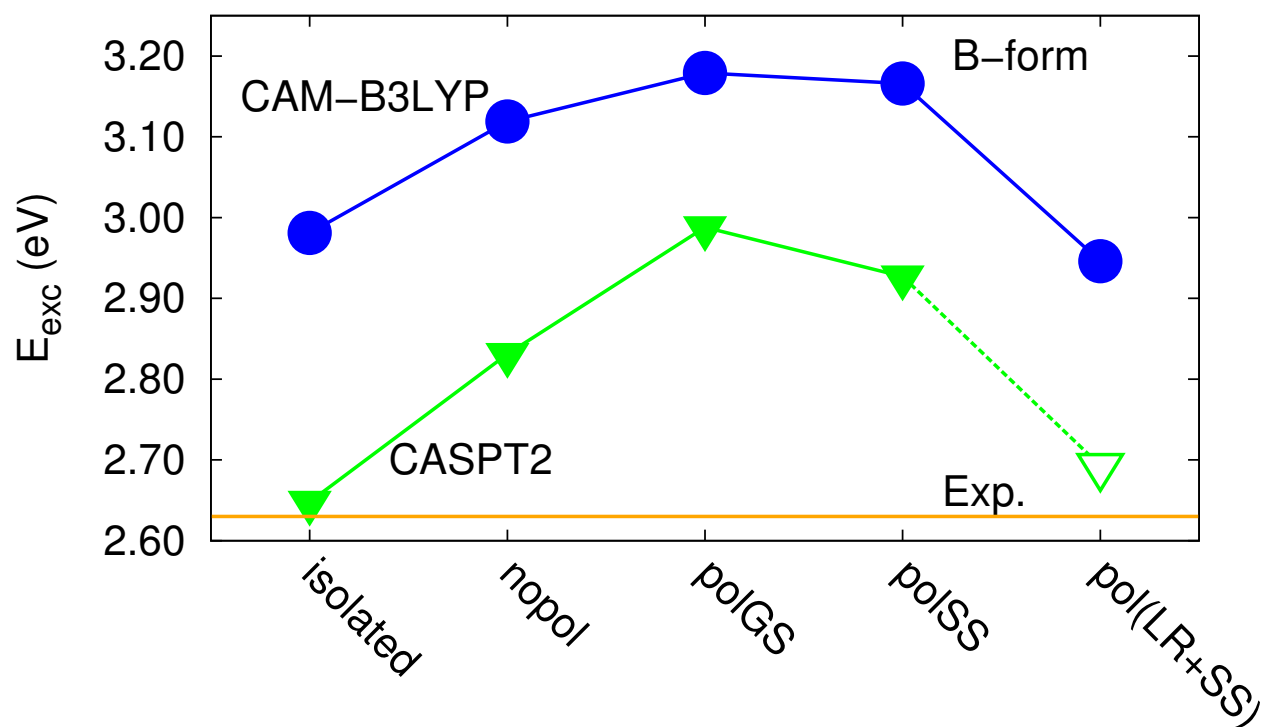


Figure 14: B form: Excitation energies computed with CASPT2 on the chromophore only (isolated) or surrounded by the environment described with static point charges (nopol), with point charges and dipoles polarized within CASSCF/MMpol in the ground state (polGS), in the excited state (polSS) or using an estimate based on the linear-response (polLR) TDDFT result (see text). The calculations are performed on the annealed frame.

quantum cluster. Furthermore, we have done so only at the TDDFT level. Since both the SS and the LR responses depend on the quantum method used to describe the excitation, what can we achieve with a correlated method considering that a large quantum cluster is out of reach and an LR formulation often not available? In an effort towards a more complete and accurate simulation of the excitation process in GFP, we attempt here to estimate the best possible excitation values using a CASPT2 description of the excitation process instead of TDDFT. At the CASPT2 level, one can evaluate the excitation energy in a polSS scheme using the state-specific induced dipoles obtained within CASSCF/MMpol (see Computational Details). Then, we can estimate the polLR correction as the TDDFT/MMpol (polLR) shift with respect to the corresponding polGS value and rescaling it with the ratio squared of the CASPT2 and TDDFT transition dipole moments.

