University of Pisa
Department of Chemistry and Industrial Chemistry

Doctoral School in Chemistry and Material Science

Doctoral Thesis

The dynamics of photoactive pigment-protein complexes and its influence on structural and spectroscopic properties

Author:
Felipe Cardoso Ramos

Supervisor:
Prof. Benedetta Mennucci

April 2021

"Nothing in life is to be feared, it is only to be understood. Now is the time to understand more, so that we may fear less."

Marie Curie

University of Pisa

Abstract

Doctoral Thesis

Doctoral School in Chemistry and Material Science

The dynamics of photoactive pigment-protein complexes and its influence on structural and spectroscopic properties

by Felipe Cardoso Ramos

The photochemistry and photophysics of organic chromophores in photoresponsive proteins can be tuned by non-covalent protein-chromophore interactions, which are experimentally difficult to access with a temporal resolution. Here, we have combined classical molecular dynamics simulations and multiscale QM/MM(pol) calculations to study light-harvesting antennae from purple bacteria (LH2) and synthetic flavin-based fluorescent proteins (FbFPs), derived from plant LOV (light, oxygen, and voltage) sensing domains, with improved fluorescence (iLOV). For the LH2 complexes, we found out that the mechanisms that govern their adaptation (and spectral tuning) to different light conditions exploit the different H-bonding environment around the bacteriochlorophyll pigments to dynamically control both internal and inter-pigment degrees of freedom. The results obtained for the FbFPs indicate that the H-bonding dynamics impact the excitation and emission vertical energies of the studied systems, but they cannot explain the increased fluorescence. Instead, such an improvement (increased fluorescence) is reached by a better packing of the flavin binding-site induced by specific mutations which stabilize the chromophore through optimized van der Waals interactions.

Acknowledgments

I would like to thank the University of Pisa, the Department of Chemistry and Industrial Chemistry (DCCI), the Doctoral School in Chemistry and Material Science (DSCM), and everyone from the MoLECoLab group, in particular Michelle Nottoli and Lorenzo Cupellini, who collaborated with me in this study. I also want to thank all my colleagues from the 33rd cycle of the PhD course, especially Nicola di Fidio, who helped me with the practical and bureaucratic issues of the doctorate program, and the doctorate school coordinator, Professor Gennaro Pescitelli, for all his work on behalf of the PhD students.

I would also like to show my sincere appreciation to my friends Eduarda Sangiogo, Ingrid Prandi, and Davide Accomasso, as well as my family, especially my parents and grandma, and all those who supported me directly or indirectly in my personal PhD journey. Furthermore, I want to express my gratitude to Gabriela Alcantara for the help regarding the PhD selection, visa and student permit applications, as well as other issues releted to Italian documents.

Last, but not least, I deeply thank Professor Benedetta Mennucci for her work as my supervisor over those three years. I am really grateful for the opportunity to be part of her research group, both during my undergraduate internship and during my PhD. Professor Mennucci, thank you for all the advice, constructive criticism, and teachings you so generously shared with me. I am sure that the lessons I learned from you and at the MoLECoLab group will be decisive for my professional career.

To my parents Nonato and Helena,
and my grandmother Neidirce.

Contents

1 Introduction **1**

2 Molecular Dynamics: General Aspects and its Application to Proteins **6**

2.1 Molecular Mechanics 8
2.2 Initial configuration 11
2.3 Solving the equations of motion 12
2.3.1 The Verlet method 13
2.3.2 The Velocity Verlet algorithm 14
2.3.3 Defining the integration time 15
2.4 Periodic boundary conditions 16
2.5 Simulation protocol 18
2.6 Trajectory analysis 20

3 Multiscale Methods: Combining Quantum Chemistry and Molecular Mechanics **23**

3.1 The QM/MM energy 24
3.1.1 Subtractive QM/MM Schemes 25
3.1.2 Additive QM/MM Schemes 26
3.2 The polarizable formulation of QM/MM 29
3.3 QM/MM applications in computational spectroscopy 32
3.3.1 Single chromophore systems 33

3.3.2 Multichromophoric systems 34

4 Case Study I: LH2 Complexes from Purple Bacteria 39

4.1 Introduction 41
4.2 Computational details 46
4.2.1 Structures 46
4.2.2 Molecular dynamics 47
4.2.3 Excitonic states 48
4.2.4 Site energies and couplings 49
4.2.5 Coupling to charge-transfer states 49
4.2.6 Geometry optimizations 50
4.3 The picture from the combined MD and QM/MMPol approach 51
4.4 Dissecting the possible origins of the spectral differences . . . 58
4.5 Why is PucD so blue-shifted? 67
4.6 Conclusions 68

5 Case Study II: LOV-based Fluorescent Proteins 74

5.1 Introduction 76
5.2 Computational details 80
5.2.1 Molecular dynamics 80
5.2.2 QM/MM(Pol) calculations 81
5.2.3 Lineshape calculations 84
5.3 Results and Discussion 86
5.3.1 Structural analysis 86
5.3.2 Spectroscopic analysis 94
5.4 Conclusions 98

6 Conclusions 99

A Supporting information for Chapter 4 102

A.1 Additional figures 103
A.2 Additional tables 113

B Supporting information for Chapter 5 **116**
B.1 Additional figures 117

Bibliography **128**

Chapter 1

Introduction

The maintenance of life depends on a constant search for internal chemical and physical balance. To stay in such a balance, the living organisms must be able to deal with the changes in the surrounding environment. Proteins are the effector macromolecules of the changes undergone by the cells in response to external stimuli. They can act in many stages of chain processes that lead to a given adaptive modification. Such modifications can be, for example, changes in the functionality of the cell membrane, allowing the entry of an alternative metabolite or the recognition of a different chemical signal involved in cell communication. Other changes can occur at the genetic level, acting on the regulation of gene expression. However, in the intricate metabolic and signaling networks inside living cells, the most common situation is that these adaptation processes are dynamically interconnected and mediated by different types of proteins (e.g. receptors, enzymes, transporters, DNA and RNA-binding proteins, and so on) [1, 2]. This central role played by proteins explains why it is so important to understand their function and how it is related to their structure.

From 1900, the number of protein structures deposited the Protein Data Bank (PDB) began to increase dramatically and that growth continues to increase [3]. Currently, the PDB has more than 170,000 available entries. Most of these structures were determined by X-ray crystallographic (88 %),

following by the nuclear magnetic resonance, NMR (8 %), and 3D electron microscopy, 3DEM (4 %), techniques [4].

The availability of such an amount of structural information, along with increasing data processing capacity, promoted the emergence of the structural bioinformatics and opened up new possibilities for protein simulation studies based on theoretical and computational chemistry methods. The first is manly related to the study of the structural variability found in the databases and to the development of structural prediction methods for proteins not yet deposited. The second focuses on structural processes of dynamic nature, which are extremely important to understand the function of proteins in the cellular environment [5].

Similarly to how the DNA sequence of the genome cannot reveal everything about a cell's physiology, the structure of a protein cannot be accurately inferred only from its amino acid sequence, just as it is not prudent to describe how a given protein works simply by observing its crystal structure. This is because the *modus operandi* of most proteins depends on dynamic factors related to their structure. The way the proteins interact with other molecules (small ligand molecules, other proteins, DNA, ecc.) depends on how they fit the surface of these molecules. Such docking and undocking ability is only possible due to proteins conformational plasticity.

Some regulatory proteins, for example, can be extremely flexible. This is the case of the intrinsically disordered proteins (IDPs) that account for nearly 40 % of our genome and lack a defined 3D structure so that they should be described as a dynamic ensemble [5]. Other proteins have only regions of high flexibility, as in the case of antibodies that present high flexible hinge regions [6], which have important functional implications for antigen recognition [7]. In addition, it is important to take into account the flexibility of the amino acid side chains that can significantly influence the protein function, even in proteins with rigid backbone structures [2].

Currently, the diversity of known proteins is such that cannot be covered only through the experimental methods currently available. It is also worth mentioning that some proteins are not able to form crystals capable

of being analyzed by X-ray crystallography. In addition, crystallographic studies can provide important structural insights but do not allow an accurate dynamic analysis of the protein structures already solved. These factors have made computational approaches in structural protein studies increasingly common.

In this Thesis, we present a computational investigation of proteins involved in different photoinduced processes [8–11]. More specifically, we studied proteins that initiate the photosynthesis through the so-called light-harvesting step, and proteins that act as photoreceptors and regulate a series of cellular processes such as stress response and phototaxis. To achieve such a goal we combine classical Molecular Dynamics (MD) with quantum-chemistry based multiscale approaches.

Classical Molecular Dynamics (MD) are crucial to investigate protein dynamics with a time scale resolution [5, 12–14]. They offer the opportunity to access conformational changes in the protein backbone or side chains, which are difficult to describe atomistically through experimental methods. Moreover, when combined with Quantum Mechanics (QM), classical MD becomes a powerful tool for describing photoinduced processes in biological systems [15, 16]. The classical MD simulation, as well as the hybrid approaches that combine classical and quantum methods are discussed in more details in the Chapters 2 and 3, respectively.

In Chapters 4 and 5, we present the two selected Case Studies (Fig. 1.1).

In the Case Study I, we focused on the light-harvesting (LH) complexes of purple bacteria. In photosynthetic organisms, LH complexes contain an aggregate of pigments which are used as antennae to absorb sunlight and transfer the excitation energy to the reaction centers, where the primary electron transfer events occur. While the excitation energy transfer is extremely efficient, there is a limit to the amount of energy that can be processed in the reaction center. This is a clear problem, especially in high-light conditions when the excess excitation energy can become extremely dangerous for the photosynthetic organism. The various organisms have developed different

strategies to adapt to the different light conditions. Here we have investigated the case of purple bacteria where the LH complexes expressed in low- and high-light conditions present significant different spectroscopic properties so to maintain a correct functioning of the photosystem.

In the case study II, we analyzed synthetic flavin-based fluorescent proteins (FbFPs) derived from plant LOV (light, oxygen, and voltage) sensing domains. The LOV domains are monocromophoric proteins which integrate the photosensitive region of larger proteins that are involved in many cell signaling pathways. Here, we focused on the study of synthetic LOV domains, which through mutagenesis, have improved their fluorescence properties. One of them (iLOV), is currently a popular imaging biomarkers. The goal is to clarify how mutations in the protein sequence alter the dynamics of the system, as well as to investigate the influence of such modifications on the spectroscopic properties.

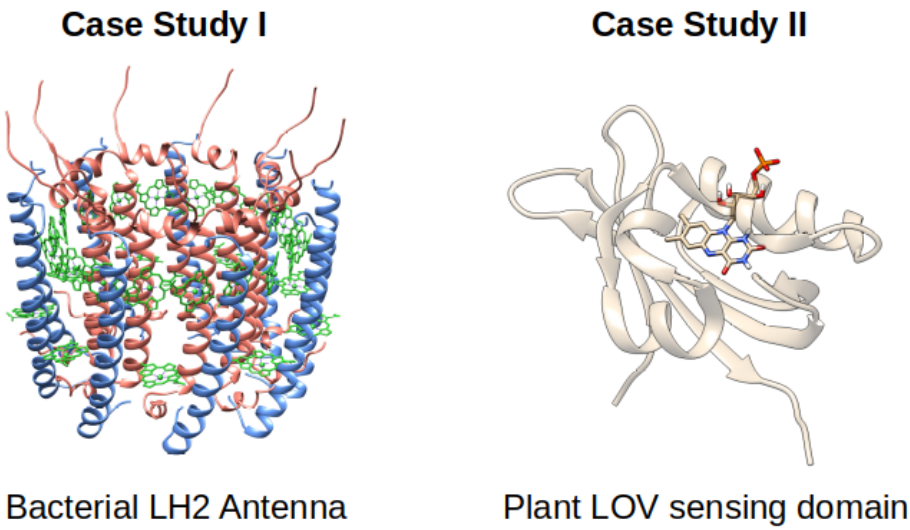

Figure 1.1: The proteins studied in the Case Studies I and II. The figure shows the LH2 antenna from purple bacteria (right) and the plant LOV sensing domain (left). See Chapters 4 and 5, as well as the Appendices A and B for more details about the systems studied and the obtained results.

Finally, in Chapter 6, we report some final remarks and in Appendices A and B we provide support information for Chapters 4 and 5, respectively.

Chapter 2

Molecular Dynamics: General Aspects and its Application to Proteins

In a very general sense, Molecular Dynamics (MD) is a computational method to study the time evolution of interacting atoms or molecules. This is achieved by generating a sequence of configurations of the system, or trajectory, based on a given equation of motion in combination with selected initial conditions. In the most common formulation of MD, classical Molecular Mechanics (MM) is used to model the interactions and the trajectory is obtained through the numerical resolution of the differential equations of motion embodied in the Newton's second law.

The main steps of the MD simulation are summarized in the flowchart of Fig. 2.1. Basically, starting from an initial configuration of the system, the forces acting on them are calculated from the MM potential. The atoms are then moved by integrating the equations of motion and a new configuration for the system is generated. The new configuration is used as a starting point in a successive MD step and the entire cycle is repeated until the chosen time window has been completed.

Each configuration for a system of N particles can be defined by a set of

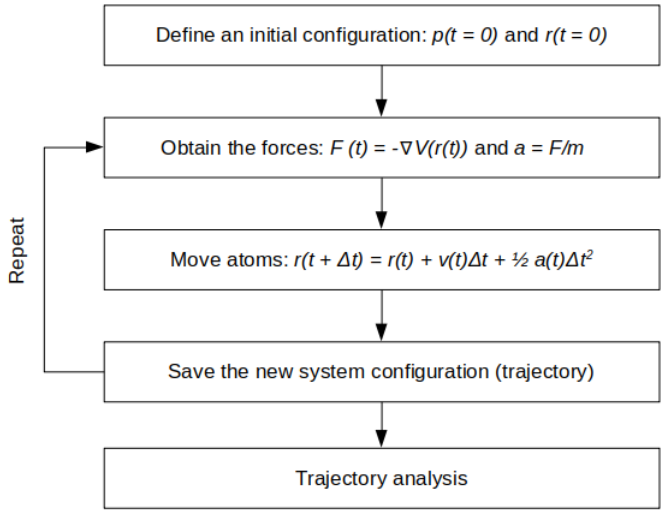

Figure 2.1: Simplified flowchart showing the main steps involved in MD simulation.

atomic positions, r^N , and momenta, p^N. As each position has 3 coordinates and each momentum has 3 components, for N atoms would exist $6N$ degrees of freedom. We can think of one particular set of coordinates and momenta as a point in this $6N$ dimensional phase space.

Some sets of macroscopic parameters can be kept constant during the simulation, such as the number of atoms in the system N, the volume V (or the pressure P), the temperature T or the total energy E. The set of constant parameters characterizes what is called the statistical ensemble. An ensemble contains a large group of replicas for a given system which differ from each other in the particle coordinates (r^N) and momenta (p^N) so that each replica occupies a region in the $6N$ dimensional phase space [17].

If the phase space is completely covered after a certain time of simulation, the system is considered to be ergodic and the temporal average of a propriety X can be computed as the average of the ensemble. But, in practice, average properties are obtained as the weighted average calculated for a set of M representative system configurations:

$$\langle X \rangle = \frac{1}{M} \sum_{t=1}^{M} X_t(\Gamma)\rho(\Gamma) \tag{2.1}$$

where $\rho(\Gamma)$ is the probability density associated with a particular point $\Gamma(r^N, p^N)$ in phase space [18].

In particular, an isolated system, without any external perturbation, evolves by conserving its total energy. The resulting ensemble is called microcanonical ensemble (NVE). The NVE is considered to be the natural ensemble of the MD method. However, for the trajectories obtained in NVE to be representative of the real system, temperature or both temperature and pressure (NPT) control is necessary.

In the following sections a presentation of the main ingredients of a classical MD simulation will be given, starting from a description of the MM potential. In such a general presentation, particular emphasis will be given to the application to proteins.

2.1 Molecular Mechanics

In Molecular Mechanics (MM), atoms are represented explicitly but they lose their quantum nature being described as classical particles of a given mass and a charge. The atom-atom interaction potential (V_{total}) can be defined as the sum of two terms:

$$V_{total} = V_{bonded} + V_{non-bonded} \tag{2.2}$$

where V_{bonded} results from interactions between close atoms in the covalent structure and $V_{non-bonded}$ from interactions between atoms which are at a distance of more than two bonds in a given molecule or they belong to two different molecules.

The potential V_{bonded} has terms referring, for example, to bond stretching (V_l), bending (V_θ) and torsional (V_ϕ) motions. Each term is described by a simple mathematical function. For both V_l) and V_θ harmonic potentials

are commonly used while for V_ϕ an expansion of periodic functions is the standard choice:

$$V_l = k_l(l - l_0)^2 \tag{2.3}$$

$$V_\theta = k_\theta(\theta - \theta_0)^2 \tag{2.4}$$

$$V_\phi = \sum_n V_n/2 \cdot [1 + cos(n\phi - \gamma)] \tag{2.5}$$

In Eqs. 2.3 and 2.4, l and θ are, respectively, the bond length and the bond angle, whereas l_0 and θ_0 are the corresponding minimum energy values and k_l and k_θ correspond to the force constants. In Eq. 2.5, ϕ is the dihedral angle, n is the periodicity of the cosine function, V_n the corresponding amplitude of the energy barrier and γ the phase angle.

The interaction potential $V_{non-bonded}$ can be decomposed into a van der Waals (V_{vdW}) and an electrostatic term (V_e). For the former a Lennard-Jones (LJ) potential is commonly used, whereas the latter is expressed as a Coulomb interaction between point charges, namely:

$$V_{LJ}(r_{ij}) = 4\epsilon_{ij}[(\sigma_{ij}/r_{ij})^{12} - (\sigma_{ij}/r_{ij})^6] \tag{2.6}$$

$$V_e(r_{ij}) = q_i q_j / r_{ij} \tag{2.7}$$

where r_{ij} is the distance between atoms i and j.

The Lennard-Jones potential is composed of an attractive term that varies with r^{-6} and a repulsive term, proportional to r^{-12}. The attractive term is associated with dispersion interactions whereas the repulsive term represents the strong short-range repulsion caused by the overlapping of the electronic clouds from two atoms, according to Pauli exclusion principle. The combination of these terms causes the Lennard-Jones potential to assume a potential well shape, as shown in Fig. 2.2. Setting the derivative equal to zero in order to find the minimum in such a potential we see that $r_{ij}^{min} = 2^{1/6}\sigma_{ij}$ and $V_{LJ}(r_{ij}^{min}) = -\epsilon_{ij}$.

The Coulomb potential describes the electrostatic interactions for each pair of atoms i and j, with partial charges q_i and q_j, respectively. Since

the Coulomb potential is proportional to r^1, its decay is much slower compared to the Lennard-Jones potential and therefore it gives rise to long-range interactions.

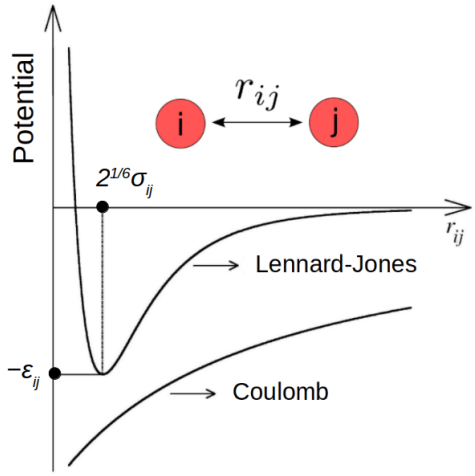

Figure 2.2: The Lennard-Jones e Coulomb potentials.

To summarize, a typical MM potential can be represented in the following way:

$$\begin{aligned} V_{total} = &\sum_{bonds} k_l(l-l_0)^2 + \sum_{angles} k_\theta(\theta-\theta_0)^2 + \\ &+ \sum_{dihedral}\sum_{n} V_n/2 \cdot [1 + cos(n\phi - \gamma)] + \\ &+ \sum_i \sum_{j>i} 4\epsilon_{ij}[(\sigma_{ij}/r_{ij})^{12} - (\sigma_{ij}/r_{ij})^6] + \sum_i \sum_{j>i} q_i q_j / r_{ij} \end{aligned} \tag{2.8}$$

The specific mathematical form of V_{total} together with the values of the numerical parameters defines what is called *force field*. Several MM force fields have been developed and applied in different areas. In the case of biological systems, the most commonly used force fields are CHARMM [19], OPLS [20], AMBER [21] and GROMOS [22].

2.2 Initial configuration

Once the force field is defined, it is necessary to obtain an initial configuration for the system, that is, the initial coordinates and velocities of each atom.

In the case of systems involving proteins, crystallographic structures are generally used to describe the initial atomic coordinates and they can be obtained from structural biology databases such as the Protein Data Bank (PDB). Hydrogen atoms, however, need special attention because they are difficult to locate by X-ray crystallography and hence their coordinates are very often missing from protein crystal structures. In such a case, the missing H atoms must be added to the system before the MD simulation. This is a delicate and crucial step because it involves understanding what is the most likely protonation state for the protein residues. This task is usually done with the help of web servers like H++ [23]. In some cases, the structure of the protein to be studied has not been characterized but it is still possible to obtain an initial model by using *ab initio* [24] or comparative structural prediction methods based on the sequence similarity [25].