The final picture emerging from this approximate CASPT2 estimate of the effects of the environment is illustrated for the B form in Figure 14 where the different CASPT2 excitation energies computed with static point charges, ground-state and state-specific induced dipoles are compared to the extrapolated polLR+polSS value. As expected from the results in the previous Section, the response at the correlated level is stronger than in TDDFT when the MM description is added and then enhanced with induced dipoles in a polGS and, subsequently, polSS formulation. The estimated polLR correction is instead rather similar to the TDDFT shift since the ratio of the CASPT2 and TDDFT transition dipole moments is in this case very close to one. That the extrapolated CASPT2 energy correlates very well with experiments is an appealing result, which should however not be overstated. One should apply a similar analysis to the other correlated methods used in this study, which further blue-shift the excitation energy with respect to CASPT2 but also respond more strongly to changes in the description of the environment (a perfect agreement between vertical excitation energies and absorption maximum is anyhow a misguided expectation also for the fluorescent B form). However, our extrapolation clearly shows that environmental effects on the excitation processes in GFP are the result of different interactions in the ground and the excited state; only by adopting a polarizable approach which can account for both state-specific relaxation and “resonance” coupling we can get a semiquantitatively correct description.

## Acknowledgement

We thank Johannes Neugebauer for useful discussions. C.D. is supported by an ECHO grant (712.011.005) and we received support from NWO for the use of the SARA supercomputer facilities, and from the COST Action CODECS. C.C. acknowledges support from the Ministerio de Economía y Competitividad (MINECO) of Spain (grants CTQ2012-36195 and RYC2011-08918), the Agència de Gestió d’Ajuts Universitaris i de Recerca from the Generalitat de Catalunya (GENCAT) (SGR2014-1189), and computational resources provided by the Consorci de Serveis Universitaris de Catalunya.

## Supporting Information Available

Additional computational details. Dependence of the excitation energies on the choice of basis sets, geometry, and other relevant parameters in the different methods (e.g. active space or DFT functional). Excitation energies computed with VMC, DMC, and with state-specific CASPT2/DFT and TDDFT/DFT. Comparison with previous TDDFT/MMpol calculations in the literature. This information is available free of charge via the Internet at <http://pubs.acs.org/>.

## References

- (1) Senn, H. M.; Thiel, W. *Ang. Chem. Int. Ed.* **2009**, *48*, 1198–1229.
- (2) Neugebauer, J. *Phys. Rep.* **2010**, *489*, 1–87.
- (3) Mennucci, B. *Phys. Chem. Chem. Phys.* **2013**, *15*, 6583–6594.
- (4) Karplus, M. *Ang. Chem. Int. Ed.* **2014**, *53*, 9992–10005.
- (5) Levitt, M. *Ang. Chem. Int. Ed.* **2014**, *53*, 10006–10018.
- (6) Warshel, A. *Ang. Chem. Int. Ed.* **2014**, *53*, 10020–10031.
- (7) Send, R.; Kaila, V. R. I.; Sundholm, D. *J. Chem. Phys.* **2011**, *134*, 214114.