In most cases, the protein has to be simulated in aqueous solution in order to reproduce the conditions found in living systems. When the solvent is explicitly treated (most of cases), its components are added to the system with the help of auxiliary softwares which perform the solvation of the protein (or protein-ligand complex). An example of the resulting simulation box is reported in Fig. 2.3. The water molecules could be added randomly to the system, but this is a risky procedure as artificially large interactions between the solvent molecules are possible. Therefore, the initial arrangement of the water molecules is usually defined by its crystalline structure [18]. Ions can also be added by deleting water molecules to create empty spaces where the charged atoms (like Na^+ and Cl^-) can be inserted. This is usually done to mimic the ionic strength found in biological cells, as well as to neutralize the system total charge.

Before the MD simulations starts, it is usually necessary to carry out

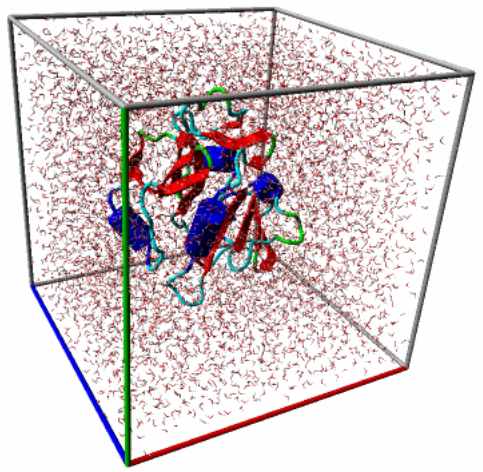

Figure 2.3: Simulation box containing a protein molecule and water as solvent.

an energy minimization step, which can be performed in several stages by optimizing different parts of the system separately. The goal here is to eliminate possible artificially high interactions, as well as to reduce the forces as much as possible in order to provide a feasible starting point for MD.

The initial velocities in a MD simulation are assigned according to the Maxwell-Boltzmann distribution for the desired temperature T:

$$p(v_{ix}) = \sqrt{\frac{m_i}{2\pi kT}} \exp\left(-\frac{m_i v_{ix}^2}{2kT}\right) \tag{2.9}$$

where $p(v_i)$ is the probability that an atom i of mass m_i has a velocity v_{ix} in the x direction. The algorithm makes sure that the randomly assigned set of initial velocities sticks to such probability distribution.

2.3 Solving the equations of motion

Once defined the model for calculating the potential (and the corresponding forces) and obtained the initial configuration of the system, it is possible to solve the differential equations of motion for each atom, accord-

ing to Newton's second law. Below, the Eq. 2.10 describes the movement of an atom i of the system, with mass m_i and coordinates $\vec{r_i}$, under the action of a force $\vec{F_i}$ at time t:

$$d^2\vec{r_i}/dt^2 = \vec{F_i}(t)/m_i \tag{2.10}$$

where $d^2\vec{r_i}/dt^2$ is the atom acceleration ($\vec{a_i}$) and $\vec{F_i}$ is given by:

$$\vec{F_i}(t) = -\nabla V_{total}(t) = -\partial V_{total}/\partial\vec{r_i}(t) \tag{2.11}$$

where V_{total} is the potential energy at time t, calculated by the MM force field which depends on the coordinates of each atom of the system.

In principle, the integration of Eq. 2.10 should allow to determine the new coordinates and velocities for the atom in the subsequent step, that is, at $t + dt$. By successively applying this process, it should be possible to obtain the trajectory, which is nothing more than the sequence of system configurations generated along the MD simulation. However, as the movements of the atoms are coupled, the resulting many-bodies problem cannot be solved analytically. Then, the differential equations need to be solved using finite difference methods [17].

Several integration algorithms, based on finite difference methods, can be used in MD. These algorithms assume that the integration can be divided into a series of small steps, separated by a fixed time interval (δt) and that positions and velocities can be expanded into Taylor series. One of the most used methods for this purpose is the Verlet method, which will be discussed below.

2.3.1 The Verlet method

The Verlet method allows to obtain the coordinates of a particle i in the times $t \pm \delta$, as follows:

$$r(t + \delta t) = r(t) + \delta t v(t) + \frac{1}{2}\delta t^2 a(t) + ... \tag{2.12}$$

$$r(t-\delta t) = r(t) - \delta t v(t) + \frac{1}{2}\delta t^2 a(t) + ... \tag{2.13}$$

where $v(t)$ is the velocity (or the first derivative of the position with respect to time) and $a(t)$ the acceleration (or second derivative).

By adding the Eqs. 2.12 and 2.13, we obtain the next truncated expression:

$$r(t+\delta t) = 2r(t) - r(t-\delta t) + \delta t^2 a(t) \tag{2.14}$$

where $a(t)$ can be computed from eq. 2.10. Note that, in the first run, $r(t-\delta t)$ is unknown and must be estimated using the following equations:

$$r(t-\delta t) = r_0 - v_0 \delta t \tag{2.15}$$

where r_0 and v_0 are, respectively, the initial position and velocity (at $t = 0$), needed to start the process.

Although $v(t)$ does not appear explicitly in Eq. 2.14, it is required to compute the kinetic energy and can be estimated as:

$$v(t) = \frac{r(t+\delta t) - r(t-\delta t)}{2\delta t} \tag{2.16}$$

It is worth to mention that solving the equations 2.14 and 2.16 lead to errors in the order of $O(\delta t^4)$ and $O(\delta t^2)$, respectively. Where $O(\delta t^n)$ means the n^{th}-order error due to the truncation of the series. Also, a consequence of having to estimate the velocity using the Eq. 2.16 is that, at the end of each run, the velocity is a step behind the position since the velocity is obtained for t and the position is obtained for $t + \delta t$. Another disadvantage of the Verlet method is that variables of different steps must be processed simultaneously so that its implementation can be memory-consuming and compromise the simulation performance.

2.3.2 The Velocity Verlet algorithm

In practice, it is more convenient to implement the Verlet method through the Velocity Verlet algorithm, which allows the explicit incorpora-

tion of velocity in the equations of motion. With this approach, the position and velocity are calculated in the same time step so that $r(t - \delta t)$ does not need to be estimated in the first step of simulation. The standard scheme used in this algorithm can be summarized in three steps, as follow:

1. Calculate $r(t + \delta t) = r(t) + v(t)\delta t + \frac{1}{2}a(t)\delta t^2$;
2. Derive $a(t + \delta t)$ from the interaction potential employing $r(t + \delta t)$;
3. Then, calculate $v(t + \delta t) = v(t) + \frac{1}{2}(a(t) + a(t + \delta t))\delta t$.

The use of Velocity Verlet algorithm generates errors in the same order of that obtained with the basic Verlet method. However, the Velocity Verlet implementation is simpler because it is necessary to keep track of just one set of variables at every time step during the simulation [18].

2.3.3 Defining the integration time

In addition to the truncation of Taylor's expansion, another approximation of integration methods is the fact that both velocities and accelerations are considered to be constant during the integration step (δt). These errors are reduced by reducing δt. However, the smaller the integration step, the greater the computational cost for the system to satisfactorily explore the phase space. On the other hand, a very long integration step could generate instabilities resulting from the close contact between atoms, which often lead to the interruption of the simulation. With an appropriate integration step, these instabilities caused by the increase in the repulsive term of Lennard-Jones's potential are smoothed and phase space can be efficiently sampled at a practicable computational cost (Fig. 2.4).

The choice of the time step value depends on the system we are simulating. The time steps should be a few orders of magnitude smaller than that of the fastest vibrational frequency of the system under study which is often the covalent bond to an hydrogen atom (about 9-12 fs). Therefore time steps of 1 fs (femtosecond), or a few femtoseconds, are typically utilized in protein DM simulations.

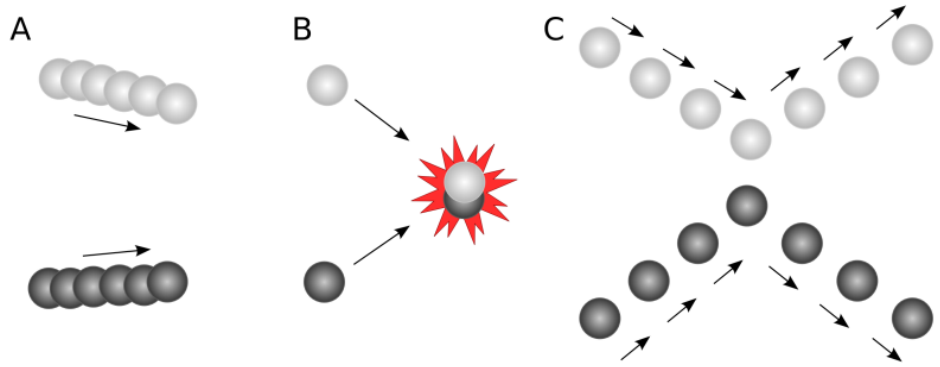

Figure 2.4: Effect of the choice of the integration step: (A) With an integration step very small, the phase space is slowly sampled. (B) One step of integration too large results in instabilities. (C) With an appropriate integration step, phase space is efficiently sampled and instabilities are smoothed. Figure adapted from Leach, 2001 [17].

2.4 Periodic boundary conditions

Biological systems such as proteins contain a very large number of atoms. In addition, the protein is generally simulated in solution. This obviously implies that the total number of atoms (protein+solvent) rapidly become too large to be explicitly simulated. To overcome this problem, a trick is normally applied, which falls under the name of periodic boundary conditions (PBC).

Within this framework, one starts by creating the simulation box (e.g., cubic or orthorhombic) containing the N atoms of the system including the solvent. Then, identical copies of such a box are arranged around the main one. That means that all the atoms in the system move exactly like those in the simulation box and also that the movement of an atom is not limited by the box walls. In such a way we can avoid the artefacts of the finiteness of the simulation box still keeping limited the number of atoms to be followed in time.

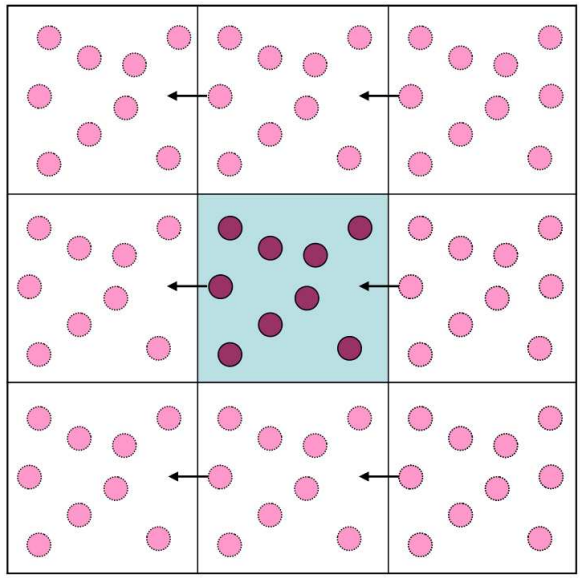

Figure 2.5: Simulation box with periodic boundary condition. Identical replicas of the simulation box are arranged around the main cell (central box) and the movement of a particle is not limited by the cell walls.

By the use of PBC, the number of interactions to be calculated for defining the forces acting on each atom of the simulation box, grows quadratically with the number of atoms. This number can rapidly become too large and some approximations are commonly introduced to reduce the computational cost. For LJ interactions, a spherical cut-off radius is generally introduced, whereas the Ewald summation is usually applied to compute the electrostatic (long-range) interactions. The main insight of this method is that, using a convergence function and a Fourier transform, a single slowly convergent sum can be transformed into two rapidly convergent sums. Due to their rapid convergence, these sums can be evaluated to high accuracy, thereby avoiding the computational artifacts introduced by more approximate cutoff-based methods.

2.5 Simulation protocol

A MD simulation is made of different steps which define the so-called protocol of simulation which is summerized in Fig. 2.6.

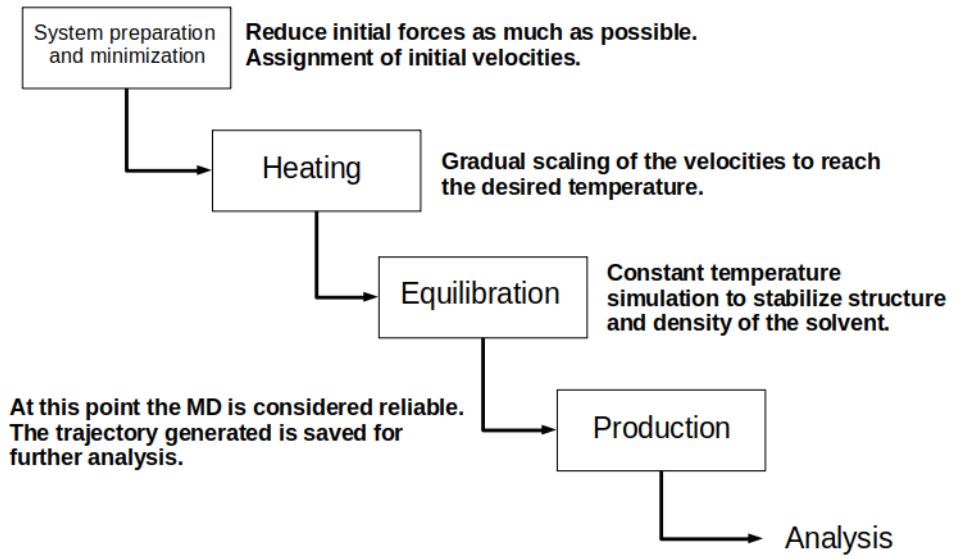

Figure 2.6: A generic protocol commonly employed for protein MD simulations.

In the previous sections we have already discussed the very first step in which we define the initial conditions for the atomic positions and velocities. We have also said that, by construction, an MD simulation should represent an NVE ensemble, where the number of atoms, the volume and the total energy is conserved while the temperature is a property that changes in time.

However, experiments are usually performed at constant temperature and volume (i.e., the canonical ensemble) or constant pressure and temperature (i.e., the isobaric-isothermal ensemble), so it is often desirable to simulate these conditions. During a simulation at constant energy, the temperature will be observed to fluctuate due to the spontaneous interconversion of the kinetic and potential components of the total energy. The instantaneous

temperature may be evaluated from the atomic velocities using:

$$\left\langle \sum_i \frac{1}{2} m_i v_i^2 \right\rangle = \frac{3}{2} NkT \tag{2.17}$$

where k is Boltzmann's constant, m_i and v_i are the mass and velocity of atom i, respectively, and N is the total number of atoms.

If we want to simulate a protein in its natural conditions, we have to control the temperature and this is achieved by the so-called heating step of the protocol. The purpose of the heating step is to reach the reference temperature T_{target}. The simplest way to do so is rescaling all velocities by an appropriate factor in order to match T_{target} even if nowadays much more robust and accurate algorithms can be used to control the temperature.

After the heating, the system is allowed to evolve until it reaches the equilibrium. This is tested by monitoring macroscopic properties of the system such as temperature and pressure, which should oscillate around average values and close to the real conditions (Fig. 2.7). Only when the equilibration has been reached, the real production step can start.

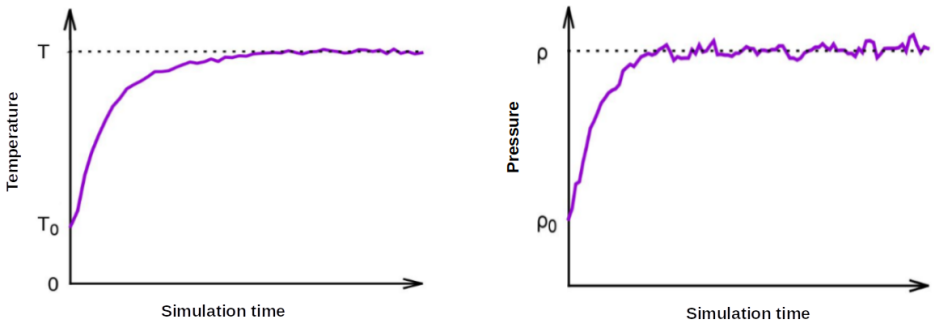

Figure 2.7: Heating step in NVT (left) and equilibration step in NPT (right). In both cases, the convergence to specific values of must be observed.

As a matter of fact, different combinations of ensembles can be employed during the heating, equilibration and production steps (see Fig. 2.8) [26]. However, the simulation conditions during the equilibration step must

be as closest as possible to the one that will be employed during the production step. Moreover, it is important to preserve the part of interest of the system from undergoing abrupt structural changes before the production step. Therefore, it is safer to impose harmonic constraints to the protein during the heating and early stages of the equilibration.

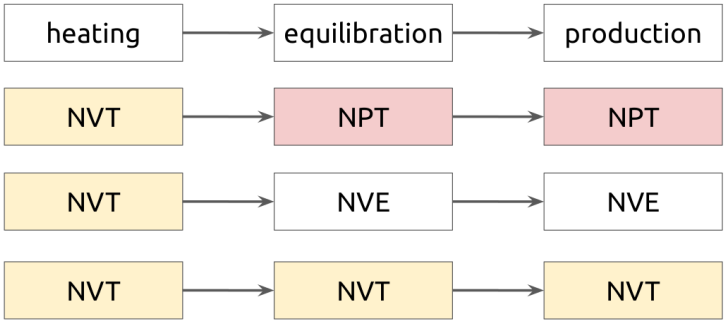

Figure 2.8: Set of the conserved properties for the heating, equilibration and production step of an MD simulation.

2.6 Trajectory analysis

A wide variety of analyses can be performed using the equilibrated trajectories obtained in the production step. These analyses allow the calculation of a series of structural, thermodynamic and dynamic properties [17]. The choice of the given analysis to be made depends on the system and on the objective of the study. Some examples can be found in the results section of Chapters 4 and 5.

Here, instead it is important to recall one of the most important structural analysis, namely the one based on the Root Mean Square Deviation (RMSD). RMSD is in fact a very popular tool to access the mobility of the protein structure along the trajectory. The RMSD allows to estimate the average atomic displacement of a system with respect to a reference structure [17]. For example, considering X_i and Y_i as positions of the same atom in two different configurations of a protein and N the total number of atoms,

we can obtain the RMSD value as follows:

$$RMSD = \sqrt{\frac{1}{M}\sum_{i=1}^{N} m_i(X_i - Y_i)^2} \tag{2.18}$$

where m_i is the mass of the atom and M is the total mass of the protein. One can, for example, calculate the RMSD for each of the various conformations of a protein generated during a MD simulation, using the starting structure as reference (usually the crystal structure is employed as a reference). In this manner, it is possible to obtain a mobility profile of the protein structure, which will allow to predict the occurrence of conformational changes or instabilities in its structure. These instabilities can then be confirmed by analyzing the trajectories using visualization programs. The RMSD can also provide an estimate of how long simulation time is needed for the study system to achieve the structural equilibrium (Fig. 2.9).

Again, there are many possibilities for analysis. The RMSD analysis can include the entire protein molecule or part of its sequence. In addition, it can be made for all atoms of the analyzed sequence or only for the protein backbone. The RMDS analysis can also be performed for each protein residue separately, in order to map the structural instabilities along the amino acid sequence. Moreover, the RMSD analysis can be applied not only to the protein. In fact, it might be useful to access the mobility of a ligand molecule with respect to the protein active-site, or even the internal mobility of the ligand itself.

In many cases, the analyses of the MD trajectories can be done directly using the simulation program, such as AMBER [21] or a molecular visualization program such as VMD [27]. Many data analysis softwares like *gnuplot*, or even programming languages like *python* and *R* are of great help in the presentation of the results.

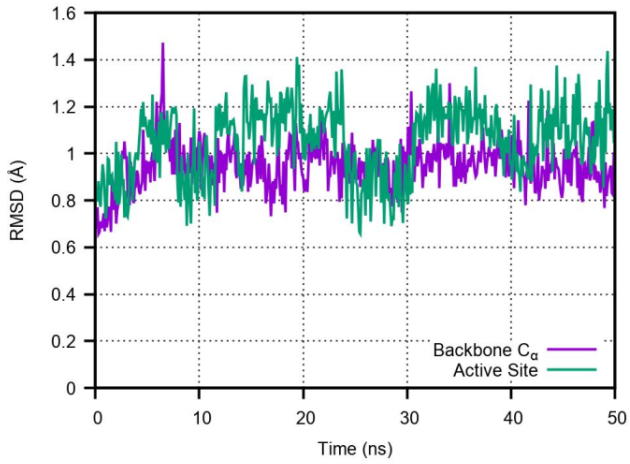

Figure 2.9: Example of RMSD analysis performed for lysozyme along the production step and computed with respect to the crystallographic structure. The analysis was done for the C_α atoms of protein backbone (purple line) and for the active-site residues (green line). The convergence of the RMSD value indicates that the structural equilibrium has been reached or that the protein molecule is exploring a local minimum on the potential energy surface.

Chapter 3

Multiscale Methods: Combining Quantum Chemistry and Molecular Mechanics

In the previous chapter we have presented a fully classic description that can be effectively used to predict properties and processes of complex systems. However, for specific cases, where the property or the process of interest is determined by electronic interactions, it is necessary to go beyond a fully classical description. This is the case, for example, of the study of chemical reactions, charge transfer processes or, as in the case of the present thesis, electronic excitations. Such a description of the electronic structure of atoms and molecules is only possible through methods based on Quantum Mechanics (QM). This requirement can be a real limitation in the case of biological systems, such as pigment-protein complexes which generally involve a large number of interacting atoms: in that case, the application of QM methods becomes impracticable.

Fortunately, an alternative multiscale strategy is possible. In general, multiscale methods allow to describe a system through the combination of

different models. In the present Thesis, such a combination involves Quantum Mechanics (QM) and classical Molecular Mechanics (MM), commonly referred to as QM/MM methods (Fig. 3.1) [28–33]. In conventional QM/MM approaches the system is divided in two parts, but multilayer schemes are also possible by introducing a high level and a low level QM region. Furthermore, more sophisticated approaches allow the implementation of dynamic boundaries between the different layers of the system, which will not be covered here.

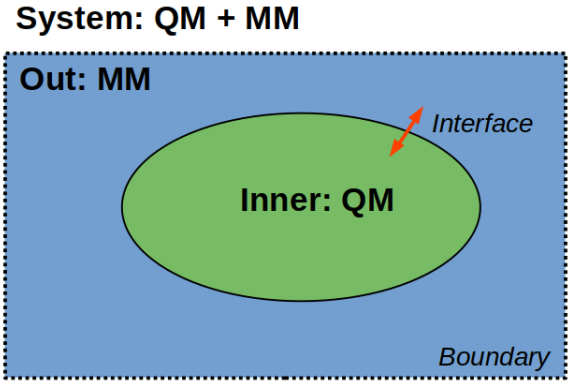

Figure 3.1: System partitioning employed in the conventional QM/MM approach. The system is divided into two subsystems. The inner subsystem is treated through quantum methods and the out subsystem is represented classically. The double arrows indicate the QM/MM interface and the dotted rectangle represent the system boundary.

3.1 The QM/MM energy

In general, the hybrid QM/MM energy contains three classes of interactions: interactions between atoms in the QM region, interactions between atoms in the MM region (described by empirical potentials, see Eq. 2.8 in Chapter 2) and interactions between QM and MM atoms. Due to the QM-MM interactions, the energy of the total system cannot be simply described as the sum of the energies obtained for the two subsystems (QM and MM

regions). Thus, associated terms are considered to describe the union between these subsystems and several approaches have been proposed to this end. These approaches can be classified into two categories: subtractive and additive coupling schemes.

3.1.1 Subtractive QM/MM Schemes

In the subtractive scheme, the energy of the system is obtained in three steps. First, the energy of the total system, consisting of both QM and MM regions, is evaluated at the MM level. The QM energy of the isolated QM subsystem is added in the second step. Third, the MM energy of the QM subsystem is computed and subtracted. The last step corrects for including the interactions within the QM subsystem twice. The subtractive form of the QM/MM potential can be expressed as:

$$E_{QM/MM} = E_{MM}(\mathbf{QM + MM}) + E_{QM}(\mathbf{QM}) - E_{MM}(\mathbf{QM}) \qquad (3.1)$$

where the subscribed acronyms indicate the kind of the potential (hybrid, quantum or classical), whereas the acronyms in bold and inside the parenthesis represent the system (QM + MM) and subsystems (QM or MM) adopted in the partitioning scheme. The most widely used subtractive QM/MM scheme is the ONIOM method, developed by Morokuma and coworkers [34, 35].

The main advantage of the subtractive QM/MM scheme is that no communication is required between the QM and the MM calculations. This clearly makes the implementation straightforward. However, compared to the more advanced schemes that are discussed below, there are also disadvantages. A clear disadvantage is that a force field is required for the QM subsystem, which may not always be available. A further important drawback of this method is the fact that the coupling between the QM and the MM subsystems is described at fully classical level. This can be particularly problematic for the modelling of electronic excitations of pigments in biological systems, since the involved electronic states are usually differently

influenced by the interactions with specific parts of the biological matrix. For an accurate description of such processes a more consistent treatment of the interactions between the electrons and their surrounding environment is needed.

3.1.2 Additive QM/MM Schemes

In additive schemes, the QM system is embedded within the larger MM system, and the potential energy for the whole system is a sum of MM terms, QM terms and QM/MM coupling terms:

$$E_{QM/MM} = E_{QM}(\mathbf{QM}) + E_{MM}(\mathbf{MM}) + E^{int.}_{QM/MM}(\mathbf{QM + MM}) \quad (3.2)$$

In this case, only the interactions within the MM region are described classically, $E_{MM}(\mathbf{MM})$. Different from the subtractive scheme, the interactions between the two subsystems, $E^{int.}_{QM/MM}(\mathbf{QM + MM})$, are treated explicitly.

The QM/MM coupling generally includes bonded, van der Waals, and electrostatic interactions between QM and MM atoms (Fig. 3.2):

$$E^{int.}_{QM/MM}(\mathbf{QM + MM}) = E^{B}_{QM/MM} + E^{vdW}_{QM/MM} + E^{E}_{QM/MM} \quad (3.3)$$

The first term, $E^{B}_{QM/MM}$, is present any time the QM and the MM subsystems are chemically linked. A common way to deal with the cut region is to introduce a link atom (L) at the QM/MM interface (See Fig. 3.1). The link atom replaces the atom directly bound to the QM region which will be treated at MM level. L is generally a hydrogen atom but any monovalent atom or group might be used to cap the QM subsystem [30]. The energy associated to the bond which will be replaced by L is treated classically through empirical force field parameters. To reduce at most possible artefacts due to the inclusion of the L atom, a proper choice of the bond to be cut is needed and usually such a bond should be far from the atoms involved in the investigated electronic processes.

The second term of 3.3, $E^{vdW}_{QM/MM}$, is related to the QM-MM van der

Waals interactions. In the most used formulations of QM/MM methods, these interactions are treated at a fully classical level, generally through a Lennard-Jones potential (See Eq. 2.6 in Chapter 2). This implies that both the MM part and the QM part of the system need specific parameters.

The third term of Eq. 3.3, $E^E_{QM/MM}$, is usually the most important one for the correct description of the environmental effects. The different methods for obtaining $E^E_{QM/MM}$ are commonly indicated as "embedding". The simplest formulation is the so-called *mechanical embedding* the QM subsystem is replaced by a set of point atomic charges which interact with MM region through the standard Coulomb interactions:

$$E^E_{QM/MM} = \sum_a^{N_{QM}} \sum_i^{N_{MM}} \frac{q_a q_b}{r_{ab}} \tag{3.4}$$

where the index a runs on the QM atoms and b on the MM charges.

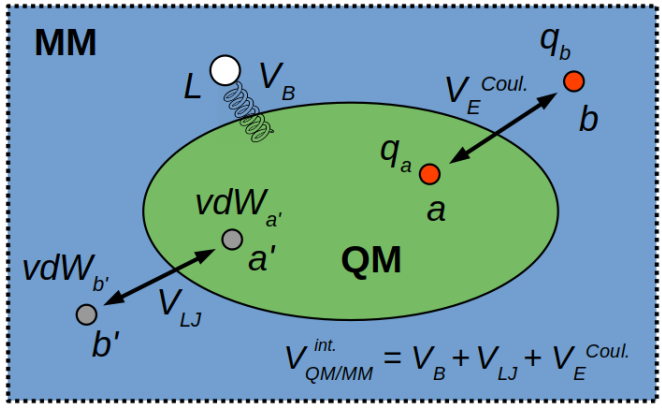

Figure 3.2: In the additive QM/MM scheme, QM/MM coupling term is obtained through the sum of V_B, V_{LJ} and V_E potentials. In the figure, a means the sites in the QM region and b means the sites in the MM region. The first potential (V_B), is defined as a classical harmonic potential that describes the covalent bond with the atom link (L). The other two are the Lennard-Jones (V_{LJ}) and electrostatic (V_E) potentials, respectively. In the mechanical embedding method V_E is defined as the classical Coulomb interaction between partial charges from QM and MM regions.

The *electrostatic embedding*, instead, provides a more accurate approach to calculate $E^E_{QM/MM}$. Here, the influence of partial charges from the MM region is introduced in the quantum Hamiltonian for the QM region. This is done by adding an one-electron interaction term to obtain the interaction energy between electrons of the QM region and the MM partial charges:

$$\hat{H}^E_{QM/MM} = \sum_M \sum_b \frac{Z_M q_b}{R_{Mb}} - \sum_b q_b \hat{V}^{QM}(r_b) \tag{3.5}$$

where M and b are, respectively, the number of QM nuclei and MM charges and $\hat{V}^Q M(r_b)$ is the electrostatic potential operator due to the electrons of the QM subsystem. In this manner it is possible to take into account the polarization effect of the MM environment on the QM electronic charge distribution (ρ_{QM}) (Fig. 3.3).

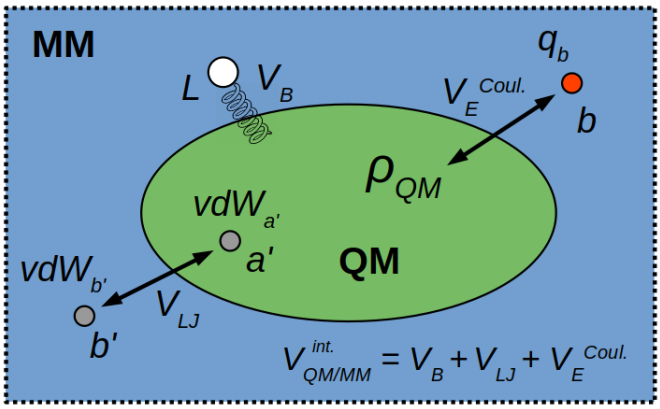

Figure 3.3: In the electrostatic embedding method V_E is defined as the Coulomb interaction between the electronic density calculated for the QM part (ρ_{QM}) and the partial MM charges. This model allows the QM part to be polarized by the MM part.

3.2 The polarizable formulation of QM/MM

The electrostatic embedding is the most common QM/MM formulation, however it still lacks one potentially important effect, namely the mutual polarization between the QM and the MM systems. The atomic charges used to represent the MM atoms are in fact fixed parameters and they do not change in response to the presence of the QM subsystem. To go beyond such a description, it is necessary to make the MM system polarizable. Various methods have been proposed so far to include mutual polarization effects [36]. A strategy is to introduce the effect of polarization using "ab initio" FFs which are fully built on first principles and require no fitted parameters [37, 38]. An alternative approach is to use polarizable FFs [39] based either on Drude oscillators [40–42], fluctuating charges [43–45] or induced point dipoles (IPD) [28, 46–49]. In this Thesis, we will focus on the IPD formulation of polarizable embedding QM/MM methods that from now on will be indicated as QM/MMpol.

Within this framework, each MM atom is endowed with a point charge and a polarizability. In more sophisticated polarizable FFs, such as AMOEBA [50], higher order multipoles are considered as well.

The polarizability allows the MM atoms to respond to an electric field, generated by the charges on the classical atoms and by the QM density, by creating an induced dipole. The QM/MMpol electrostatic and polarization interaction energy can be written as follows [48, 51]:

$$\begin{aligned} E^{\mathrm{QM/MMpol}} &= \frac{1}{2}\sum_b q_b V_b^{\mathrm{MM}} + \sum_b q_b V_b^{\mathrm{QM}} + \\ &+ \frac{1}{2}\sum_b \left(\alpha_b^{-1}\mu_b^2 + \sum_{b'\neq b} \mu_b \mathcal{T}_{bb'} \mu'_b \right) - \sum_b \mu_b (E_b^{\mathrm{MM}} + E_b^{\mathrm{QM}}) \end{aligned} \tag{3.6}$$

In Eq. 3.6, the first term is the interaction among the MM charges, where V_b^{MM} is the potential generated by all other charges at site b, the second term is the interaction between the MM charges and the QM density, V_b^{QM}

being the QM potential at site b. These two terms are common to standard electrostatic embedding QM/MM. The third and fourth terms instead are specific of the polarizable embedding models. They are the self-interaction of the induced dipoles, which can be understood as the work needed to induce the dipole itself, and the repulsion between all induced dipoles, respectively. In the fourth term, $\mathcal{T}_{bb'}$ is the effective dipole-field tensor, namely [50, 52]:

$$[\mathcal{T}_{bb'}] = \frac{f_e}{r_{bb'}^3}\mathbf{I} - \frac{3f_t}{r_{bb'}^5}\begin{bmatrix} x^2 & xy & xz \\ yx & y^2 & yx \\ zx & zy & z^2 \end{bmatrix} \tag{3.7}$$

where $\mathbf{I}$ is the unit matrix and f_e and f_t are distance dependent damping functions originally introduced by Thole [53] to avoid the so-called "polarization catastrophe", e.g. the divergence of the Coulomb interaction between two point dipoles when they get too close.

The energy in Eq. 3.6 is a variational functional [51] of the induced dipoles themselves. The minimum of the energy corresponds to the situation where the induced dipoles maximize the favorable interaction energy with the MM charges and QM density, while at the same time minimizing the repulsion among themselves. The equations for the induced dipoles are obtained by differentiating Eq. 3.6 and read:

$$\alpha_b^{-1}\mu_b + \sum_{b' \neq b} \mathcal{T}_{bb'}\mu'_b = E_b^{\mathrm{MM}} + E_b^{\mathrm{QM}} \tag{3.8}$$

Equations 3.6 and 3.8 are the main constituents of a polarizable QM/MM implementation. Assuming the QM subsystem is described at a Self-Consistent Field (SCF) level of theory, such as Hartree-Fock (HF) or Kohn-Sham (KS) Density Functional Theory (DFT), the polarizable QM/MM energy can be computed by assembling the energy in Eq. 3.6 and adding to the Fock or KS matrix the additional contributions, that are the derivatives of Eq. 3.6 with respect to the density matrix. In the atomic

orbitals (AO) basis:

$$V_{\mu\nu}^{\text{pol}} = \sum_b q_b \langle \chi_\mu(r)| \frac{1}{|r - r_b|} |\chi_\nu(r)\rangle - \sum_b \mu_b \langle \chi_\mu(r)| \frac{r - r_b}{|r - r_b|^3} |\chi_\nu(r)\rangle \quad (3.9)$$

The first term in Eq. 3.9 is independent of the QM charge density and can be summed with the one-electron Hamiltonian, as it is commonly done in EE QM/MM. The second term, on the other hand, represents at the same time the main advantage and the main complication introduced by a polarizable model. The induced dipoles, that are computed by solving the linear system 3.8, depend on the QM density, as the QM electric field appears in the right-hand side. Therefore, the Fock matrix contribution that stems from the induced dipoles is density dependent. This reflects the fact that the QM density and the induced dipoles are mutually polarized, i.e., the classical environment is able to adapt to changes in the QM density, which is the main strength of polarizable models. On the other hand, this mutual polarization introduces a non-linearity in the QM/MM interaction, which means that the QM and induced dipoles equations need to be solved iteratively.

The QM/MMpol approach is schematized in Fig. 3.4.

It is important to remember that the IPD formulation requires the use specific force fields for the MM part. The conventional force fields are not suitable in this case as, in their parameterization, they already take into account the average polarization of molecules [52, 54]. In addition, the inclusion of the QM/MM mutual polarization can increase the computational cost. However, the development of new formulations and algorithms, as well as the improvement of modern computers, have allowed to decrease the computational cost associated with the polarizable embedding approach.

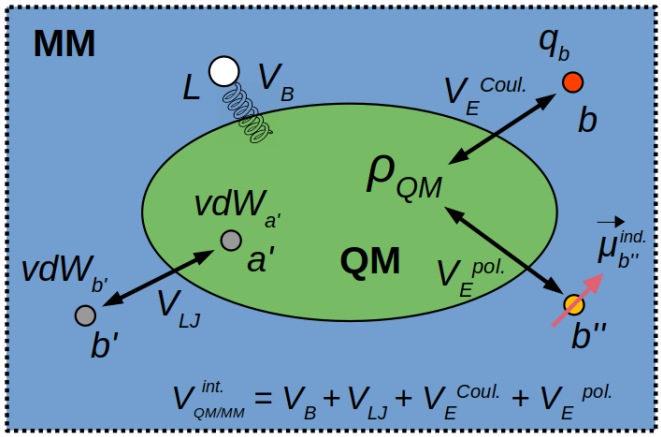

Figure 3.4: In the polarizable embedding approach, V_E has to terms: $V_E^{Coul.}$ and $V_E^{pol.}$. The first electrostatic term ($V_E^{Coul.}$) represents the Coulomb interactions between ρ_{QM} and the MM charges. The second electrostatic term ($V_E^{pol.}$) is introduced in V_E potential the to describe the interaction between ρ_{QM} and induced dipole moments in the MM part. Here, the QM and MM regions can polarized each other mutually.

3.3 QM/MM applications in computational spectroscopy

Spectroscopy in the UV-visible spectral region (UV-Vis) is among the most important experimental techniques employed in chemical and biochemical laboratories. It is possible to have important insights into a given biological system through the analysis of the absorption or fluorescence spectra of its chromophoric regions. However, the interpretation of the spectra requires knowledge about the structure and dynamics of the studied system.

The computational modelling can provide valuable information that allow the validation or even the reinterpretation of experimental data. In such theoretical approaches, the correct description of the environment is crucial since the excitation of chromophoric molecules can be strongly influenced by the surrounding environment [55, 56].

In this Thesis we studied chromophoric molecules in protein environments, which can have a very dynamic nature. To take this dynamics into account, a common approach is to combine classical MD trajectories with QM/MM vertical excitation energy calculations. Within this framework, the QM region is usually described by the time-dependent density-functional theory (TDDFT) method, due to the favorable accuracy/cost ratio. The MM part, in turn, is described by conventional or polarizable force fields.

In the next sections of this chapter, we briefly discuss how these multiscale methods can be used to obtain UV-Vis spectra. For the first case, we will consider both the case of a single chromophore and the one of a multichromophoric aggregate with coupled excitations.

3.3.1 Single chromophore systems

From a computational point of view, the most common approach for the simulation of absorption spectra, is based on the computation of the excitation energies of the chromophore in its ground-state equilibrium geometry. These vertical excitation energies are then compared to the maxima of the corresponding experimental absorption bands. Despite its simplicity and computationally effectiveness, this approach can easily become inadequate due to the large approximations it involves. An important drawback is the lack of the coupling of electronic transitions to nuclear degrees of freedom. This coupling can in fact generates a characteristic vibronic progression, and can result in an asymmetrical lineshape even when the vibronic structure is hidden in homogeneous and inhomogeneous broadening. For molecules in condensed phase, peaks positions and intensities in the absorption spectra are also affected by the interactions with the inhomogeneous environment, owing to the many different configurations that it adopts around the molecule. The resulting spectra are thus broadened by the so-called static disorder.

Two main approaches are nowadays widely used to go beyond the single vertical excitation approximation [33, 36, 57].

The first one is the “ensemble” based approach, where the many environment-molecule configurations are explicitly sampled and then used as

geometries for excitation energy calculations. These configurations are often sampled along a molecular dynamics simulation, and excitation energies are computed with QM/MM methods generally within the electrostatic embedding formulation. This approach is particularly suited for chromophores embedded in complex and dynamic environments such a protein matrix. When classical MD are used for sampling the configurational space, the performance of the approach will strongly depends one the accuracy of MM force field used to generate the MD trajectory. Most FFs are generally well suited to sample the structural fluctuations of the biomatrix, but they are not accurate enough to properly describe the internal motions of the chromophore. A possible solution is to precede the excitation energy calculations with a geometry optimization of the chromophore within a frozen environment.

The second approach explicitly considers the vibrational wave functions of the ground and excited electronic states, which lead to the vibronic fine structure of the spectrum. In this context, within the Frank-Condon approximation, the intensity of the vibronic transitions can be computed from the overlap of the ground and excited-state vibrational nuclear wave functions, obtained from the shape of the ground state and excited state potential energy surfaces. Contrary to the ensemble approach, this approach is usually suited for rigid molecules, whose absorption spectra are dominated by vibronic features. A common drawback of this vibronic approach is the lack of an explicit inclusion of environment effects in the determination of the broadening. Such an effect can be artificially added through a Gaussian broadening which simulates the effect of a static disorder.

3.3.2 Multichromophoric systems

We define a multichromophoric system as a system constituted by two or more interacting chromophoric groups. The different chromophoric groups can be located on different molecules or within the same covalent structure. Aggregates of chlorophylls in pigment-protein complexes are an example of the first case, while nucleobases of a DNA strands are an example of the second case.

To understand what happens in such a kind of systems, let's consider a model system made of two identical molecules characterized by a single electronic excitation (see Fig.3.5). Let's also assume that the two molecules are placed at an intermolecular distance large enough not to allow significant changes in the ground state of each molecule but short enough to make their excitations coupled.

As a result of such a coupling, the dimeric system will generate two new excitations (called *excitons*) at two different energies with respect to the one characterizing each single molecule. The energy gap between the two new excitations (also called *exciton splitting*) depends on the strength of the coupling between the electronic transitions in the two molecules. In the absorption spectrum of the coupled system, there is the formation of two peaks, one at lower and other at higher energy with respect to the monomer absorption. The relative intensity of the bands depends on the relative position and orientations of the two interacting molecules.

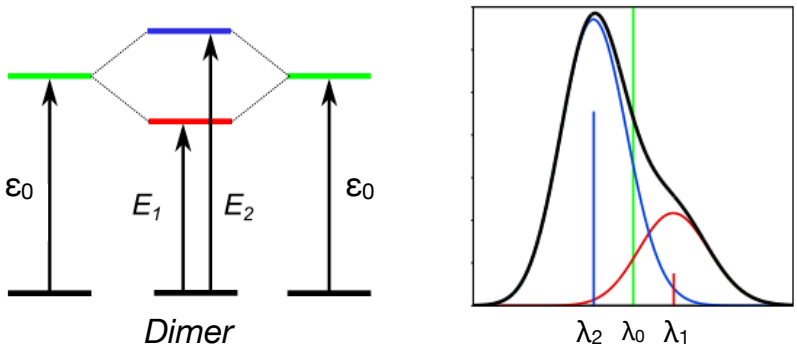

Figure 3.5: Left panel: Energy level diagram of the coupled dimer system (degeneracy is assumed for the monomers). ϵ_0 is the excitation energy of the monomer (site energies) whereas E_1 and E_2 are the excitation energies in the dimer. Right panel: Absorption spectrum: vertical green line represents the excitation energy of the monomer (λ_0); red and blue signals are referred to the two excitations of the dimer; black curve is the final spectra of the dimer.