- (8) Isborn, C. M.; Götz, A. W.; Clark, M. A.; Walker, R. C.; Martinez, T. J. *J. Chem. Theory Comput.* **2012**, *8*, 5092–5106.
- (9) Filippi, C.; Buda, F.; Guidoni, L.; Sinicropi, A. *J. Chem. Theory Comput.* **2012**, *8*, 112–124.
- (10) Kaila, V. R. I.; Send, R.; Sundholm, D. *Phys. Chem. Chem. Phys.* **2013**, *15*, 4491–4495.
- (11) Valsson, O.; Campomanes, P.; Tavernelli, I.; Rothlisberger, U.; Filippi, C. *J. Chem. Theory Comput.* **2013**, *9*, 2441–2454.
- (12) Amat, P.; Nifosi, R. *J. Chem. Theory Comput.* **2013**, *9*, 497–508.
- (13) Schwabe, T.; Beerepoot, M. T.; Olsen, J. M.; Kongsted, J. *Phys. Chem. Chem. Phys.* **2015**, *17*, 2582–2588.
- (14) Thomsson, M. A. *J. Phys. Chem.* **1996**, *100*, 14492–14507.
- (15) Curutchet, C.; Muñoz-Losa, A.; Monti, S.; Kongsted, J.; Scholes, G. D.; Mennucci, B. *J. Chem. Theory Comput.* **2009**, *5*, 1838–1848.
- (16) Olsen, J. M.; Aidas, K.; Kongsted, J. *J. Chem. Theory Comput.* **2010**, *6*, 3721–3734.
- (17) Slipchenko, L. V. *J. Phys. Chem. A* **2010**, *114*, 8824–8830.
- (18) Steindal, A. H.; Olsen, J. M.; Ruud, K.; Frediani, L.; Kongsted, J. *Phys. Chem. Chem. Phys.* **2012**, *14*, 5440–5451.
- (19) Beerepoot, M. T.; Steindal, A. H.; Kongsted, J.; Brandsdal, B. O.; Frediani, L.; Ruud, K.; Olsen, J. M. *Phys. Chem. Chem. Phys.* **2013**, *15*, 4735–4743.
- (20) Beerepoot, M. T.; Steindal, A. H.; Ruud, K.; Olsen, J. M.; Kongsted, J. *Comp. Theor. Chem.* **2014**, *1040-1041*, 304–311.
- (21) Pikulska, A.; Steindal, A. H.; Beerepoot, M. T. P.; Pecul, M. *J. Phys. Chem. B* **2015**, *119*, 3377–3386.

- (22) Sneskov, K.; Schwabe, T.; Christiansen, O.; Kongsted, J. *Phys. Chem. Chem. Phys.* **2011**, *13*, 18551–18560.
- (23) Svensson, M.; Humbel, S.; Froese, R. D. J.; Matsubara, T.; Sieber, S.; Morokuma, K. *J. Phys. Chem.* **1996**, *100*, 19357–19363.
- (24) Pereira Gomes, A. S.; Jacob, C. R.; Visscher, L. *Phys. Chem. Chem. Phys.* **2008**, *10*, 5353–5362.
- (25) Gomes, A. S. P.; Jacob, C. R. *Annu. Rep. Prog. Chem., Sect. C* **2012**, *108*, 222–277.
- (26) Daday, C.; König, C.; Valsson, O.; Neugebauer, J.; Filippi, C. *J. Chem. Theory Comput.* **2013**, *9*, 2355–2367.
- (27) Send, R.; Suomivuori, C.-M.; Kaila, V. R. I.; Sundholm, D. *J. Phys. Chem. B* **2015**, *119*, 2933–2945.
- (28) Day, R. N.; Davidson, M. W. *Chem. Soc. Rev.* **2009**, *38*, 2887–2921.
- (29) Newman, R. H.; Fosbrink, M. D.; Zhang, J. *Chem. Rev.* **2011**, *111*, 3614–3666.
- (30) Hell, S. W. *Science* **2007**, *316*, 1153–1158.
- (31) Patterson, G. H. *Semin. Cell Dev. Biol.* **2009**, *20*, 886–893.
- (32) Hasegawa, J.-Y.; Fujimoto, K.; Swerts, B.; Miyahara, T.; Nakatsuji, H. *J. Comput. Chem.* **2007**, *28*, 2443–2452.
- (33) Marques, M. A. L.; López, X.; Varsano, D.; Castro, A.; Rubio, A. *Phys. Rev. Lett.* **2003**, *90*, 258101.
- (34) Laino, T.; Nifosi, R.; Tozzini, V. *Chem. Phys.* **2004**, *298*, 17–28.
- (35) Sinicropi, A.; Andruniow, T.; Ferré, N.; Basosi, R.; Olivucci, M. *J. Am. Chem. Soc.* **2005**, *127*, 11534–11535.