The multichromophoric system may have a very large dimension in term of the total number of atoms. An accurate description of the electronic

excitations of such a large system is still challenging, even for TDDFT, and approximated approaches have to be introduced.

These models are called excitonic strategies or excitonic models. They allow the reconstruction of the system excitonic Hamiltonian from calculations of vertical energies of each single chromophore (called *site energy*) and excitonic couplings between each pair of excitations.

In the standard formulation of the model, the electronic excited states of the entire system are expanded on a basis of diabatic states $|i\rangle$, which represent an excitation localized on a single chromophoric unit:

$$|\psi_K\rangle = \sum_i c_i^K |i\rangle \tag{3.10}$$

It is commonly assumed that the states $|i\rangle$ are Hartree products of the excited state of one chromophore with the ground states of all other chromophores ($|0, \ldots, 1, 0, \ldots\rangle$). Physically, this corresponds to neglecting the GS exchange interactions between different chromophores. Consistently, the global ground state $|g\rangle$ is represented as the Hartree product of all chromophores' ground states.

The electronic Hamiltonian of the entire system is then written on the basis of states $|i\rangle$, namely:

$$\hat{\mathcal{H}}_{el} = E_g |g\rangle \langle g| + \sum_i \mathcal{E}_i |i\rangle \langle i| + \sum_{i \neq j} V_{ij} |i\rangle \langle j| \tag{3.11}$$

where $\mathcal{E}_i$ are the excitation energies of the isolated chromophores (*site* energies), and V_{ij} are the electronic couplings between the different chromophores' states. E_g is the energy of the ground state, which is taken as the zero of the energy scale whenever one is only interested in the excitation energies. The electronic energies and coefficients of the expansion in Eq. (3.10) are obtained by diagonalization of the Hamiltonian matrix. Here, for the sake of simplicity, we have assumed that each chromophore i contributes with only one excited state to the diabatic basis set, but the essential details of the model do not change when considering more than one excited state per

chromophore.

The power of the exciton model lies in the possibility of computing its parameters $\mathcal{E}_i$ and V_{ij} only from calculations of single-chromophore properties. In particular, the electronic coupling V_{ij} between two excited states localized on different chromophores can be very well approximated as the following Coulomb interaction:

$$V_{ij} = \int dr \int dr' \frac{\rho_i^T(r)\rho_j^T(r')}{|r - r'|} \tag{3.12}$$

where $\rho_i^T(r)$ ($\rho_j^T(r')$) is the transition density corresponding to the excitation at site i (j). The integral in Eq. (3.12) can be evaluated numerically, or analytically using a basis set expansion.

A common simplification of this interaction is the point-dipole approximation (PDA). Within this framework the transition densities are approximated in terms of transition dipoles and the coupling reduces to:

$$V_{ij}^{\mathrm{PDA}} = \frac{\mu_i \cdot \mu_j R_{ij}^2 - 3(\mu_i \cdot \mathbf{R}_{ij})(\mu_j \cdot \mathbf{R}_{ij})}{R_{ij}^5} \tag{3.13}$$

where $\mu_{i/j}$ is the transition dipole for transition i or j, $\mathbf{R}_{ij}$ is the vector distance between the centers of transitions i and j, with magnitude R_{ij}. It is well known that the point-dipole approximation breaks down at close interchromophoric separation, namely, when R_{ij} is smaller than the dimensions of the chromophores.

The standard exciton model (Eq. (3.10)) only considers states where the excitation is localized on single chromophores. As such, it neglects all the charge transfer (CT) states $|i^+j^-\rangle$ where the hole and electron reside on two different chromophores i and j. Generalizations of the Hamiltonian (3.11) to include CT configurations are however possible as it will be reported in Chapter 4.

In sections 3.1-3.2 we have summarized the most common approaches used in quantum chemistry to include the effects of the environment. The same approaches can be extended to multichromophoric systems and used

to calculate the excitonic parameters (site energies and couplings) needed to build the excitonic Hamiltonian of the embedded system.[16]

The electronic couplings are affected by the environment through two mechanisms. Firstly, the environment can change both the geometrical and the electronic structure of the chromophores and modify their transition properties, i.e. transition dipoles and transition densities. These changes will be "implicitly" reflected in a change of the Coulomb coupling (Eq. 3.12) which will be generally enhanced due to the electrostatic effect of the environment. The second effect is due to the polarizable environment, which mediates the interaction among chromophores' excitations. The resulting "explicit" effect generally reduces the magnitude of the direct (Coulomb) coupling. For this reason it is common to say that the coupling is "screened" by the environment.

Both electrostatic and polarizable embedding QM/MM formulations give the first implicit effect while only polarizable models include the explicit effect. Within the induced dipole formulation of QM/MM, an atomic polarizability is added to the fixed charge to describe each atom of the MM environment and the explicit term to the coupling becomes [48]:

$$V_{ij}^{\mathrm{MMPol}} = -\sum_{b} \left[\int dr \rho_i^T(r) \frac{(r_b - r)}{|r_b - r|^3} \right] \mu_b(\rho_j^T) \tag{3.14}$$

where the transition density ρ_j^T induces a response in the environment which is represented by the induced dipoles μ_b.

Chapter 4

Case Study I: LH2 Complexes from Purple Bacteria

This Chapter is focused on the light-harvesting complexes 2 (LH2) from purple bacteria and their mechanism of adaptation to light. In particular, we studied LH2 complexes expressed in low-light (LL) and high-light (HL) conditions in order to explain the spectral differences between then. The study was based on the combination of classical molecular dynamics (MD) of each complex in a lipid membrane and excitonic calculations by employing multiscale quantum mechanics/molecular mechanics (QM/MM) approach including a polarizable embedding. From the comparative analysis, it comes out that the mechanisms that govern the adaptation of the complex to different light conditions use the different H-bonding interactions around the bacteriochlorophyll a (BChl) pigments to dynamically control both internal and inter-pigment degrees of freedom. While the former have a large effect on the site energies, the latter significantly change the electronic couplings, but only the combination of the two effects can fully reproduce the tuning of the final excitons and explain the observed spectroscopic differences.

The results presented here are based on Ref. [58], the graphical abstract

for this study is shown in Fig. 4.1. This work was published on the *Chemical Science* journal and can be accessed freely through the link: `https://doi.org/10.1039/c9sc02886b`.

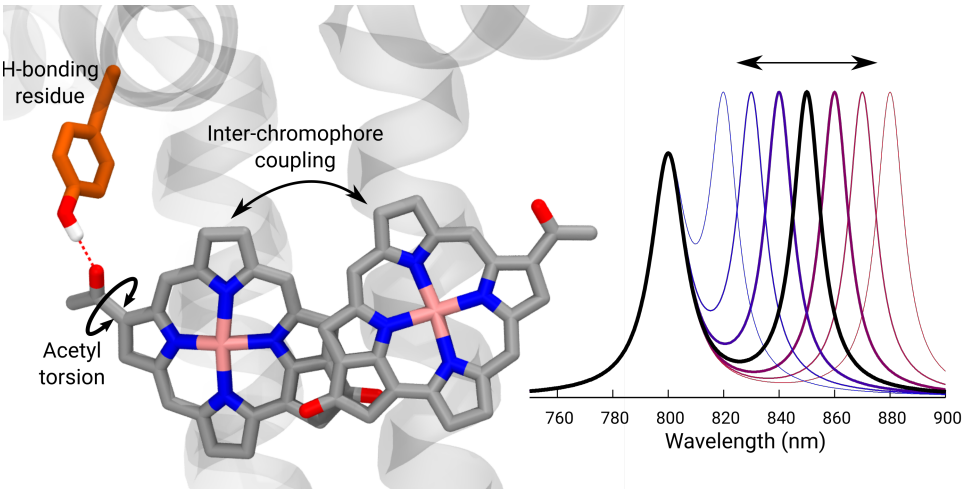

Figure 4.1: Graphical abstract of the paper that is the basis of this Chapter (see Ref. [58]). The obtained results demonstrate that the spectral tuning of LH2 antenna complexes arises from H-bonding, the torsion of bacteriochlorophyll a (BChl) acetyl group, and the inter-chromophore couplings. Copyright: `https://doi.org/10.1039/c9sc02886b`.

4.1 Introduction

Purple bacteria are photosynthetic organisms with a unique light-harvesting (LH) apparatus, which involves two types of pigment-protein complexes, the so called LH2 and LH1 [8, 59–61]. In the photosynthetic membranes, the LH1 complexes surround the reaction center (RC), while the LH2 complexes are arranged more peripherally around the LH1-RC complex; the ratio of the LH2 complexes to LH1-RC complexes is regulated by the incident light intensity [62, 63]. Moreover, in some species of purple bacteria, an additional adaptation to changing light conditions is possible thanks to the modular composition of the LH complexes. In LH2 complexes, for example, the units consist of pairs of hydrophobic, low-molecular-weight polypeptides, called α and β: by expressing different genes that encode such polypeptides, different LH2 complexes are obtained showing different spectroscopic properties in the NIR spectral range [64–66].

The α and β chains, in fact, noncovalently bind a number of Bacteriochlorophylls a (BChl) which present a bright excitation (the so-called Q_y excitation) at around 790 nm when isolated, and in the range 800-890 nm when arranged in the multichromophoric aggregates present in the different forms of LH2 and LH1 complexes. These shifts to longer wavelengths are due to excitonic interactions among the Q_y excitations of the BChls which, in all complexes, are arranged in circular, elliptical, or horseshoe aggregates (from now called rings) presenting different dimensions and symmetries.

In particular, one of the species showing different forms of LH2 at different light conditions, the *Rps. acidophila*, is characterized by a nonameric circular symmetry which corresponds to two absorption bands centered at 800 and 850 nm when the cells are grown under high light (HL) conditions. The same organism, when grown under low light (LL) conditions, presents a different form of LH2, whose longer wavelength band is blue-shifted at 820 nm while the other band remains unshifted. Moreover, the shifted band is significantly broadened and less intense. This different spectroscopic form of LH2 is also known in the literature as B800-B820 or LH3. The available

crystallographic data clearly show that the HL LH2 (from now on HL-LH2) [67, 68] and its LL analog (from now on LL-LH2) [69] present exactly the same multichromophoric structure with a nonameric symmetric repetition of the basic building block, the $\alpha\beta$ unit which accommodates three BChl pigments (and one carotenoid molecule). As a result, two circular rings, made of 9 and 18 BChls respectively, are obtained (Fig. 4.2a-b). The only significant differences are due to the different local environment around the BChls of the 18-meric ring. In particular, the protein residues adjacent to the BChls in HL-LH2, Tyr44 and Trp45, in LL-LH2 are replaced by phenylalanine and leucine, respectively. While the former ones form hydrogen bonds with the acetyl moiety of the BChl, the latter ones do not. It has also to be noted that in LL-LH2 a new hydrogen bond is however activated from the acetyl group to another tyrosine (Tyr41). This hydrogen bond is not present in the HL-LH2 complex where the residue at the equivalent position is phenylalanine. These changes in the H-bonding patterns between the two complexes are shown in Figs. 4.2d-f and A3.

Other studies have showed that another purple bacteria species (*Rps. palustris*), when grown in LL conditions, can express different forms of LH2 with multiple gene pairs encoding the LH2 apoproteins [8, 70–72]. It was suggested that LL LH2 complexes of *Rps. palustris* have a heterogeneous peptide composition [71], whereas Papiz and co-workers suggested that these complexes mainly contain the *puc*$\mathrm{BA_d}$ gene pair sequence [70]. Recently, a quadruple deletion mutant, containing only the *puc*$\mathrm{BA_d}$ gene pair, was characterized. The LH2 complex produced by this mutant, called PucD, presents an absorption spectrum with a single band and a strong circular dichroism signal in the 800 nm region, similar to the wild type LL LH2 of *Rps. palustris* [72]. Unfortunately, such studies were not able to generate a high resolution structural model and different structures have been proposed. In the study by Papiz and co-workers, [70] the electron density at 7.5 Å resolution of the LL grown LH2 complex of *Rps. palustris* suggests a larger pigment density than what found in the LH2 complexes from *Rps. acidophila*, and an octameric structure with exact 8-fold rotational symmetry. On the

contrary, the PucD mutant presents a similar pigment composition to the LH2 complexes of *Rps. acidophila* and a similar nonameric structure, the main differences being once more in the local environment of the BChls [72].

Many experimental and theoretical studies have tried to identify the origin of the spectroscopic specificities of the different complexes but none of them has yet given a conclusive answer. From the available structural data some hypotheses have been proposed; which can be summarized as follows:

- Hydrogen bonding: hydrogen bonds to the acetyl group conjugated with the BChl macrocycle ring is expected to cause a red-shift in the Q_y excitation relative to the non-hydrogen bonded case. Hence, the loss of the hydrogen bond in the BChls of LL-LH2 may be the origin of the overall blue-shift [73, 74].

- Rotation of the acetyl group: when the acetyl group is in the plane of the BChl macrocycle ring, it adds one more double bond to the conjugated system and the Q_y excitation is red-shifted. Due to the different H-bonding, the acetyl groups are rotated further out of the plane in the BChls of LL-LH2 and this may be the origin of the overall blue-shift [69, 75, 76].

- Deformation of the structure of the macrocycle: none of the macrocycle rings of the BChl molecules in LH2 are completely planar in the available crystal structures. The BChls of the 9-meric ring are slightly domed, while the ones in the 18-meric ring are differently distorted (see Fig. 4.2c where the bowed βBChl is shown). Deformation of the macrocycle ring has been suggested to give large shifts of the Q_y absorption band [77].

- Electrostatic and polarization effects: the residues around the BChls can produce large shifts in absorption, with the magnitude and direction of the shift depending upon the electric fields and the polarization

effects due to their specific nature and the disposition in the different complexes [78, 79].

- Charge transfer (CT) effects: the coupling between higher energy CT states between BChls in the 18-meric ring and the locally excited Q_y states influences the exciton structure of the complex [80, 81], lowering the energy of the first bright exciton state, which gives rise to the longest wavelength band in the absorption spectrum. Our recent computational study suggested that the structural differences between the two complexes correspond to a reduction of the CT–Q_y coupling values in LL-LH2 with respect to HL-LH2 [82].

In the present systematic study we have tested each of the proposed hypothesis by analyzing the LL and HL LH2 complexes from *Rps. acidophila* as well as an artificial complex analogous to the PucD complex from *Rps. palustris* which has been predicted by homology modeling (HM-PucD). The latter, despite having the polypeptide sequences of PucD, was built based on the same nonameric structure of LH2 from *Rps. acidophila*. The resulting complex is characterized by a further reduction of H-bonds with respect to LL-LH2, as the Tyr41, which is H-bonded to αBChl in LL-LH2, is now replaced by a Valine (see Fig. 4.2d-f for a comparison of the different hydrogen-bonding pattern in the three LH2 complexes). For the three complexes we have performed classical molecular dynamics (MD) in a lipid membrane to generate equilibrated systems and include the effects of thermal fluctuations. The structures coming from the MD trajectories have then been used to construct the excitonic Hamiltonian to generate the exciton states and the final absorption spectra. All excitonic calculations have used a hybrid quantum mechanics/molecular mechanics approach including a polarizable embedding (QM/MMPol).

The results reproduce the blue-shift of the excitons of the 18-meric ring going from HL to LL-LH2, in agreement with spectroscopic data while they fail to give the expected further shift from LL-LH2 to PucD. By identifying and quantifying the reasons of the successes and the limits of the adopted

computational strategy, an explanation of the mechanisms that govern the adaptation to different light conditions is suggested in terms of a delicate interplay between the H-bonding network around the 18-meric ring and "intra" and "inter" pigment mobility. This explanation opens a new scenario for a structure-based mutagenesis strategy which has the goal to control the relative energy of the excitons.

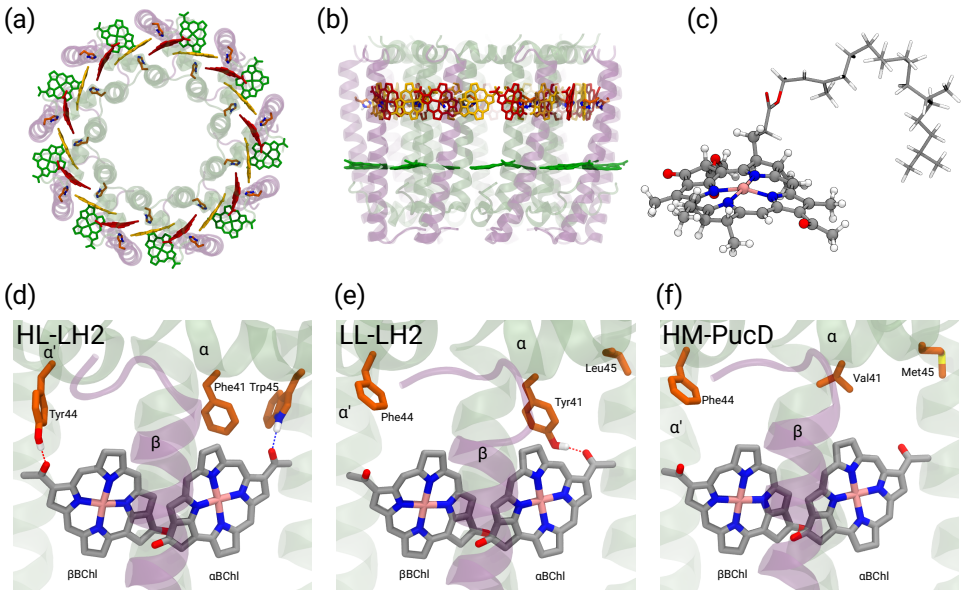

Figure 4.2: (a) and (b) Structure of LH2 complex from *Rps. acidophila*. The LH2 complexes are circular structures formed by nine dimeric units, the α and β chains, binding one catotenoid molecule (not shown) and three bacteriochlorophyll a (BChl) molecules (α and βBChl, and B800). (c) Representation of the βBChl extracted from the LL-LH2 crystal structure. The atoms represented as spheres were included in the QM part in the multiscale calculations (see Methods), whereas the other atoms were described with the MMPol method. (d), (e) and (f) Schematic representation of the local protein environment for α and βBChl within the same $\alpha\beta$ unit inside the three complexes: HL-LH2, LL-LH2 and HM-PucD, respectively. Copyright: `https://doi.org/10.1039/c9sc02886b`.

4.2 Computational details

4.2.1 Structures

The initial structure for the *Rps. acidophila* LH2 complexes were obtained from the Protein Data Bank (PDB): entries 1NKZ [68] for the HL-LH2 and 1IJD [69] for the LL-LH2.

For the PucD it was necessary to employ a structural prediction process based on a homology modeling (HM) approach [83]. The sequences for α and β chains of the PucD complex were obtained from GenPept at National Center for Biotechnology Information (NCBI): entries WP_011158559 and WP_011158560, respectively. Due to the larger overall sequence identity with LL-LH2 (70% and 64% for chains α and β, respectively) than with HL-LH2 (57% and 64%) we used the crystal structure of LL-LH2 as template. Firstly, we produced a 2D alignment between each target sequence and its respective template, and then generated 10 homology models for each chain (α and β) of PucD complex. Next, the models were evaluated based on both their stereochemical quality and structural overlap with the crystal structure; the best model for each chain was selected and used to construct the PucD $\alpha\beta$-apoproteins. To avoid disrupting protein-cofactor interactions, the vicinity of conserved interacting residues was built by inserting mutant fragments taken from the predicted model into the crystal structure of LL-LH2. The same procedure was done for the additional residues present in the LL-PucD protein sequence. The N-terminal end of chain α and C-terminal end of chain β were supposed to be flexible and non-structured; thus, we adjusted some torsions in order to prevent close contacts between monomers in the final nonameric complex. Finally, the α and β chains were assembled with the cofactors (BChls and Cars) from LL-LH2 complex and the C9 symmetry was applied to the system to obtain the entire PucD structure.

4.2.2 Molecular dynamics

To reproduce the membrane environment, we performed molecular dynamics simulations of the three different forms of LH2 in 1,2-dioleoyl-*sn*-glycero-3-phosphocholine (DOPC) bilayers. Using the input-generator tool of CHARMM-GUI server [84], we built a DOPC membrane with about 800 lipid molecules in total, solvated with a water layer of 40 Å on both sides and at 0.1 M NaCl. The membrane was then preequilibrated by applying the same procedure used by Dickson *et al* [85]. The next step was to insert the complexes into the preequilibrated membrane, and to eliminate the close contacts by deleting all lipid molecules up to 1.0 Å from the external part of the complex (the lipid core was preserved). Then, by using the *tleap* module of AmberTools [21], we added extra Na^+ ions so as to achieve system charge neutrality. The final systems contained approximately 290 000 atoms with the simulation box of dimensions approximately 171 Å × 165 Å × 108 Å. The minimization was done by first minimizing the lipid core molecules, next all the lipid tails and finally the whole system.

For the MD simulations we used a protocol very similar to that employed in our previous work [86]. Briefly, we first performed a heating from 0 to 100 K (5 ps in the NVT ensemble) constraining all the system but not the lipid tails with a harmonic potential (10.0 kcal mol^{-1} $Å^{-1}$) and then from 100 to 300 K (100 ps in the NPT ensemble) constraining just the protein and cofactors. Next, a 10 ns NTP equilibration at 300 K was performed initially applying the same constraining on the protein and cofactors but releasing the harmonic force constant by 1 kcal mol^{-1} $Å^{-1}$ at each 1 ns. Finally, a production step of 200 ns at 300 K in the NTP ensemble was performed for each system. In all MD simulations, the time step was set to 2 fs. For system temperature and pressure control we employed a Langevin thermostat and an anisotropic barostat, both implemented in the Amber16. The particle-mesh Ewald algorithm [87] was used to describe the long-range electrostatic interactions. The MD analysis was performed by using both the *cpptraj* [88] module of AmberTools and locally developed tools. For trajectories visual-

ization we employed the Visual Molecular Dynamics (VMD) software [89].

Both minimization and MD simulations were performed using the Amber16 program employing the *ff14SB* [90] force field for protein and *lipid14* [85] for lipids. The parameters for the BChls were taken from the literature [91] and for carotenoids we applied a DFT-based strategy developed in our group, previously described by Prandi *et al* [92]. Water molecules were described by the TIP3P model.

4.2.3 Excitonic states

We described the excitons of the multichromophoric system as linear combinations of locally excited (LE) states and charge-transfer (CT) states. The LE state of each BChl, is the lowest bright excitation (Q_y).

The BChls in the 18-meric ring are usually labeled as either α or β, depending on the noncovalent binding to the α or β chain, respectively and they alternate in the ring structure. For the CT states we considered all the possible charge transfers between BChls in adjacent $\alpha\beta$ pairs of the ring, so that, for every pair, there are two possible transfers: $\alpha \rightarrow \beta$ or $\beta \rightarrow \alpha$. By combining the local excitations and the CT states, the following Hamiltonian can be obtained:

$$\begin{aligned}\hat{H} = \sum_{i}^{27} \epsilon_i \left|i\right\rangle \left\langle i\right| + \sum_{ij}^{27} V_{ij} \left|i\right\rangle \left\langle j\right| + \sum_{m}^{36} \epsilon_m^{\mathrm{CT}} \left|m\right\rangle \left\langle m\right| \; + \\ + \sum_{i}^{18} \sum_{m}^{36} \left(V_{im}^{\mathrm{CT}} \left|i\right\rangle \left\langle m\right| + h.c. \right)\end{aligned} \tag{4.1}$$

where the indices i and j run on the locally excited states, ϵ_i is the excitation energy of the i–th BChl and V_{ij} is the electronic coupling between the i–th and j–th excitations. The index m runs on the CT states, ϵ_m^{CT} is the energy of the m–th CT state and V_{im}^{CT} is the coupling between the i–th locally excited state and m–th CT state. The excitonic analysis was performed by using the EXAT program [93].

4.2.4 Site energies and couplings

From the MD trajectories, we extracted 50 equally spaced frames from 100 to 200 ns, and for each of them, we computed the site energy (Q_y state) for the 27 BChls and the corresponding electronic couplings using the polarizable embedding QM/MM method (from now on QM/MMpol) described in Chapter 3. Within this framework, the BChls of interest are treated with a QM method and the rest of the environment as a set of classical point charges and atomic polarizabilities. As such, this approach allows the QM part and the classical part to mutually polarize. For the QM part, we used time dependent density functional theory (TD-DFT) at the B3LYP/6-31+G(d) level [94, 95]; for the classical part we used the Wang force field [52]. We excluded the phytyl tail of the BChl from the QM part as it does not influence the Q_y transition but we account for its electrostatic and polarization effects describing its atoms as polarizable MM sites (the QM/MM cut is shown in Fig. 4.2c).

We computed the electronic couplings using a method based on the transition densities of the interacting BChls [48, 96, 97]. That is, for each pair of BChls we computed the Coulomb interaction as:

$$V_{ij} = \int\int \frac{\rho_i^{\mathrm{tr}}(\vec{r})\rho_j^{\mathrm{tr}}(\vec{r}')}{|\vec{r}-\vec{r}'|} d^3r' d^3r - \sum_k \left(\int \rho_i^{\mathrm{tr}}(\vec{r}) \frac{\vec{r}_k - \vec{r}}{|\vec{r}_k - \vec{r}|^3} d^3r \right) \vec{\mu}_k^{\mathrm{ind}}(\rho_j^{\mathrm{tr}}) \qquad (4.2)$$

where ρ_i^{tr} and ρ_j^{tr} are the transition density of the i–th and j–th interacting pigments. The sum in the second term runs on the polarizable sites located at the $\vec{r}_k$ positions, where $\vec{\mu}_k^{\mathrm{ind}}(\rho_j^{\mathrm{tr}})$ is the dipole induced by the transition of the j–th pigment. The first and second terms of Eq. 4.2 are respectively the bare Coulomb coupling between the pigments (V_{Coul}) and the effect of the environment (V_{MMPol}).

4.2.5 Coupling to charge-transfer states

We computed the effect of CT states with the method proposed in a previous work [82, 98]. From 10 of the 50 MD frames, we selected all the possible pairs of adjacent BChls, divided among the intra-chain and inter-chain

dimers, for a total of 90 calculations for each dimer. The phytyl tail of each BChl was excluded from the calculation. All couplings within each dimer are computed with the multi-FED-FCD diabatization scheme [82], which combines the Fragment Excitation Difference (FED) [99] and Fragment Charge Difference (FCD) [100, 101] methods. Using appropriate additional operators, the adiabatic Hamiltonian of the dimer is transformed into a diabatic basis, in order to subsequently extract the LE-LE and LE-CT couplings from the diabatic Hamiltonian matrix. The CT energies are corrected a posteriori with the corrected Linear Response (cLR) formalism to account for the state-specific response of the environment, which is needed when a large density redistribution upon excitation occurs.

As here we have to compute both LE and CT states, the dimeric calculations were performed with a tuned (ω= 0.195) long-range corrected BLYP functional [102] and the 6-31G(d) basis set. The robustness of this approach with respect to basis sets, functionals, polarization and charge cutoffs has been already validated in our previous work [98]. The energy of the LC-BLYP/6-31G(d) results was decreased by 1008 cm^{-1} to match the locally excited states found with B3LYP/6-31+G(d). For the HM-PucD we used the LL-LH2 CT results as we expect them to be very similar.

4.2.6 Geometry optimizations

In addition to MD simulations we have also performed some tests on the crystal structures. In these cases, due to the limitation in the resolution of the crystal structures, the geometry of the BChls was optimized with a QM/MM method within the ONIOM scheme [34, 35]. For the QM BChl, we used the B3LYP/6-31G(d) level and for the MM part, we used the AMBER force field. The BChl phytyl tail was included entirely in the QM part. To be consistent with the previous work [82], for HL-LH2 we have used the high-resolution X-ray structure of *Rps. acidophila* determined by Roszak et al. (unpublished results) instead of the crystal structure used as a starting point for the MD (PDB entry 1NKZ) [68]. The optimization was repeated for both α and β BChl of HL and LL-LH2 by keeping all the rest frozen, moreover we

also froze all the dihedral coordinates of the BChl at their crystal values.

4.3 The picture from the combined MD and QM/MMPol approach

As described in the Introduction, the three LH complexes have been investigated through a combination of classical MD simulations and excitonic QM/MMPol calculations.

In Table 4.1 we summarize a selection of the calculated excitonic parameters obtained as averages over the configurations coming from the MD simulations of the three complexes. The complete set of parameters is reported in Tab. A1 of Appendix A.

Table 4.1: Average site energies of the three not equivalent BChls and the two largest couplings within the 18-meric ring. Here $V^1_{\alpha\beta}$ and $V^2_{\alpha\beta}$ indicate the coupling between adjacent BChls belonging to chains of different units (inter-chain), and of the same unit (intra-chain), respectively. All values (in cm^{-1}) are reported together with the relative standard deviations computed along 50 frames of the MD trajectories.

	HL-LH2	LL-LH2	HM-PucD
αBChl	13527 (276)	13639 (290)	13790 (266)
βBChl	13556 (279)	13693 (297)	13715 (286)
B800-BChl	13783 (326)	13735 (312)	13767 (328)
$V^1_{\alpha\beta}$	266 (55)	149 (89)	127 (74)
$V^2_{\alpha\beta}$	298 (35)	281 (37)	285 (35)

As expected, the site energies of the BChl in the 9-meric ring (B800) remain almost identical in all complexes. On the contrary, significant changes are found in the α and βBChls of the 18-meric ring where a blue-shift is obtained moving from HL to LL and HM-PucD. In particular, the excitation of αBChl shows a blue-shift of 112 cm^{-1} going from HL to LL and a further shift of 151 cm^{-1} when moving to HM-PucD. In βBChl instead, the blue-shift

is 127 cm^{-1} from HL to LL but only 32 cm^{-1} when moving to HM-PucD. As it will be better detailed in the next section, these shifts are clearly correlated with the change in the H-bonding patterns: in fact, moving from HL to LL, αBChl loses an H-bond with Trp45 but it gains an H-bond with Tyr41 while βBChl loses the H-bond with Tyr44. Further moving from LL-LH2 to HM-PucD, αBChl loses the H-bond which this time is not replaced, whereas βBChl does not significantly change its local environment.

Also the couplings show significant differences in the three complexes. In particular, the inter-chain $V^1_{\alpha\beta}$ is reduced of ca. 44% when moving from HL to LL whereas for the intra-chain analog ($V^2_{\alpha\beta}$) the reduction is only of 5%; we note that the same trend was previously found by Montemayor et al [74], and also suggested from experiments[79, 103]. It is also worth noting that almost no differences are found between LL-LH2 and HM-PucD. The large difference in the $V^1_{\alpha\beta}$ arises from the Coulomb component of the coupling which decreases from HL-LH2 to LL-LH2 and HM-PucD whereas the effect due to the polarizable environment is similar in the three complexes Fig. A9. To further investigate the origin of this difference, we computed two parameters which are expected to affect the coupling, namely the BChl–BChl distance and their mutual orientation (here quantified in terms of the orientation factor κ from Förster theory).

As shown in Fig. 4.3, for the inter-chain pairs (corresponding to the $V^1_{\alpha\beta}$ coupling), the distance remains the same moving from the HL-LH2 to the other two complexes; on the contrary, the mutual orientation changes becoming less favorable for a large coupling in both LL-LH2 and HM-PucD. For the intra-chain pairs (coupled through $V^2_{\alpha\beta}$), instead, both distance and mutual orientation remain the same in all complexes. These observations lead to the following conclusion: the adjacent α and βBChls within the same unit have correlated fluctuations regardless of their H-bond network, while the fluctuations of α and βBChls across the two units (i.e. the inter-chain pairs) are correlated only in the HL-LH2 system, where there is an H-bond connecting the βBChl to the neighboring unit (Fig. 4.2d).

Let us now analyze how these changes in the excitonic parameters are

reflected in the exciton states and in the final absorption spectra.

For the sake of clarity, we recall that in a perfectly symmetric 18-meric cyclic aggregate, the collective exciton states that originate from the coupled Q_y transitions of individual BChls form a manifold of 18 states where there are two non-degenerate and eight pairwise degenerate excitons. The structure of the aggregates showing transition dipoles of individual BChls almost in the ring plane and the selection rules determine that only the lowest and the highest state pairs (commonly assigned with the quantum numbers $k = \pm 1$ and $k = \pm 8$, respectively) can be optically excited. The remaining 14 states are dark, i.e. not accessible optically from the ground state. Within this picture, an important measurable quantity is the so-called exciton width, namely the difference between the ± 8 and ± 1 states [103–105]. Static (as well as dynamic) disorder relaxes these symmetry-controlled selection rules by randomly shifting the states and removing their degeneracy. Moreover, disorder affects the exciton dipole strengths allowing their redistribution from optically allowed states of symmetric aggregates ($k = \pm 1$ and $k = \pm 8$) to adjacent dark states. This is exactly what happens in our simulations based on room temperature MD trajectories; however, it is still possible to define an exciton width by averaging the exciton Hamiltonian over the MD trajectory, and averaging all the site energies and couplings that are equivalent by symmetry. In this way, we obtain the properties of the (average) perfectly homogeneous rings. The resulting exciton widths, together with the shift of $k \pm 1$ exciton energy with respect to HL-LH2, are reported in Table 4.2 and compared with experimental data where available.

The simulated and experimental spectra are reported in Fig. 4.4 and the energies of the bright excitonic states are reported in Tab. A3 of the Appendix A. The lineshape was simulated as a convolution of Lorentzian functions centered on the exciton energies, whose widths were selected to match the experimental data. More specifically, we used the following half-width at half-maximum (HWHM): HL-LH2 150 cm^{-1}, LL-LH2 and HM-PucD 280 cm^{-1} for the low energy band, and 180 cm^{-1} for the high energy band. We note that with this procedure we describe all (homogeneous and

inhomogeneous) sources of broadening with a single lineshape. Although there are more sophisticate (and more reliable) approaches to include spectral broadening, an accurate description of the lineshapes is out of the scope of this work. Nonetheless, we note that the symmetric lineshapes used here could slightly skew the visual interpretation of the spectra, because the various LH2 complexes present asymmetrical lineshapes. However, such a bias is expected to be quite small, and properly accounting for the lineshape would not change the following interpretation.

Table 4.2: Calculated and experimental exciton widths for the three investigated systems at room temperature and shift of $k \pm 1$ exciton energy with respect to HL-LH2. The calculated values reported in square parenthesis are obtained with the inclusion of CT states. The exciton width is the difference between the $\pm$ 8 and $\pm$ 1 states of the 18-meric ring. The experimental exciton width for HL-LH2 is from Ref. 104. All values are in cm^{-1}.

		Exciton width	Shift
HL-LH2	Calc	1098 [1179]	-
	Exp	1259	-
LL-LH2	Calc	856 [899]	277 [312]
	Exp	N.A.	501
HM-PucD	Calc	842 [888]	352 [386]
	Exp	N.A.	806

As it can be seen from the tables and the graphs, the simulations give an accurate description of HL-LH2 and they reproduce the blue-shits when moving to the other two complexes. However, the calculated shifts are underestimated; this is particular evident for HM-PucD which remains too similar to LL-LH2. As a further test, we have applied the approach used in the previous study for HL-LH2: [86] we calculated the excitonic Hamiltonian for each of the 50 frames and we obtained the resulting exciton states and absorption spectra which were finally averaged. The resulting absorption spectra of the three complexes are shown in Fig. 4.2 of the Appendix A, whereas the energies of the excitonic states are reported in Tab. A3 of the

Appendix A. These data show exactly the same trend as the one obtained from the average Hamiltonian.

Before moving to a detailed analysis of the results of our calculations, it is useful to consider the experimental results on genetically modified LH2 complexes from *Rps. sphaeroides* [78, 107]. By constructing single (Tyr44, Tyr45 $\rightarrow$ Phe, Tyr) and double (Tyr44, Tyr45 $\rightarrow$ Phe, Leu) site-specific mutants of wild-type (WT) LH2, these studies found that the absorbance of the B850 band at 77K was blue-shifted by about 220 cm^{-1} and 450 cm^{-1}, respectively. As the single mutation corresponds to the loss of one H-bond of βBChl, and the double mutation to the simultaneous loss of two H-bonds (on β and α BChl), these data seem to show additivity in the H-bonds effects. However, it has to be noted that these shifts are about a half of the ones measured for LL-LH2 and PucD. It is true that the mutants refer to another organism, and that their spectra have been measured at a much lower temperature. However, from the comparison, one can conclude that the mutations localized on the H-bonding residues preserve the main excitonic characteristics of WT (HL) LH2 much more than in the low-light complexes. Moreover, it is interesting to note that the shifts in the mutants are of the same order of magnitude of those calculated for the two complexes. This consideration suggests that our calculations properly include the main effect of the H-bonds on site energies and couplings, but they miss some additional effects.

To better investigate this suggestion, in the following section we separately analyze all the possible effects (environmental, structural, etc) that have been proposed in the literature to explain the spectroscopic changes in the three complexes.

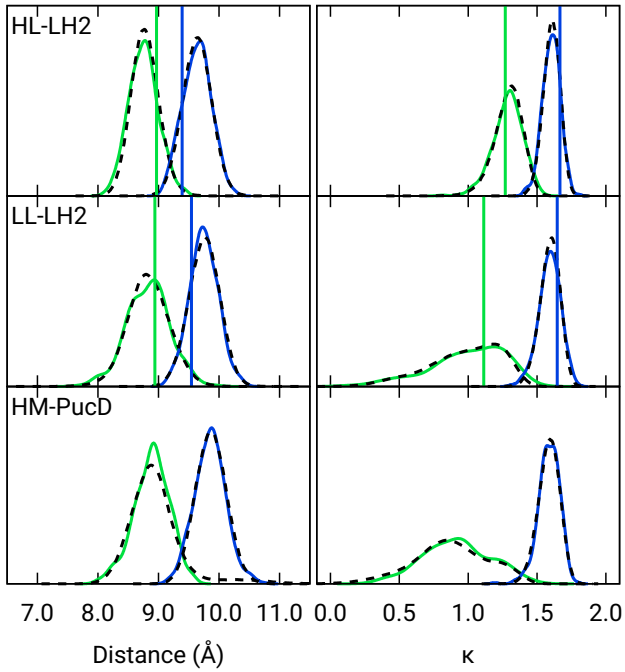

Figure 4.3: Distribution of the BChl-BChl distance (left) and the orientation factor κ (right) for the $V^1_{\alpha\beta}$ (green) and $V^2_{\alpha\beta}$ (blue) couplings for the three complexes. The BChl-BChl distance is here calculated on the basis of the effective center of the macrocycle ring defined as the average of the four nitrogen atom positions. The orientation factor κ is defined according to the Förster theory. We plotted the distribution over the whole MD (dashed black line) and over the select 50 frames (colored solid lines). The vertical sticks indicate the values measured on the crystal structures. Copyright: `https://doi.org/10.1039/c9sc02886b`.

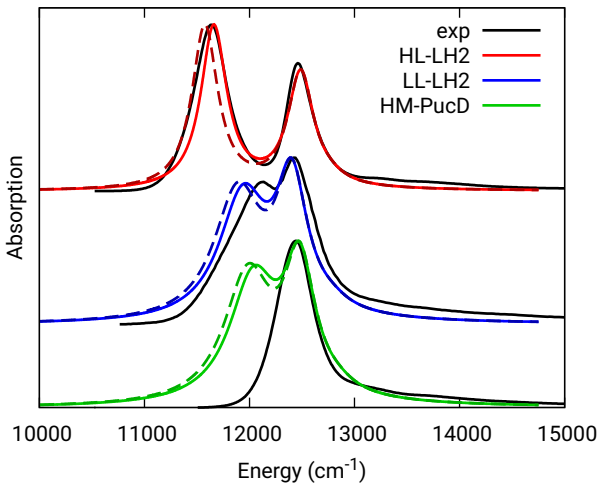

Figure 4.4: Comparison between simulated and experimental absorption spectra of the three investigated complexes. The spectra computed with the inclusion of CT states are drawn as dashed lines. The experimental spectra for the HL and LL-LH2s have been adapted from Ref. 106 while the PucD spectrum has been adapted from Ref. 72. In the plots, we shifted all the simulated spectra by -1247 cm^{-1} to match the experimental B800 band: this shift is consistent with the typical error of the QM model (here TDDFT) in describing excitation energies. The lineshape was simulated using a convolution of Lorentzian functions centered on the excitonic states, the widths of the Lorentzian functions were selected to match the experimental data.

4.4 Dissecting the possible origins of the spectral differences

As a first analysis, we consider the effect of the composite environment (namely the protein, the membrane and the water solvent). In Fig. 4.5 we report the contribution of selected residues/cofactors on the Q_y excitation of the α and βBChls of the 18-meric ring in the three complexes (to have a compact notation the one letter code is used for the residues). These contributions have been obtained by repeating the calculation of the excitation energy of each BChl by "switching off" the selected residue, i.e., setting to zero its charges and polarizabilities. We calculated the contribution of each residue as the difference between the excitation energy computed in the full MMPol environment and that calculated after the switch-off. A more complete analysis including all the residues within 6 Å from the BChls is shown in Fig. A7 of the Appendix A.

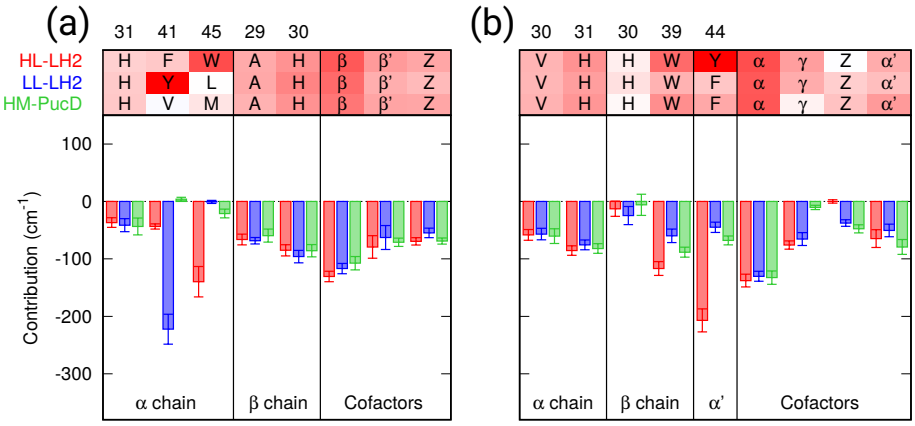

Figure 4.5: Contributions of the residues/cofactors to the Q_y excitation of α (a) and β (b) BChls in the three complexes. The average value and the standard error bar are shown; the results are computed on 20 frames of MD. Amino acids are displayed using the one letter code (H=his, F=phe, Y=tyr, V=val, W=trp, L=leu, M=met, A=ala), the two adjacent BChls in the ring are displayed using "α" and "β" whereas Z refers to the closest carotenoid and γ to the B800 BChl. Copyright: `https://doi.org/10.1039/c9sc02886b`.