- (36) Bravaya, K. B.; Khrenova, M. G.; Grigorenko, B. L.; Nemukhin, A. V.; Krylov, A. I. *J. Phys. Chem. B* **2011**, *115*, 8296–8303.
- (37) Bravaya, K. B.; Grigorenko, B. L.; Nemukhin, A. V.; Krylov, A. I. *Acc. Chem. Res.* **2011**, *45*, 265–275.
- (38) Grigorenko, B. L.; Nemukhin, A. V.; Morozov, D. I.; Polyakov, I.; Bravaya, K. B.; Krylov, A. I. *J. Chem. Theory Comput.* **2012**, *8*, 1912–1920.
- (39) Petrone, A.; Caruso, P.; Tenuta, S.; Rega, N. *Phys. Chem. Chem. Phys.* **2013**, *15*, 20536–20544.
- (40) Grigorenko, B. L.; Nemukhin, A. V.; Polyakov, I.; Morozov, D. I.; Krylov, A. I. *J. Am. Chem. Soc.* **2013**, *135*, 11541–11549.
- (41) Laio, A.; VandeVondele, J.; Röthlisberger, U. *J. Chem. Phys.* **2002**, *116*, 6941–6947.
- (42) Laio, A.; VandeVondele, J.; Röthlisberger, U. *J. Phys. Chem. B* **2002**, *106*, 7300–7307.
- (43) CPMD v3.13.1., Copyright IBM Corp, 1990-2008; Copyright MPI für Festkörperforschung Stuttgart, 1997-2001; <http://www.cpmd.org/> (access date: 01/07/2015).
- (44) van Gunsteren, W. F.; Billeter, S. R.; Eising, A. A.; Huenenberger, P. H.; Krueger, P.; Mark, A. E.; Scott, W. R. P.; Tironi, I. G. *Biomolecular Simulation: The Gromos96 Manual and User Guide*; 1996, Hochschulverlag an der ETH Zurich.
- (45) Duan, Y.; Wu, C.; Chowdhury, S.; Lee, M. C.; Xiong, G.; Zhang, W.; Yang, R.; Cieplak, P.; Luo, R.; Lee, T.; Caldwell, J.; Wang, J.; Kollman, P. *J. Comput. Chem.* **2003**, *24*, 1999–2012.
- (46) Jorgensen, W. L.; Chandrasekhar, J.; Madura, J. D.; Impey, R. W.; Klein, M. L. *J. Chem. Phys.* **1983**, *79*, 926–935.
- (47) Perdew, J. P.; Burke, K.; Ernzerhof, M. *Phys. Rev. Lett.* **1996**, *77*, 3865–3868.
- (48) Perdew, J. P.; Burke, K.; Ernzerhof, M. *Phys. Rev. Lett.* **1997**, *78*, 1396.

- (49) Troullier, N.; Martins, J. L. *Phys. Rev. B* **1991**, *43*, 1993–2006.
- (50) Berendsen, H. J. C.; Postma, J. P. M.; van Gunsteren, W. F.; DiNola, A.; Haak, J. R. *J. Chem. Phys.* **1984**, *81*, 3684–3690.
- (51) Nosé, S. *J. Chem. Phys.* **1984**, *81*, 511–519.
- (52) Hoover, W. G. *Phys. Rev. A* **1985**, *31*, 1695–1697.
- (53) Aquilante, F.; Vico, L. D.; Ferré, N.; Ghigo, G.; Malmqvist, P.-Å.; Neogrády, P.; Pedersen, T. B.; Pitonák, M.; Reiher, M.; Roos, B. O.; Serrano-Andrés, L.; Urban, M.; Veryazov, V.; Lindh, R. *J. Comp. Chem.* **2010**, *31*, 224–247.
- (54) Huang, P.; Carter, E. A. *J. Chem. Phys.* **2006**, *125*, 084102.
- (55) Sharifzadeh, S.; Huang, P.; Carter, E. A. *Chem. Phys. Lett.* **2009**, *470*, 347–352.
- (56) Widmark, P.-O.; Malmqvist, P.-Å.; Roos, B. O. *Theor. Chem. Acc.* **1990**, *77*, 291–306.
- (57) Widmark, P.; Malmqvist, P.; Roos, B. O. *Theor. Chem. Acc.* **1990**, *77*, 291–306.
- (58) Dunning Jr, T. H. *J. Chem. Phys.* **1989**, *90*, 1007–1023.
- (59) Aquilante, F.; Malmqvist, P.-Å.; Pedersen, T. B.; Ghosh, A.; Roos, B. O. *J. Chem. Theory Comput.* **2008**, *4*, 694–702.
- (60) Ghigo, G.; Roos, B. O.; Malmqvist, P.-Å. *Chem. Phys. Lett.* **2004**, *396*, 142–149.
- (61) Forsberg, N.; Malmqvist, P.-A. *Chem. Phys. Lett.* **1997**, *274*, 196–204.
- (62) Wang, J.; Cieplak, P.; Kollman, P. A. *J. Comput. Chem.* **2000**, *21*, 1049–1074.
- (63) Wang, J.; Cieplak, P.; Li, J.; Hou, T.; Ray, L.; Yong, D. *J. Chem. Phys. B* **2011**, *8*, 3091–3099.
- (64) Wang, J.; Cieplak, P.; Li, J.; Wang, J.; Cai, Q.; Hsieh, M.; Lei, H.; Luo, R.; Duan, Y. *J. Chem. Phys. B* **2011**, *8*, 3100–3111.