As expected, the main contributors to the Q_y energy are the H-bonded residues, namely, Trp45 for αBChl and Tyr44 for βBChl in HL-LH2, and Tyr41 for αBChl in LL-LH2. All of them lead to a red-shift ranging between 140 cm^{-1} (Trp45) to 220 (Tyr41). Notably, the effect of changing the H-bonding network from HL to LL-LH2 is not only a blue-shift for βBChl, due to the loss of the H-bond with Tyr44, but also a (small) red-shift for αBChl, due to the replacement of Trp45 with Tyr41. The latter is in fact a stronger H-bond donor than the former [108]. As a further interesting note, we observe that for all the BChls in the three complexes a not negligible source of red-shift is given by the adjacent BChls.

From this analysis it appears that each H-bond is responsible for a red-shift of the Q_y excitation of about 140-220 cm^{-1}. In previous computational studies, different values were suggested for the same contribution, and in some cases values as large as 500 cm^{-1} were proposed [109]. However, we have to note that those calculations were performed using a different model (not involving a polarizable environment) and different structures with respect to the ones here used.

To analyze the possible effect of the structures, in Fig. 4.6, we report the H-bond length distribution for αBChl-Trp45 and βBChl-Tyr44 of HL-LH2 and for αBChl-Tyr41 of LL-LH2 along the MD trajectory, compared with the results from two differently relaxed crystal structures. The first of these relaxed structures was taken from our previous work [82] and it was obtained at ONIOM(QM:MM) level where the BChl (without the tail) and the residues directly interacting with it, namely the axially coordinating histidine and the hydrogen-bonded residues, were included in the QM region and allowed to move. The rest of the environment was instead kept frozen at the crystal structure, including the other BChls. The second relaxed geometry (from now on indicated as "constrained optimization") is also obtained at the ONIOM(QM:MM) level, but this time only the BChl (without the tail) was included in the QM layer. Moreover, all its dihedral angles were kept frozen at the crystal values together with the positions of all the atoms of the environment.

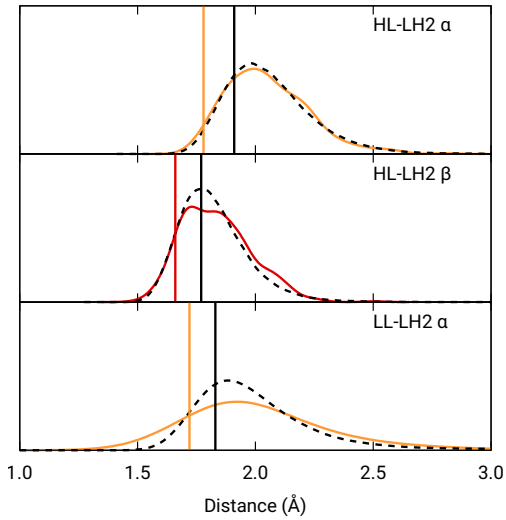

Figure 4.6: Hydrogen bond length distributions for αBChl-Trp45 and βBChl-Tyr44 of HL-LH2 and for αBChl-Tyr41 of LL-LH2 along the MD compared with the results from the full optimization from previous work (black sticks) [82] and from the present constrained optimization (colored sticks) of the BChls. We reported the distribution from the whole MD (18000 frames) as a black dashed line and from the 50 frames selected for excitonic calculations as a colored line. Copyright: `https://doi.org/10.1039/c9sc02886b`.

As it can be seen from Fig. 4.6, the maximum of the distribution of the H-bond distances for the three investigated pairs agrees well with a fully relaxed structure while the constrained relaxation (which remains closer to the original crystal structure) gives somehow shorter distances. Nonetheless, the differences are around 0.1 Å and their effects on the induced shift on the Q_y excitation are less than 40 cm^{-1}.

To further confirm the robustness of our results on the description of H-bond effects, we have performed two tests. In the first test, we have validated the accuracy of the selected MMPol parameters (charges and polarizability). To do so, we have compared the H-bond contributions reported in Fig. 4.5 with the ones computed on the same 20 structures extracted from the MD trajectory, this time using a QM description also for the H-bonded residue

under investigation, while leaving the rest of the environment at MMPol level. The correlation between the QM/MMPol and the QM/QM/MMPol results are reported in Fig. 4.7.

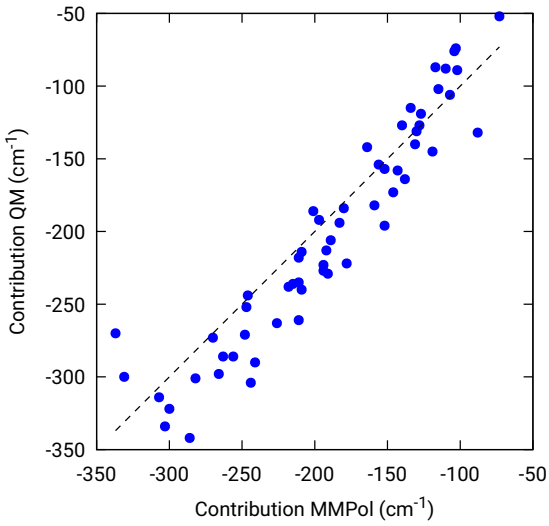

Figure 4.7: Benchmark of the MMPol description of the H-bonding residue against a QM description. The measured property is the effect of the residue (Tyr or Trp) on the BChl Q_y excitation energy. For this test, 20 structures extracted from the MD trajectory were used. Copyright: `https://doi.org/10.1039/c9sc02886b`.

The good correlation tells us that in the present system, the H-bond effects are well described by a classical model including both electrostatic and polarization effects, whereas possible non classical effects (dispersion and/or charge-transfer) are negligible.

In the second test, we assess the quality of the selected DFT functional in describing the H-bond effects on the Q_y excitation. For this test we have repeated the calculations of the QM/MMPol H-bond contributions with five different functionals selected among the most successful ones to describe electronic excitations in large chromophores (PBE0, CAM-B3LYP, M062X,

LC-BLYP, ωB97XD). The correlations between B3LYP and these functionals are reported in Fig. A8 of the Appendix A. The obtained results show an almost perfect correlation with PBE0, while all the other (long-range corrected) functionals give a larger spread of values, the maximum differences being for ωB97XD. Thus, the sensitivity of the excitation energy to the H-bond increases with the amount of exact exchange in the DFT functional. However, there is no systematic bias between one functional and another, and on average the results are quite consistent.

All these tests confirm that our description of the H-bond effects on the Q_y excitation energy is sufficiently robust.

The changes in the H-bonding network observed in the three LH complexes also correspond to changes in the acetyl dihedral angle. As already commented in the Introduction, the acetyl (C=O) double bond is conjugated with the macrocycle ring: we thus expect that moving the acetyl out of the macrocycle plane could lead to a blue-shift of the Q_y excitation, and indeed this effect has been proposed as responsible for the spectroscopic changes observed from HL to LL-LH2 [75].

As done for the H-bond distances, also here we compare the distribution for the dihedral angles of α and βBChl in the three different complexes along the MD trajectory with the results from the two differently relaxed crystal structures. This comparison, reported in the top panel of Fig. 4.8, clearly shows that for all the investigated complexes, the calculated distributions are centered on a planar structure for both α and βBChl. This finding disagrees with what found in the constrained optimizations, which, we recall, coincide with the crystal data for all dihedral angles. The latter in fact indicate a dihedral angle of *ca* $\pm 20°$ for α and βBChl of HL-LH2, which changes to -30° in LL-LH2 (no crystal data are available for PucD). The out of plane rotation of the acetyl group is also confirmed by the full optimizations of the BChls obtained in the previous work, even though for αBChl in LL-LH2 the rotation is smaller.

To check the origin of the planarization in the MD simulations, we run an optimization of the BChls described at the same MM level within a frozen

environment. Also in this case, the acetyl group planarizes (all the dihedral values are reported in Tab. A4 of the Appendix A), so we can conclude that the MM force-field here used for the BChl overstabilizes planarity in the acetyl orientation. This has the consequence of reducing the differences in the site energies of the different complexes, as their BChls present roughly the same planar acetyl orientation.

To quantify the real effect of the torsion of the acetyl group on the Q_y excitation, we performed a relaxed scan of the βBChl geometry within the LL-LH2, and finally calculated the Q_y excitation energy on the corresponding structures. This scan has been obtained at ONIOM(B3LYP/MM) level where the QM subsystem is made of the βBChl together with the coordinating histidine and the close by phenylalanine while the MM subsystem (the phytyl chain of the BChl, the other BChls and cofactors and the protein) is kept frozen in the crystal configuration. For each torsional angle, the QM subsystem has been allowed to relax. From the results reported in the right panel of Fig. 4.8, we can obtain a rough estimate of what we are missing in our MD descriptions, where the acetyl dihedral angle always averages to zero. If we assume valid the crystallographic estimates of the dihedral angles, a further blue-shift of about 220 cm^{-1} should be considered for the αBChl (for which the crystal dihedral angle changes from 20° to -30°) and 120 cm^{-1} for the βBChl (for which the dihedral angle change from -25° to -36°). We note that if we add these shifts to the Q_y energies calculated from the MD and combine them with the corresponding couplings (the dihedral angle in fact does not significantly affect the coupling) we obtain that the energy differences in the $k = \pm 1$ between LL and HL-LH2 increases to 452 cm^{-1} (486 cm^{-1} if we also include the CT effect). These results, compared with the experimental shift of 501 cm^{-1}, seem to show that an accurate prediction of the dihedral angle could be the missing piece in our simulation to fully reproduce the spectroscopic differences between HL and LL forms of LH2. This analysis is exemplified in Fig. 4.9, where we report the same comparison reported in Fig. 4.4, but now the simulated spectra of LL-LH2 and HM-PucD have been obtained by correcting the site energies by the additional blue-shift induced

by the different out-of-plane distorsion of αBChl and βBChl as predicted by crystal data (for HM-PucD we have assumed the same distorsion as LL-LH2).

This conclusion in some way goes against what reported in recent computational studies by De Vico and coworkers [77, 110], who used a multistate multiconfiguration restricted active space with second-order perturbation theory correction (MS-RASPT2) to calculate the excitation energies of the BChl at their crystal structure. In the same study, instead, a different source of the blue shift was suggested, that is, the change in the BChl macrocycle ring curvature. In particular, by using the geometry of α and βBCHls taken directly from the crystal structure (without any relaxation) of HL and LL-LH2 huge blue-shifts of the order of 0.23 eV (ca. 1850 cm^{-1}) are found for both BChls.

In order to test this hypothesis, we have compared the unrelaxed crystal structures with the two sets of refinements of the same crystal structures already considered in the previous analysis, namely, (i) the ones obtained from a constrained optimization of the BChls where all their dihedral angles were kept frozen to the crystal values, and (ii) the full optimization where all the internal degrees of freedom of the BChl and of the close-by residues were allowed to relax.

The obtained results, reported in Table A5 of the Appendix A, clearly show that, if we use the crystal structures without any relaxation, a very large (and nonphysical) blue-shift of ca. 2000 cm^{-1} is found for both α and βBChls when moving from LHL to LL, exactly as found by De Vico et al [77]. However, as soon as we relax the bond lengths (and bond angles), still keeping the dihedral angles frozen (and hence the macrocycle curvature), these differences almost disappear: they reduce to 40 cm^{-1} for and 60 cm^{-1} when calculated for the isolated α and βBChls, respectively, and to 65 and 186 cm^{-1} when the effect of the MMPol environment is included. If we further relax all the internal degrees of freedom together with the close by residues, we do not see any further significant change, showing that the bond lengths play a major role in determining the excitation energy. As a matter of fact this result was expected due to the conjugated nature of the macrocycle

ring.

To reach a more detailed picture of the role of bond lengths in the excitation energy shifts, we trained a linear regression model, using all bond lengths in the macrocycle ring as explanatory variables for the Q_y excitation energies computed along the MD trajectory. As shown in Fig. A10, the prediction based on bond lengths explains more than 60% of the variability in excitation energies. Therefore, bond lengths are the main factor determining Q_y excitation energies of BChls in LH2. We then employed the parameters of the linear regression to predict the excitation energies in crystal and optimized structures. These structures were not employed for the fitting, but nonetheless their excitation energy is consistently predicted by the bond lengths model. In particular, this model correctly predicts that both α and β unrelaxed structures of HL-LH2 are strongly red-shifted. We can thus conclude that the bond lengths are ultimately responsible for the nonphysical excitation energy in the HL-LH2 1NKZ crystal structure, which also explains the huge blue-shift found for both α and βBChls when moving from HL to LL.

These data, together with the ones obtained from the MD (where the macrocycle ring is almost planar), indicate that the curvature of the macrocycle as described by the crystal structures does not significantly contribute to the blue-shift of the Q_y excitation energies. In addition, this analysis clearly shows the limit of the crystal data for bond lengths (and angles) especially when a conjugated pigment is involved.

As a last analysis, we considered the effect of higher energy CT states between adjacent BChls, which can couple with the Q_y excitations and finally lead to changes in the exciton energies.

To do that, we have calculated the four Q_y-CT couplings in the inter-chain and intra-chain BChl dimers of HL and LL-LH2 along the respective MD trajectories. Due to the very close similarity observed up to now in the simulation of LL-LH2 and PucD, this analysis of CT has been limited to the former complex only.

As shown in Fig. 4.10, the present results seem to indicate that the

effect of CT couplings is more modest than estimated in our previous study where the effects of the fluctuations were not included [82], Within that static picture, the Q_y-CT couplings were systematically smaller for LL-LH2 with respect to HL-LH2, with some of the inter-chain couplings dropping to almost zero. However, when averaging these couplings along the MD trajectory, the differences between the two systems are reduced, even though the LL-LH2 inter-chain couplings are significantly smaller than those for HL-LH2. For the intra-chain couplings, instead, the picture obtained with MD is reversed with respect to the static picture (Fig. 4.10b), and the couplings in LL-LH2 become larger than those of HL-LH2.

In order to ascertain that these differences do not arise from some geometrical bias of the BChls, we recomputed the Q_y-CT couplings using the geometries from the constrained optimization described before: these calculations (see cOPT vales in Fig. 4.10) confirm the same picture obtained in the previous work (fOPT values) showing that the differences found in the couplings from MD structures are due to the different average configuration of the 18-meric ring (and of the embedding protein) with respect to the crystal structure. In fact, the results obtained along the MD trajectory stem from the variety of distances and orientations undertaken by the BChl dimers, which, combined with the sensitivity of CT couplings to the geometry [98], give rise to a large variability of coupling values (see also the average values and the standard deviations reported in the Table A2 of the Appendix A. Notably, the variance of Q_y-CT is larger in LL-LH2 than in HL-LH2, possibly as a consequence of the larger geometrical freedom of the β BChl. As a final note we observe that the presence of dark states (identified as charge transfer states and/or polaron pairs) have been also investigated experimentally in the HL-LH2 and PucD forms from *Rps. palustris* using two-dimensional electronic spectroscopy [111], These experimental observations seem to suggest that such states are present in both complexes and they can act as strong quenchers. However, in those observed dark states likely refer to relaxed states, whose energy is much more red-shifted than what found in the present simulations where we have calculated the vertical

CT states.

4.5 Why is PucD so blue-shifted?

From the previous analysis, it came out that by removing the H-bond connecting the BChls of two different units, a significant decrease in the coupling is obtained. We have explained these effects in terms of a much larger distribution of the relative orientations explored by the inter-chain BChls in LL-LH2 and PucD complexes with respect to HL-LH2, which finally averages in smaller couplings between different units. On the contrary, the coupling between BChls of the same unit has shown to stay almost the same even when one (LL-LH2) or both (PucD) H-bonds are removed. One could explain this findings saying that that the dynamics of the $\alpha\beta$ chains of the same unit preserve the relative orientation of the BChls which are anchored to the respective chains through His residues, and the change in H-bonds has only a minor effect.

However, the validity of this assumption cannot be fully validated by our MD simulations. In fact, if a complete release of all H-bonds interactions between the BChls and the binding chain (as it happens in PucD) would lead to changes in the relative position of the BChls, this could be seen only allowing the complex to explore structures which are farther with respect to the starting (crystal) structures. In particular, the intra-chain pair could be more flexible than what revealed in our MD simulation, thus leading to a further reduction of the corresponding couplings. If this is the case, we would see a significant effect in the PucD. To have an indirect check of this suggestion, we have estimated how much the intra-chain couplings should be reduced to achieve the experimental shift: to do that we have recalculated the position of the k=$\pm$1 exciton state of PucD using the excitonic parameters obtained from our simulation and scaling only the intra-chain coupling $V^2_{\alpha\beta}$ by factors ranging from 0.65 to 0.85. The results are reported in Fig. 4.11 As it can be seen from the graph, by reducing the $V^2_{\alpha\beta}$ of 30% we get the expected spectrum showing a single band at about 800 nm. We note however,

that here we have also corrected the Qy excitations of α and βBChls so to account for the artificial planarization of the acetyl group already seen in LL-LH2. If this further effect is not included, we need to introduce a much larger scaling of the coupling (namely around 50%) which would be rather unlikely.

4.6 Conclusions

We have investigated the origin of the exciton tuning in LH2 complexes when grown in different light conditions. The study has been made possible by the combination of molecular dynamics simulations and excitonic Hamiltonians calculated through a multiscale QM/MMPol approach. In particular, we studied the high light and low light forms of LH2 from *Rps. acidophila* and a third structure encoded by the *puc*$\mathrm{BA_d}$ gene from *Rps. palustris*. Whereas for the first two structures a high-resolution crystal structure is available, for the third one it is not, so we employed the homology modeling technique to get a starting structure for the molecular dynamics. By applying the same computational strategy to the three complexes, we could achieve a detailed understanding of the origin of their excitonic and spectroscopic differences.

First, we investigated the role of the composite environment (the protein, the membrane and the solvent) in the tuning of the $\mathrm{Q_y}$ excitations of the BChls and of their electronic couplings. For the excitation energies, we found solid evidence that the electrostatic and polarization effects of the environment remain similar in the three complexes. Instead, the differences in the H-bonding residues explain a large part of the observed exciton tuning. In particular, we have estimated that the loss of each H-bond not only accounts for about 200 $\mathrm{cm^{-1}}$ of blue-shift of $\mathrm{Q_y}$ but it also indirectly affects the electronic couplings. Specifically we have seen that because the βBChl in HL-LH2 is connected to the neighboring unit by a H-bond, the thermally induced fluctuations in its orientation remain correlated to the one of the adjacent αBChl in such unit. In the other two complexes, instead, the H-bond is lost and the orientation in the two BChls belonging to different units

becomes uncorrelated thus significantly reducing their coupling.

We also investigated the role of the internal geometry of the pigments, through the acetyl torsion and the curvature of the macrocycle ring of the BChls. For the first parameter, we found that, as expected, a deviation from the planarity results in a blue-shift. Indeed, this effect, when combined with all the above described H-bond effects on site energies and couplings, could fully explain the measured spectral differences between HL and LL forms of LH2. On the contrary, the curvature of the macrocycle ring seems not to play a major role when a proper description of the bond lengths within the conjugated ring is accounted for. In particular, the huge differences found in the literature for the Q_y excitations of HL and LL-LH2 [77, 110] can be accurately explained by the fact that the crystal (1NKZ) structure of HL-LH2 displays an nonphysical conjugation pattern of the macrocycle ring due to inaccurate bond lengths. Instead, the BChls from the crystal structure of LH3 (1IJD) present a more regular pattern of bond lengths within the macrocycle ring and, as a result, a huge blue-shift is obtained when compared to the HL-LH2.

Furthermore, we investigated the role of higher energy CT states between adjacent BChls in the 18-meric ring which can couple to the Q_y excitations and finally affect the excitons. Indeed, we have found that the HL and LL forms of LH2 have different effects of CT states, but also that thermal fluctuations tend to reduce these differences with respect to a picture based on the crystal structure [82].

Finally, we have suggested a possible explanation of the measured large change in the spectrum when moving from LL-LH2 to PucD in terms of an additional reduction of the couplings, this time involving the BChls belonging to the same unit. What is difficult to say from the present simulations is whether this further decrease of the couplings in PucD involves distortions of the 18-meric ring with respect to LL-LH2 or is instead induced by a larger mobility of the BChls around their unaffected average position. In order to confirm one or the other of the two effects would in fact require to largely extend the time windows to be investigated by the MD trajectory.

This is certainly an aspect which requires further investigation both from an experimental and a computational point of view.

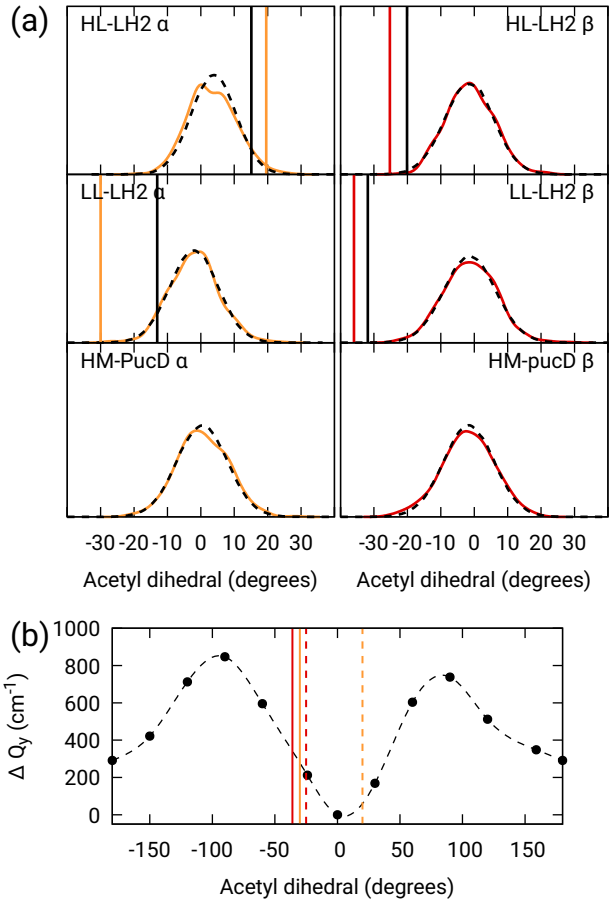

Figure 4.8: (a) Dihedral angle distributions from the MD trajectory compared with the results from the previous work full optimizations (black sticks) [82] and from the present work constrained optimizations (colored sticks). We reported the distribution from the whole MD (18000 frames) as a black dashed line and the one from the 50 frames selected for QM calculations as a colored line. (b) Blue-shift of the excitation energy along a relaxed scan of the LL-LH2 βBChl acetyl dihedral. The vertical sticks show the dihedral angles of the crystal structures: αBChls yellow, βBChls red, HL-LH2 dashed line, LL-LH2 solid line. Copyright: `https://doi.org/10.1039/c9sc02886b`.

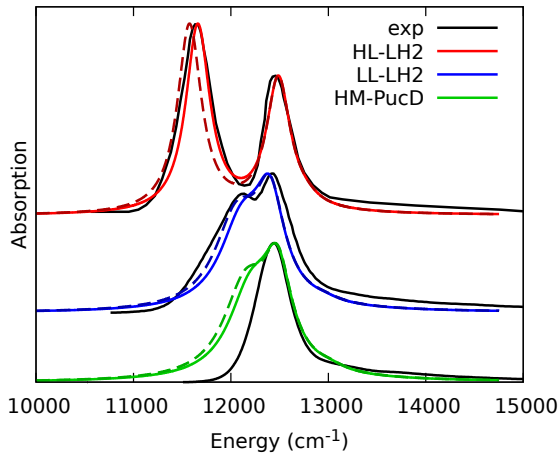

Figure 4.9: Comparison between simulated and experimental absorption spectra of the three investigated complexes. The spectrum of HL-LH2 is the same reported in Fig. 4.4 while the spectra of LL-LH2 and HM-PucD have been recalculated by correcting the site energies of αBChl and βBChl by the additional blue-shift induced by the different out-of-plane distorsion as predicted by crystal data (for HM-PucD we have assumed the same distorsion as LL-LH2). The spectra computed with the inclusion of CT states are drawn as dashed lines. Copyright: `https://doi.org/10.1039/c9sc02886b`.

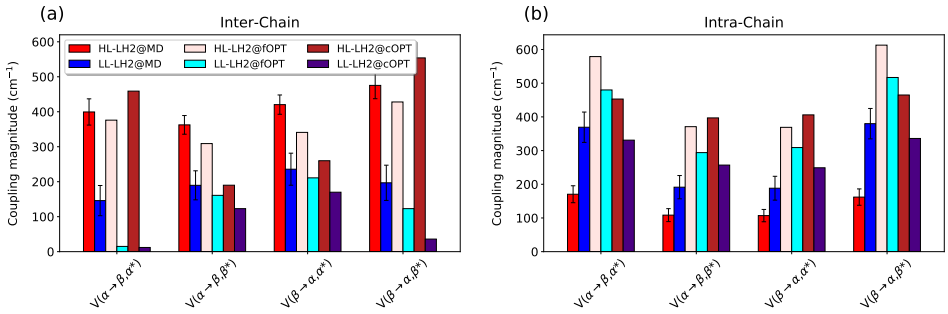

Figure 4.10: Comparison of the Q_y-CT couplings in the (a) inter-chain and (b) intra-chain BChl dimers along the MD trajectory and for the crystal structures. The results of the previous work [82] are marked as fOPT, whereas cOPT refers to the results of the constrained optimizations. Absolute values are shown (cm^{-1}). Error bars indicate 95% confidence intervals. Copyright: `https://doi.org/10.1039/c9sc02886b`.

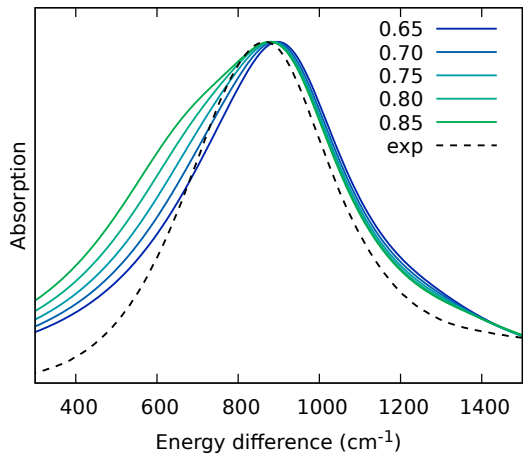

Figure 4.11: Simulated spectra of PucD models obtained for different scaling factors of the intra-chain coupling $V^2_{\alpha\beta}$. The x-axis corresponds to the energy shift with respect to the B850 peak of HL-LH2. Copyright: `https://doi.org/10.1039/c9sc02886b`.

Chapter 5

Case Study II: LOV-based Fluorescent Proteins

In this Chapter we present an investigation of synthetic flavin-based fluorescent proteins (FbFPs) developed to be applied as biomarkers in cell and molecular biology studies. FbFPs are a class of fluorescent reporters derived from light-oxygen-voltage (LOV) sensing proteins. Through mutagenesis, natural LOV proteins have been engineered to obtain improved fluorescence properties. In this study, we combined extended classical Molecular Dynamics simulations and multiscale Quantum Mechanics/Molecular Mechanics methods to clarify the relationship between structural and dynamic changes induced by specific mutations and the spectroscopic response. To reach this goal we compared two LOV variants, one obtained by the single mutation needed to photochemically inactivating the natural system, and the other (iLOV) obtained through additional mutations and characterized by a significantly improved fluorescence.

Our simulations confirmed the "flipping and crowding" effect induced in iLOV by the additional mutations and revealed its mechanism of action. We also showed that these mutations, and the resulting differences in the composition and flexibility of the binding pockets, are not reflected in significant shifts of the excitation and emission energies, in agreement with the

similarity of the spectra measured for the two systems. However, a small but consistent reduction was found in the Stokes shift of iLOV, suggesting a reduction of the intermolecular reorganization experienced by the chromophore after excitation which could slow down its internal conversion to the ground state and improve the fluorescence.

The results presented here are based on Ref. [112], the graphical abstract for this study is shown in Fig. 5.1. This work was published on the *The Journal of Physical Chemistry B* and can be accessed freely through the link: `https://dx.doi.org/10.1021/acs.jpcb.0c10834`.

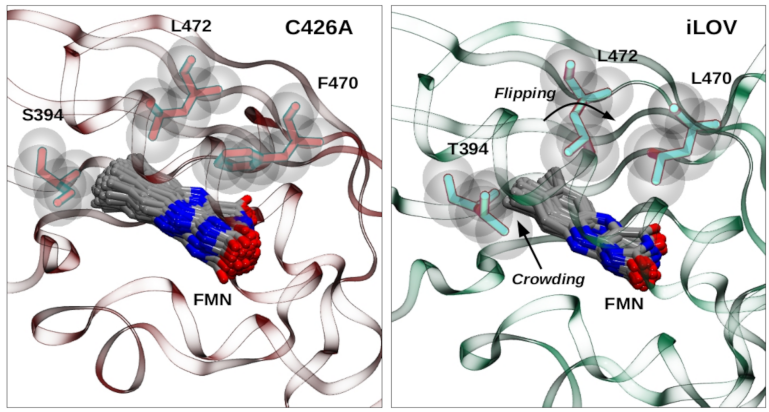

Figure 5.1: Graphical abstract of the paper that is the basis of this Chapter (see Ref. [112]). The obtained results demonstrate that the isoalloxazine ring of the flavin mononucleotide (FMN) is stabilized by the mutations present in the iLOV active-site thus reducing the extent of intermolecular reorganization experienced by the chromophore after excitation to the S_1 state. Copyright: `https://dx.doi.org/10.1021/acs.jpcb.0c10834`.

5.1 Introduction

Green fluorescent protein (GFP) has revolutionized the imaging of dynamic processes within living cells[113]. However, the use of GFP as *in vivo* reporters is limited by some environmental and cellular factors impeding either chromophore formation or fluorescence activity [114, 115]. In this context, the so-called flavin-based fluorescent proteins (FbFPs) emerged as an alternative class of FPs[116–119]. FbFPs are derived from a highly conserved family of blue light photoreceptors known as light, oxygen, and voltage (LOV) sensing proteins. In nature, LOV proteins typically associate with flavin mononucleotide (FMN) to function as blue-light photoreceptors and regulate a serie of cellular processes in both bacteria (stress response and virulence) and plants (phototaxis) [9, 10].

Natural LOV domains bind FMN noncovalently and, upon UVA/blue-light excitation, undergo a reversible photocycle involving the formation of a covalent bond between the chromophore and a conserved cysteine residue with complete loss of fluorescence [120]. The substitution of the cysteine residue abolishes LOV domain photochemistry and recovers the fluorescence of the bound FMN[121]. LOV domains photochemically inactivated in this way are inherently fluorescent; however, additional mutagenesis was needed to further improve their fluorescence and photostability [121–123]. Random and structure-based engineering methods have been combined to generate a large pool of mutants, which are expressed and then selected according to specific properties, such as fluorescence quantum yield, thermal stability and photobleaching reversibility [121–125]. In particular, large attention has been given to a specific class of LOV-based reporter variants with improved properties properties obtained from the LOV2 domain of *Arabidopsis thaliana* phototropin 2 (phot2) through a directed evolution approach based on the DNA shuffling technique and screening toward enhanced fluorescence [121].

As a first mutation, the photoactive cysteine (Cys426 of *Arabidopsis* phot2) was replaced with alanine to achieve the photochemical inactivated derivative C426A. Five additional mutations (S394T, S409G, I452T, F470L

and M475V see Five additional mutations (S394T, S409G, I452T, F470L and M475V, see Fig. 5.2) were introduced to such a derivative leading to an improved LOV (iLOV) showing a substantial increase in fluorescence intensity and emission quantum yield with respect to the single mutant C426A [121, 125].

The iLOV domain was functionally and structurally characterized by Christie et al. [122]. From the crystallographic structures, the authors proposed that the five additional mutations of iLOV stabilize its structure by increasing packing interactions, especially in the flavin-binding cavity. According to this model, the FMN isoalloxazine ring would be rigidified through a "crowding and flipping" process undergone by the T394 and L472 side chains. The crowding is related to the replacement of a serine residue with a threonine residue in the flavin-binding pocket of iLOV (S394T mutation, see Fig. 5.2) while the flipping effect is the rotation of a leucine (L472) side chain in the direction of the chromophore induced by the F470L mutation in iLOV. Both modifications are supposed to increase the stability of flavin in the pocket.

Since all the experimentally obtained FbFPs have similar spectral features (maximum absorption and fluorescence at 447 nm and 493 nm, respectively) [121, 122, 125–128], further optimization efforts have been focused on enhancing the emission colour range by introducing single-point mutations in the FMN binding site. The goal in this case is to reach the "biotransparent" window (650–900 nm) in which light easily penetrates through mammalian tissues [129, 130]. Within this research line, a fundamental role has been played by computational studies [127, 131–134]. In particular, these studies have clearly shown that a detailed conformational analysis of the protein is the necessary prerequisite for any research efforts aiming to design variants with further improved fluorescence properties [134].

Following these findings, here we have performed microsecond-scale simulations of the C426A and iLOV variants originally engineered by Chapman et al [121]. These trajectories have been used to compare the structural dynamics of the two systems and to give a statistically meaningful set of con-

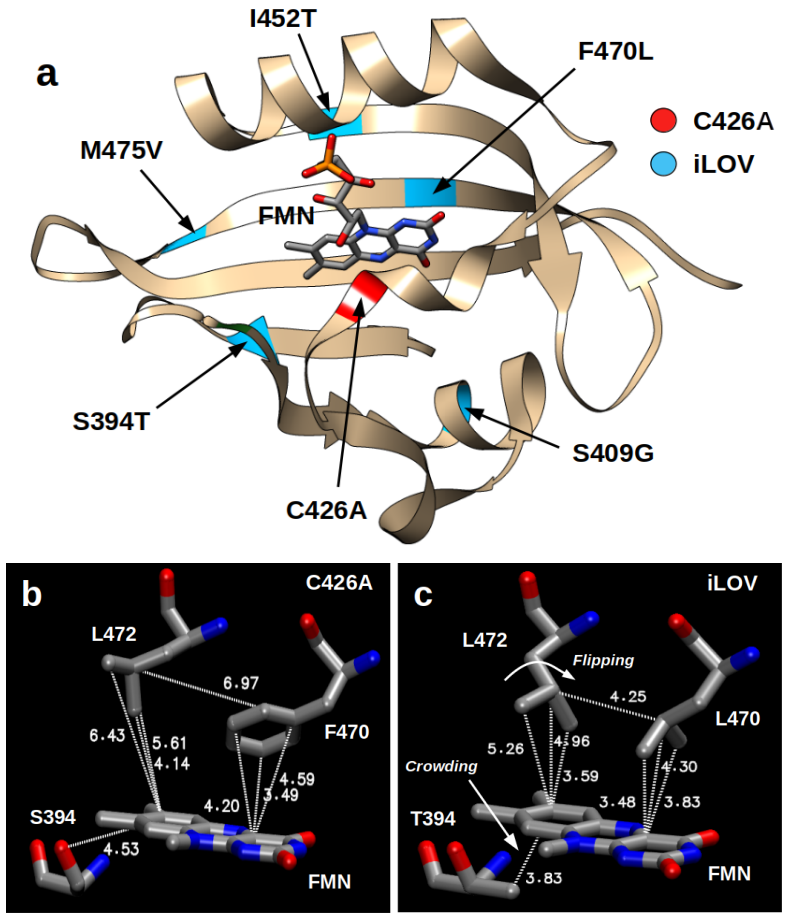

Figure 5.2: (a) Mutations present in C426A and iLOV with respect to the wild-type LOV2 domain. In the former, the photoactive cysteine was replaced with alanine and in the latter, five additional mutations (S394T, S409G, I452T, F470L and M475V) were introduced. (b) and (c) Network of residues involved in the "crowding and flipping" model. The distance values are in angstroms and were obtained from 4EER (C426A) and 4EES (iLOV) crystal structures. Residue numbering is based on the phot2 protein sequence, which is the same found in PDB structures. Copyright: `https://dx.doi.org/10.1021/acs.jpcb.0c10834`.

figurations for the simulation of excitation and emission energies and spectra. The latter step have been performed through a polarizable embedding quantum mechanics/molecular mechanics approach (QM/MMpol) which accounts for mutual polarization effects between the QM chromophore and the classical protein residues (and solvent molecules) [48].

The goal here is twofold: first we aim at verifying the "crowding and flipping" mechanism proposed to explain the fluorescence increase in iLOV [122], and secondly we attempt to provide a clear connection between structural and spectroscopic properties, which can be used in further optimizations of LOV-based fluorescent proteins.

From this investigation, it comes out that the binding pockets are characterized by a large flexibility of its H-bonding network in the μs timescale in both systems. However, this flexibility does not seem to significantly affect the excitation and emission energies. On the other hand, the additional mutation present in iLOV is responsible for a reduced mobility of the isoalloxazine ring of FMN within the pocket. This "crowding" effect also slightly reduces the extent of intermolecular reorganization experienced by the chromophore after excitation.

5.2 Computational details

5.2.1 Molecular dynamics

MD simulations were performed on the two LOV variants described by Christie et al. through crystallography analysis [122]: the photochemically inactivated LOV2 domain from *A. thaliana*, named C426A mutant (PDB: 4EER), and the iLOV protein containing 5 single-point mutations (S394T, S409G, C426A, I452T and F470L) (PDB: 4EES). In Fig. 5.2 of the Supporting Information we report the multiple sequence alignment of wild-type LOV2 domain, C426A and iLOV.

By using the *tleap* module of AmberTools [21], we carried out the preparation of the two investigated systems, their solvation within a truncated octahedron water box (with $\sim$ 23000 water molecules), as well as the addition of Na^+ and Cl^-ions at 0.15 M. Extra Na^+ ions were also add so as to achieve system charge neutrality. The two simulated systems have, in total, approximately 70000 particles.

The system minimization was done by first minimizing hydrogen atoms, next the solvent components, the protein-chromophore complex and finally the whole system. For the MD simulations, we first performed the system heating which was divided into two steps: the first one from 0 to 100 K (5 ps in the NVT ensemble) constraining all the system with a harmonic potential (4.0 kcal mol^{-1} $Å^{-1}$) and the second one from 100 to 300 K (100 ps in the NPT ensemble) constraining just the protein backbone. Next, a 5 ns NPT equilibration step at 300 K was done initially applying the same constraining on the protein backbone but releasing the harmonic force constant by 1 kcal mol^{-1} $Å^{-1}$ at each 1 ns. Then, we carried out 5 μs of production at 300 K in the NPT ensemble for both C526A and iLOV systems. Two replica MDs were performed on each system, for a total sampling time of 20 μs.

Both minimization and MD simulations were performed using the Amber16 program employing the *ff14SB* [90] force field for protein. The parameters for the FMN chromophore was obtained from the literature [135]. In all MD simulations, the time step was set to 2 fs. For system temperature

and pressure control we employed a Langevin thermostat and an anisotropic barostat, both implemented in the Amber16. The particle-mesh Ewald algorithm [87] was used to describe the long-range electrostatic interactions. The MD analysis was performed by using both the *cpptraj* [88] module of AmberTools and locally developed tools. For trajectory visualization we employed the Visual Molecular Dynamics (VMD) software [89].

The network of hydrogen bonds around the FMN chromophore was investigated by computing distances between the FMN and H-bonding residues, as well as among these residues, along both replicas of each system. A principal component analysis (PCA) on these distances, followed by a clustering with the HDBSCAN algorithm [136, 137], was used to find different conformations of the FMN pocket (clusters). Ten structures were randomly extracted from each cluster to be employed in the following calculations. The interaction of nonpolar residues with the FMN ring was quantified by computing the overlap integral between the pseudo-electronic densities of the FMN ring and the residue sidechain:

$$\mathrm{Ovlp} = \int \mathrm{d}\mathbf{r}\tilde{\rho}_A(\mathbf{r})\tilde{\rho}_B(\mathbf{r}) \tag{5.1}$$

where the pseudo-electronic densities $\tilde{\rho}_X(\mathbf{r})$ $(X = A, B)$ were computed as a sum of Gaussian distributions centered on the heavy atoms of each fragment:

$$\tilde{\rho}_X(\mathbf{r}) = \sum_{i \in X} \frac{1}{\sigma_i\sqrt{2\pi}} e^{-\frac{(\mathbf{r}-\mathbf{r}_i)^2}{2\sigma_i^2}} \tag{5.2}$$

The standard deviations of the Guassian distributions, σ_i, were taken equal to the van der Waals radii of the elements.

5.2.2 QM/MM(Pol) calculations

The multiscale calculations, for both crystal structures and MD snapshots, were divided into four steps: (**i**) ground state (S_0) geometry optimization; (**ii**) vertical excitation energy calculation; (**iii**) excited state (S_1)

geometry optimization; and (**iv**) vertical emission energy calculation. The initial coordinates for step **i** was obtained from the crystal structure and from configurations extracted from the MD trajectories, and the optimized S_0 structures were used as input for step **iii**. In all steps, only the isoalloxazine group of FMN was included in the QM region. The calculations with the crystal structures were carried without the presence of the solvent. We applied the time dependent density functional theory (TD-DFT) to obtain the full transition densities in steps involving $S_0 \rightarrow S_1$ or $S_1 \rightarrow S_0$ transitions (steps **ii**, **iii** and **iv**). All multiscale calculations have been performed using a locally modified version of Gaussian G09 [138]. All geometry optimizations (steps **i** and **iii**) were performed at ONIOM(QM:MM) level [34, 35], with the QM subsystem allowed to move and the rest of the system (the ribityl tail of FMN and the protein) kept frozen. For the configurations obtained from MD, a solvation sphere of about 20 Å of radius was defined around the chromophore. The Na^+ and Cl^- ions were stripped from the solvent. The QM part was described at the B3LYP/6-31G(d) level for ground-state optimizations (step **i**) and TD-ωB97X-D/6-31+G(d) level for the excited state optimizations (step **iii**). The MM part was described by the same force field used for the MD simulation.

All excitation and emission calculations (steps **ii** and **iv**) were performed using a three-layer model (see Fig. 5.3): (1) the QM subsystem described at TD-ωB97X-D/6-31+G(d) level; (2) the protein, the water molecules and ions within 15 Å from FMN described using a polarizable MM model (from now on MMpol), (3) the water molecules and the ions in a shell of 15-30 Å from the chromophore treated at MM level. The polarizable first shell was described with the *pol12* AL Amber force field [52, 139].