- (65) Li, Q.; Mennucci, B.; Robb, M. A.; Blancafort, L.; Curutchet, C. *J. Chem. Theory Comput.* **2015**, *11*, 1674–1682.
- (66) Frisch, M. J. et al. *Gaussian 09 Revision A.2*, Gaussian 09 Revision A.2, Gaussian Inc. Wallingford CT 2009.
- (67) Yanai, T.; Tew, D. P.; Handy, N. C. *Chem. Phys. Lett.* **2004**, *393*, 51–57.
- (68) Iikura, H.; Tsuneda, T.; Yanai, T.; Hirao, K. *J. Chem. Phys.* **2001**, *115*, 3540–3544.
- (69) Hehre, W. J.; Ditchfield, R.; Pople, J. A. *J. Chem. Phys.* **1972**, *56*, 2257–2261.
- (70) Clark, T.; Chandrasekhar, J.; Spitznagel, G. W.; Schleyer, P. v. R. *J. Comput. Chem.* **1983**, *4*, 294–301.
- (71) Frisch, M. J.; Pople, J. A.; Binkley, J. S. *J. Chem. Phys.* **1984**, *80*, 3265–3269.
- (72) Curutchet, C.; Novoderezhkin, V. I.; Kongsted, J.; Muñoz-Losa, A.; van Grondelle, R.; Scholes, G. D.; Mennucci, B. *J. Phys. Chem. B* **2013**, *117*, 4263–4273.
- (73) te Velde, G.; Bickelhaupt, F. M.; Baerends, E. J.; van Gisbergen, S. J. A.; Fonseca Guerra, C.; Snijders, J. G.; Ziegler, T. *J. Comput. Chem.* **2001**, *22*, 931–967.
- (74) Fonseca Guerra, C.; Snijders, J. G.; te Velde, G.; Baerends, E. J. *Theor. Chem. Acc.* **1998**, *99*, 391–403.
- (75) ADF2013, SCM, Theoretical Chemistry, Vrije Universiteit, Amsterdam, The Netherlands, <http://www.scm.com> (access date: 06/06/2014).
- (76) Lenthe, E. V.; Baerends, E. J. *J. Comput. Chem.* **2003**, *24*, 1142–1156.
- (77) Zhao, Y.; Truhlar, D. G. *J. Phys. Chem. A* **2006**, *110*, 13126–13130.
- (78) Lembarki, A.; Chermette, H. *Phys. Rev. A* **1994**, *50*, 5328–5331.
- (79) Perdew, J. P. *Phys. Lett. A* **1992**, *165*, 79–82.



- (80) Akinaga, Y.; Ten-no, S. *Chem. Phys. Lett.* **2008**, *462*, 348–351.
- (81) Seth, M.; Ziegler, T. *J. Chem. Theory Comput.* **2012**, *8*, 901–907.
- (82) Angeli, C.; Cimiraglia, R.; Evangelisti, S.; Leininger, T.; Malrieu, J.-P. *J. Chem. Phys.* **2001**, *114*, 10252–10264.
- (83) Angeli, C.; Cimiraglia, R.; Malrieu, J.-P. *Chem. Phys. Lett.* **2001**, *350*, 297–305.
- (84) Angeli, C.; Cimiraglia, R.; Malrieu, J.-P. *J. Chem. Phys.* **2002**, *117*, 9138–9153.
- (85) F., N. *Wiley Interdiscip. Rev.: Comput. Mol. Sci.* **2012**, *2*, 73–78.
- (86) Eichkorn, K.; Treutler, O.; Öhm, H.; Hada, M.; Häser, M.; Ahlrichs, R. *Chem. Phys. Lett.* **1995**, *240*, 283–290.
- (87) Neese, F.; Wennmohs, F.; Hansen, A.; Becker, U. *Chem. Phys.* **2009**, *356*, 98–109.
- (88) Dyal, K. G. *J. Chem. Phys.* **1995**, *102*, 4909–4918.
- (89) Angeli, C.; Cimiraglia, R.; Malrieu, J.-P. *Chem. Phys. Lett.* **2000**, *317*, 472–480.
- (90) CHAMP is a quantum Monte Carlo program package written by C. J. Umrigar, C. Filippi and collaborators.
- (91) Burkatzki, M.; Filippi, C.; Dolg, M. *J. Chem. Phys.* **2007**, *126*, 234105.
- (92) For the hydrogen atom, we use a more accurate BFD pseudopotential and basis set. Dolg, M.; Filippi, C., private communication.
- (93) We add one s and one p diffuse function on the carbon and the nitrogen using exponents from the aug-cc-pVDZ basis set, taken from EMSL Basis Set Library <http://bse.pnl.gov> (access date: 01/07/2015).
- (94) Filippi, C.; Umrigar, C. J. *J. Chem. Phys.* **1996**, *105*, 213–226, As Jastrow correlation factor, we use the exponential of the sum of three fifth-order polynomials of the electron-nuclear

(e-n), the electron-electron (e-e). The Jastrow factor is adapted to deal with pseudo-atoms, and the scaling factor  $\kappa$  is set to 0.6 a.u. The 2-body Jastrow factor includes five parameters in the e-e terms and four parameters for each atom type in the e-n terms.

- (95) Umrigar, C. J.; Toulouse, J.; Filippi, C.; Sorella, S.; Hennig, R. G. *Phys. Rev. Lett.* **2007**, *98*, 110201.
- (96) Filippi, C.; Zaccheddu, M.; Buda, F. *J. Chem. Theory Comput.* **2009**, *5*, 2074–2087.
- (97) Casula, M. *Phys. Rev. B* **2006**, *74*, 161102.
- (98) Phillips, J. C.; Braun, R.; Wang, W.; Gumbart, J.; Tajkhorshid, E.; Villa, E.; Chipot, C.; Skeel, R. D.; Kalé, L.; Schulten, K. *J. Comput. Chem.* **2005**, *26*, 1781–1802.
- (99) Heyer, L. J.; Kruglyak, S.; Yooseph, S. *Genome Res.* **1999**, *9*, 1106–1115.
- (100) Humphrey, W.; Dalke, A.; Schulten, K. *J. Mol. Graphics* **1996**, *14*, 33–38.
- (101) Chatteraj, M.; King, B. A.; Bublit, G. U.; Boxer, S. G. *Proc. Natl. Acad. Sci.* **1996**, *93*, 8362–8367.
- (102) Brejc, K.; Sixma, T. K.; Kitts, P. A.; Kain, S. R.; Tsien, R. Y.; Ormö, M.; Remington, S. J. *Proc. Natl. Acad. Sci.* **1997**, *94*, 2306–2311.
- (103) Stoner-Ma, D.; Jaye, A. A.; Matousek, P.; Towrie, M.; Meech, S. R.; Tonge, P. J. *J. Am. Chem. Soc.* **2005**, *127*, 2864–2865.
- (104) Nifosi, R.; Tozzini, V. *Proteins: Struct., Funct., Genet.* **2003**, *51*, 378–389.
- (105) Ormö, M.; Cubitt, A. B.; Kallio, K.; Gross, L. A.; Tsien, R. Y.; Remington, S. J. *Science* **1996**, *273*, 1392–1395.
- (106) Elsliger, M.; Wachter, R.; Hanson, G.; Kallio, K.; Remington, S. *Biochemistry* **1999**, *38*, 5296–5301.

- (107) Jain, R. K.; Ranganathan, R. *Proc. Natl. Acad. Sci.* **2004**, *101*, 111–116.
- (108) Ward, W. W.; Cody, C. W.; Hart, R. C.; Cormier, M. J. *Photochem. Photobiol.* **1980**, *31*, 611–615.
- (109) Niwa, H.; Inouye, S.; Hirano, T.; Matsuno, T.; Kojima, S.; Kubota, M.; Ohashi, M.; Tsuji, F. I. *Proc. Natl. Acad. Sci.* **1996**, *93*, 13617–13622.
- (110) Webber, N. M.; Litvinenko, K. L.; Meech, S. R. *J. Phys. Chem. B* **2001**, *105*, 8036–8039.
- (111) Litvinenko, K. L.; Webber, N. M.; Meech, S. R. *Chem. Phys. Lett.* **2001**, *346*, 47–53.
- (112) Mandal, D.; Tahara, T.; Webber, N. M.; Meech, S. R. *Chem. Phys. Lett.* **2002**, *358*, 495–501.
- (113) Creemers, T. M. H.; Lock, A. J.; Subramaniam, V.; Jovin, T. M.; Völker, S. *Nat. Struct. Biol.* **1999**, *6*, 557–560.
- (114) Kobayashi, R.; Amos, R. D. *Chem. Phys. Lett.* **2006**, *420*, 106–109.
- (115) Cai, Z.-L.; Crossley, M. J.; Reimers, J. R.; Kobayashi, R.; Amos, R. D. *J. Phys. Chem. B* **2006**, *110*, 15624–15632.
- (116) Tozer, D. J.; Amos, R. D.; Handy, N. C.; Roos, B. O.; Serrano-Andres, L. *Mol. Phys.* **1999**, *97*, 859–868.
- (117) Dreuw, A.; Head-Gordon, M. *J. Am. Chem. Soc.* **2004**, *126*, 4007–4016.
- (118) Wanko, M.; Garcia-Risueño, P.; Rubio, A. *Phys. Status Solidi B* **2012**, 392–400.
- (119) Olsen, S. *J. Chem. Theory Comput.* **2010**, *6*, 1089–1103.
- (120) Olsen, S.; McKenzie, R. H. *J. Chem. Phys.* **2011**, *134*, 114520.
- (121) Olsen, S. *J. Phys. Chem. B* **2015**, *119*, 2566–2575.
- (122) Armengol, P.; Gelabert, R.; Moreno, M.; Lluch, J. M. *Org. Biomol. Chem.* **2014**, *12*, 9845–9852.

- (123) Wanko, M.; Hoffmann, M.; Frauenheim, T.; Elstner, M. *J. Phys. Chem. B* **2008**, *112*, 11462–11467.
- (124) Caricato, M.; Mennucci, B.; Tomasi, J.; Ingrosso, F.; Cammi, R.; Corni, S.; Scalmani, G. *J. Chem. Phys.* **2006**, *124*, 124520.
- (125) Cammi, R.; Corni, S.; Mennucci, B.; Tomasi, J. *J. Chem. Phys.* **2005**, *122*, 104513.
- (126) Corni, S.; Cammi, R.; Mennucci, B.; Tomasi, J. *J. Chem. Phys.* **2005**, *123*, 134512.
- (127) Wesolowski, T. A.; Warshel, A. *J. Phys. Chem.* **1993**, *97*, 8050–8053.
- (128) Humbert-Droz, M.; Zhou, X.; Shedge, S. V.; Wesolowski, T. A. *Theor. Chem. Acc.* **2014**, *133*, 1405.
- (129) Wesolowski, T. A.; Weber, J. *Chem. Phys. Lett.* **1996**, *248*, 71–76.
- (130) Daday, C.; König, C.; Neugebauer, J.; Filippi, C. *ChemPhysChem* **2014**, *15*, 3205–3217.
- (131) Gagliardi, L.; Lindh, R.; Karström, G. *J. Chem. Phys.* **2004**, *121*, 4494–4500.
- (132) Fux, S.; Jacob, C. R.; Neugebauer, J.; Visscher, L.; Reiher, M. *J. Chem. Phys.* **2010**, *132*, 164101.
- (133) Manby, F. R.; Stella, M.; Goodpaster, J. D.; Miller, T. F. *J. Chem. Theory Comput.* **2012**, *8*, 2564–2568.