We note that the latter functional gives significantly blue-shifted excitation energies with respect to experiments (about +0.4 eV). A better agreement with experiments was found for B3LYP but, with that functional, the characterization of the excited state was unclear due to artificial mixing. As here the goal was to obtain consistent excitation and emission energies, to predict Stokes shifts, we preferred to use ωB97X-D and obtain well charac-

terized excited states with excitation and emission energies of comparable accuracy, even if both are blue-shifted.

All excitation and emission calculations were performed by employing a polarizable embedding QM/MM approach (from now on QM/MMpol) [48]. By using this method we were able to describe the isoalloxazine group of FMN with a QM method and the surrounding environment as a set of classical point charges and atomic polarizabilities. As such, this approach allows the QM part and the classical part to mutually polarize. In the QM/MMPol calculations, the system was divide into three parts, the QM region, the MMPol region and the external MM region (not polarizable). All residues within 15 Å from FMN was included into the MMPol region. Residues beyond 15 Å and up to 30 Å from the chromophore were included into the MM region. The QM part was described by using the TD-DFT method at the ωB97X-D/6-31+G(d) level. For a graphical representation of the three-layer scheme employed, see Fig. 5.3.

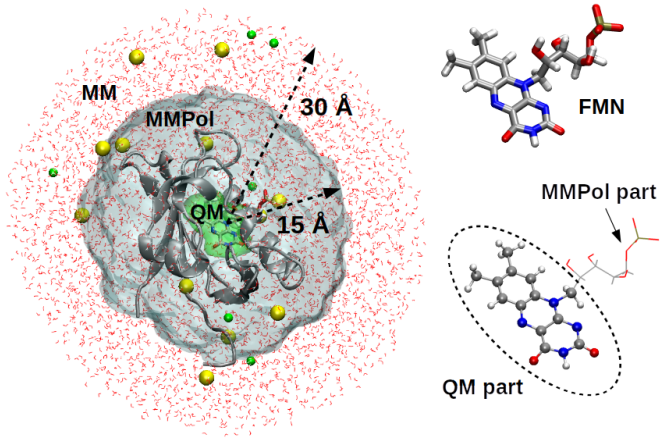

Figure 5.3: System partitioning employed in the QM/MMPol/MM calculations. Copyright: https://dx.doi.org/10.1021/acs.jpcb.0c10834.

5.2.3 Lineshape calculations

The homogeneous lineshape $\sigma(\omega-\omega_{01})$ of the flavin chromophores was computed using the second-order cumulant expansion in the displaced harmonic oscillator (DHO) formalism [140]:

$$\begin{aligned}\sigma(\omega-\omega_{01}) = \Re\int_0^\infty \mathrm{d}t \exp\left[i(\omega-\omega_{01})t + \right. \\ \left. + \int_0^\infty \mathrm{d}\omega \frac{J(\omega)}{\omega^2}\left(\coth(\frac{\beta\hbar\omega}{2})\cdot(\cos(\omega t)-1) - i(\sin(\omega t)-\omega t)\right)\right.\end{aligned} \tag{5.3}$$

where the spectral density (SD) $J(\omega)$ is defined in order to encode the vibronic coupling with all normal modes:

$$J(\omega) = \pi\sum_k S_k\omega_k^2\big(\delta(\omega-\omega_k) - \delta(\omega+\omega_k)\big) \tag{5.4}$$

where S_k is the Huang-Rhys factor along mode k. The Huang-Rhys factors were calculated on the crystal structures of both iLOV and C426A, by first computing the ground-state normal modes, in the same ONIOM scheme used

for the optimizations in step (i). The excited-state gradient computed at the B3LYP/6-31G(d) level of theory was then projected onto each normal mode to get the Huang-Rhys factor.

The final spectra were computed, for each cluster, by summing the contribution of N structures with excitation energies $\omega_{01}^{(j)}$ and transition dipoles $\boldsymbol{\mu}_{01}^{(j)}$:

$$A(\omega) = \omega \frac{1}{N} \sum_{j=1}^{N} |\boldsymbol{\mu}_{01}^{(j)}|^2 \sigma(\omega - \omega_{01}^{(j)}) \tag{5.5}$$

where $\sigma(\omega)$ is the absorption lineshape computed as described above.

5.3 Results and Discussion

5.3.1 Structural analysis

In our μs-long MD simulations we observed that both C- and N- termini are highly flexible in both C426A and iLOV domains, with up to 4Å RMSD from the crystal structure (Fig. 5.3). Excluding the termini, however, both C426A and iLOV showed a rigid protein backbone, with the average RMSD value around 2 Å with respect to the crystallographic structures (Figs. B3 and B4). No significant changes in the secondary structure were observed. In one replica, C426A showed increasing RMSD, up to 3 Å after ∼4000 ns. This deviation arises from a conformational change in a flexible loop between residues D477 and E481 (Fig. B3). As this loop is external to the protein and far from the binding pocket (Fig. 5.4a), we can exclude an influence of its conformation on the properties of the chromophore.

In order to achieve a clear description of the possible structural and dynamic specificities of the two systems which can be related to their different fluorescence behavior, we split the analysis in two parts: one focused on the H-bonding network which characterizes the FMN-binding pocket and the other investigating the specific residues involved in the "crowding and flipping" mechanism, illustrated in Fig. 5.2b,c.

As it regards the H-bonding network, a total of 22 distances (d1-22) were defined and analysed: 12 protein-FMN (d1-12), 8 protein-protein interactions (d13-20) nearby the FMN-binding pocket and 2 intra-FMN interactions (d21,d22). The full map of the H-bonding distances are reported in Fig. B5 of the Appendix B.

In general, most of the H-bonding interactions here analyzed showed to be stable in both C426A and iLOV during the whole MD simulations. In particular, all the interactions between the ribityl tail of FMN and side chains N425, R427, Q430 and R443 (d7-12) present in the crystal structures were conserved along the MD trajectories. The same occurred for the protein-protein interactions involving the same four residues (d17-20), for the two intra-FMN H-bonds (d21,22), as well as for the H-bonds between the isoal-

loxazine ring of FMN and the N458 side chain (d4,5).

On the other hand, differences with respect to the crystal structures have been found for the H-bonding interactions involving residues N390, N458, N468, A469, Q489 and A490 (d1-5,13-16) (see Fig. 5.4a,b). To give an example of these differences, in Figure 5.4c,d we compare the distributions of distances between FMN and either Q489 or N468 residues, in the two systems.

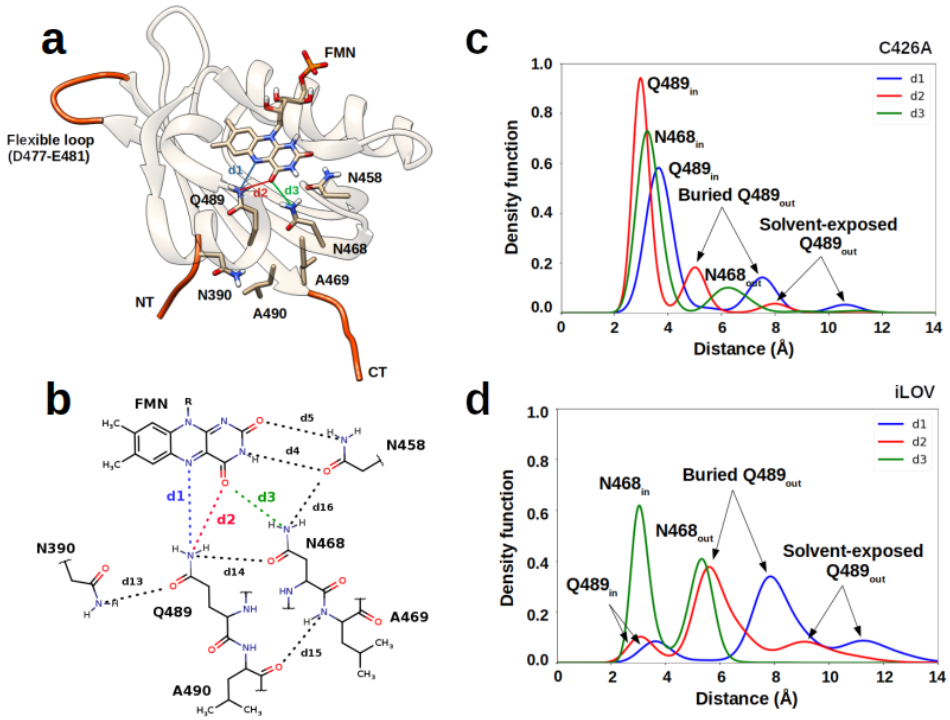

Figure 5.4: (a) Cyrstal structure of iLOV showing the conserved flexible residues in the FMN-binding site. In red color are also indicated the regions of the protein that showed great flexibility during our MD simulations (either in C426A or iLOV). NT and CT refers to C- and N-terminal ends. (b) Map of interactions indicating the distances employed in the H-bond analysis (d1-5,d13-d16). (c) and (d) Distributions obtained for FMN-Q489 and FMN-N468 H-bond distances (d1-3) for C426A and iLOV. Copyright: `https://dx.doi.org/10.1021/acs.jpcb.0c10834`.

To understand these findings we have to say that N468 and Q489 chains can adopt multiple conformations along the simulations, which correspond to different distances from FMN (Figure 5.4c,d). As it can be seen from the figures, some conformations are oriented towards the chromophore, with distances less than 4 Å; such conformations are referred to as *in*. Other conformations are instead either solvent-oriented or "buried", and present distances longer than 5 Å. Such conformations are referred to as *out*. Importantly, we have observed the occurrence of *out* conformations of Q489 after at least 300 ns for iLOV, and after more than 2 μs for C426A (Fig. B6–B7 in the Appendix B). This suggests that the crystal conformation is a local minimum in the free-energy landscape of these systems, which might have been stabilized by crystal packing. The long timescales needed to equilibrate the protein, however, suggest caution in interpreting the relative population of *in* and *out* conformations.

All these different conformations of the FMN pocket were clusterized on the basis of distances d1-5 and d13-16 via the HDBSCAN algorithm. The clustering algorithm is able to distinguish $Q489_{in}$ and several types of $Q489_{out}$ conformations (Fig. B6–B7 in the Appendix B). After excluding clusters with population $< 1\%$, we obtained three clusters for C426A (named C2, C5 and C7) and six clusters for iLOV (named C0, C1, C2, C3, C6 and C7). The larger number of iLOV clusters with respect to C426A arises from a greater population of the $Q489_{out}$ and $N468_{out}$ sidechain conformations. A comparison of the network of interactions of FMN with the residues of the binding pocket in the crystal structures and representative configurations for the different clusters are reported in Fig. 5.5 and Fig. 5.6, for C426A and iLOV, respectively. For the latter, only the three most populated clusters are shown while the other are reported in Fig. B8 in the Appendix B.

By analysing the different clusters we can notice that the $Q489_{out}$ and $N468_{out}$ orientations favor the entry of water molecules into the active site in both C426A and iLOV. These water molecules can replace Q489 and/or N468 in the H-bonding network with FMN and form a bridge between the FMN and the residues. In such a case, the Q489 side chain adopts a buried conformation

$Q489_{out}$ relatively closer to the FMN with respect to the solvent-exposed $Q489_{out}$ orientation. The two FMN-N458 H-bonds (d4,5), in turn, showed to be quite persistent in both C426A and iLOV (see Fig. 5.5, 5.6). No *out* orientations were observed for N458.

Here we will use "IN" to refer to the structures in which both $Q489_{in}$ and $N468_{in}$ occur simultaneously, similarly to what found in the crystal structures. And we will use "OUT" to refer to structures in which either $Q489_{out}$ or $N468_{out}$ are present. Using this classification, we observe that one cluster of C426A (C7) and one of iLOV (C1) are of type IN.

Concerning the OUT structures, our simulations indicate that the $Q489_{out}$ side chain can form H-bonds with N390 and N468

(d13 and d14 in Fig. B9). Intriguingly, substituting N390 with a serine was shown to improve photostability of the iLOV *in vivo*, probably by restricting the movement of Q489.[122] In addition, we observed interactions between $Q489_{out}$ and water molecules in the active site or from the bulk solvent when Q489 is solvent-exposed (see Figs. 5.5, 5.6 and B8).

Unlike Q489, $N468_{out}$ is never solvent-exposed, but it adopts a buried conformation in both systems. In most of the OUT configurations analyzed, the $N468_{out}$ conformation is characterized by water molecules bound to the FMN oxygen. Sometimes, hydrogen bond between FMN and N468 occurs through a bridging water molecule while an interaction between N458 and N468 is established (see C3 in Fig. 5.5).

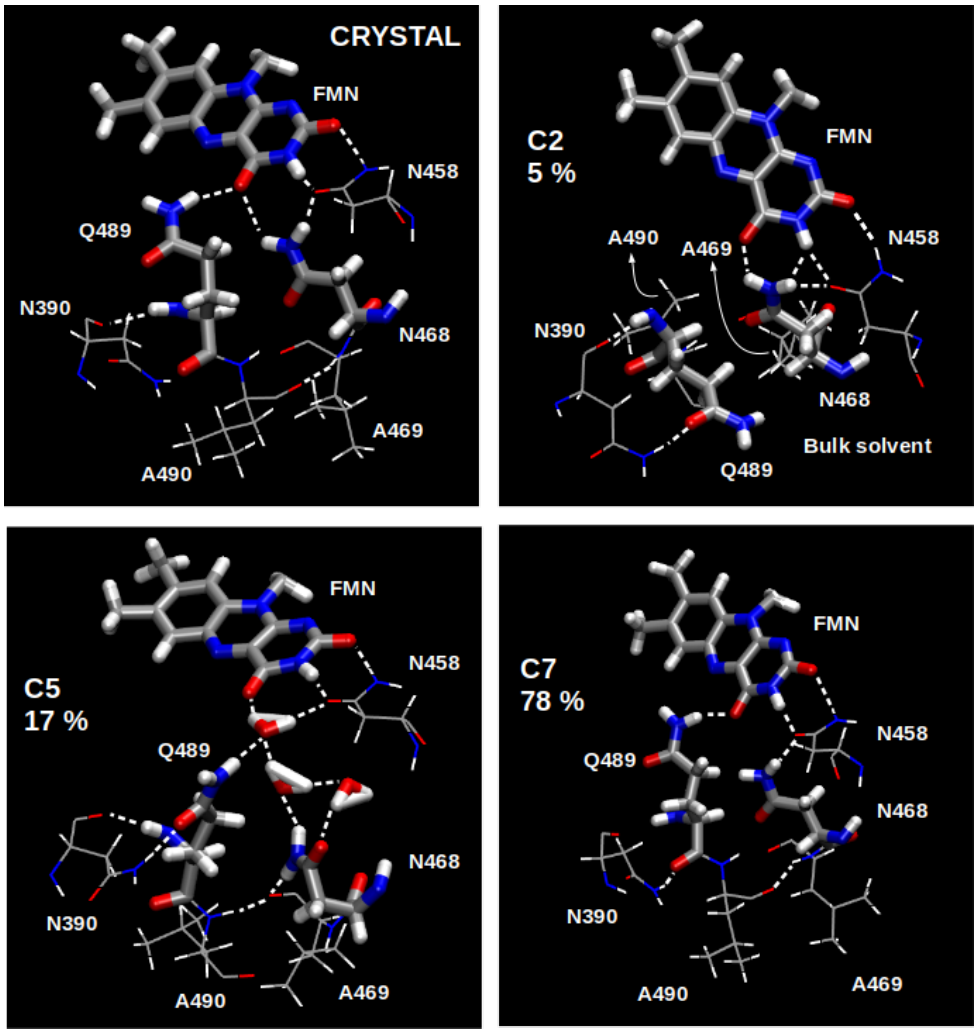

Figure 5.5: Comparison of the binding pocket in the crystal structure and in representative configurations for the three clusters obtained for C426A (C2, C5 and C7). The percentage values indicate the population of each cluster in the simulated trajectories. Copyright: `https://dx.doi.org/10.1021/acs.jpcb.0c10834`.

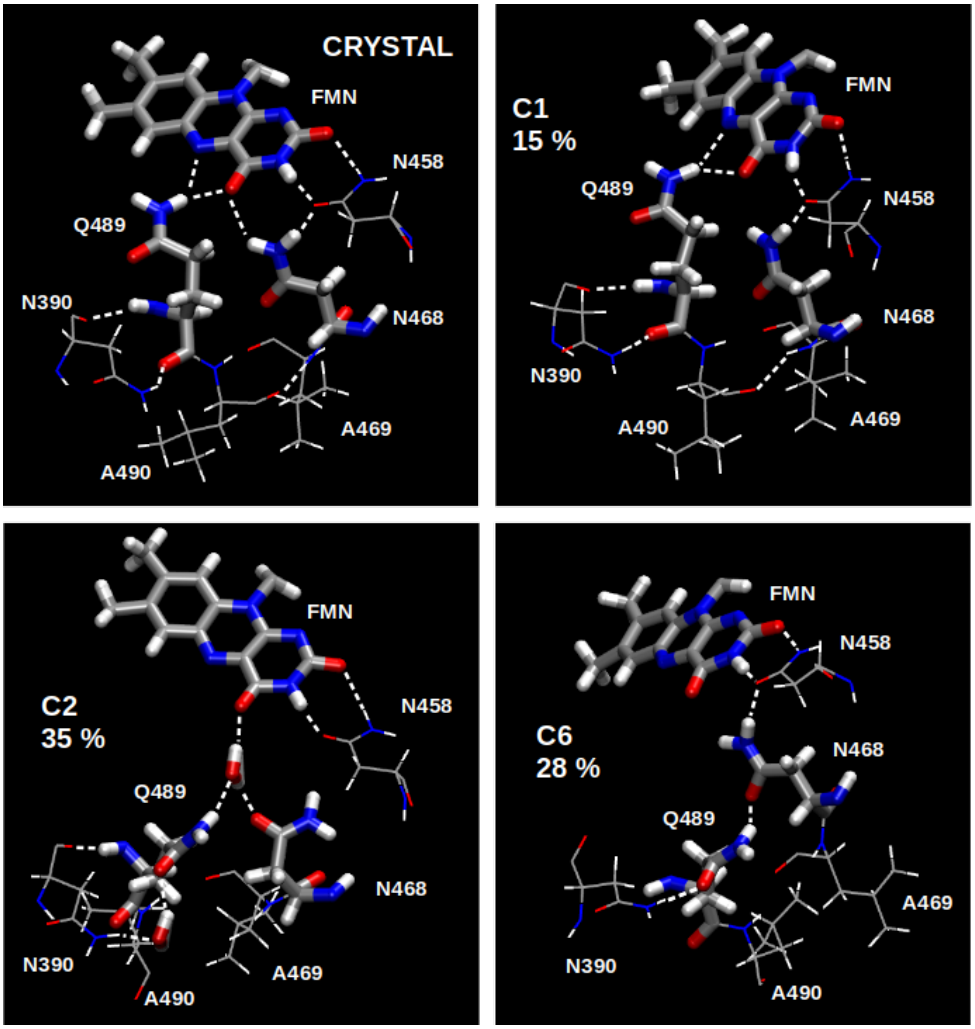

Figure 5.6: Comparison of the binding pocket in the crystal structure and in representative configurations for the three most populated clusters obtained for iLOV (C1, C2 and C6). The percentage values indicate the population of each cluster in the simulated trajectories. Copyright: `https://dx.doi.org/10.1021/acs.jpcb.0c10834`.

Previous MD studies of iLOV or natural LOV domains have also reported the change of orientation undergone by the glutamine (Q) in the active site [127, 131–133]. Studies focused on the spectroscopic optimization

of the LOV domain attempted to produce a red-shifted iLOV variant through the mutation Q489K, which puts a positively charged residue, lysine, close to the FMN chromophore [131]. Later, experiments and MD simulations showed that K489 is as flexible as glutamine, and prefers a *out* conformation [127]. MD studies on the LOV1 domain from *Chlamydomonas reinhardtii* also pointed out that the H-bonding interaction between FMN and Q120 side chain (equivalent to Q489 in iLOV) is also unstable, persisting for only 33 % of the simulation time (60 ns) [141]. Similarly, Lokhandwala et al. have observed buried and solvent exposed conformations for the Q204 side chains in MD studies involving a short LOV domain from *Trichoderma reesei* [142]. The $N468_{out}$ conformer (or its equivalent in protein sequence) was not reported in previous MD studies, probably due to the short time windows investigated by those simulations (30 to 60 ns). In our simulations, no significant conformational changes of the N468 side chain were observed before 300 ns, suggesting that long simulation times or enhanced sampling methods are needed to assess the conformation of the flavin-binding side chains.

Although largely reported in MD simulation studies, the flexibility of conserved residues in the FMN biding pocket has not been observed in the crystal structures, which consistently present a $Q489_{in}$ conformation [122]. As suggested before [131], the more compact $Q489_{in}$ conformation might be stabilized by crystal packing. The functional relevance of the dynamics of such residues in natural LOV proteins also remains unclear, but they may have implications for both decay pathways and signal transduction [141–143]. Regarding the artificial LOV domains, Khrenova et al. suggested that the Q489 side chain may be involved in both spectral and fluorescence-efficiency tuning [131].

As reported in the Introduction, Christie et al [122], proposed that the fluorescence increase in iLOV is mainly related to the optimized protein-chromophore van der Waals interactions induced by the “crowding and flipping” mechanism illustrated in Fig. 5.2b,c. As suggested from the crystal structures, the mutation of the serine (S394) with a tyrosine in iLOV results in a “crowding” of the chromophore local environment, whereas the

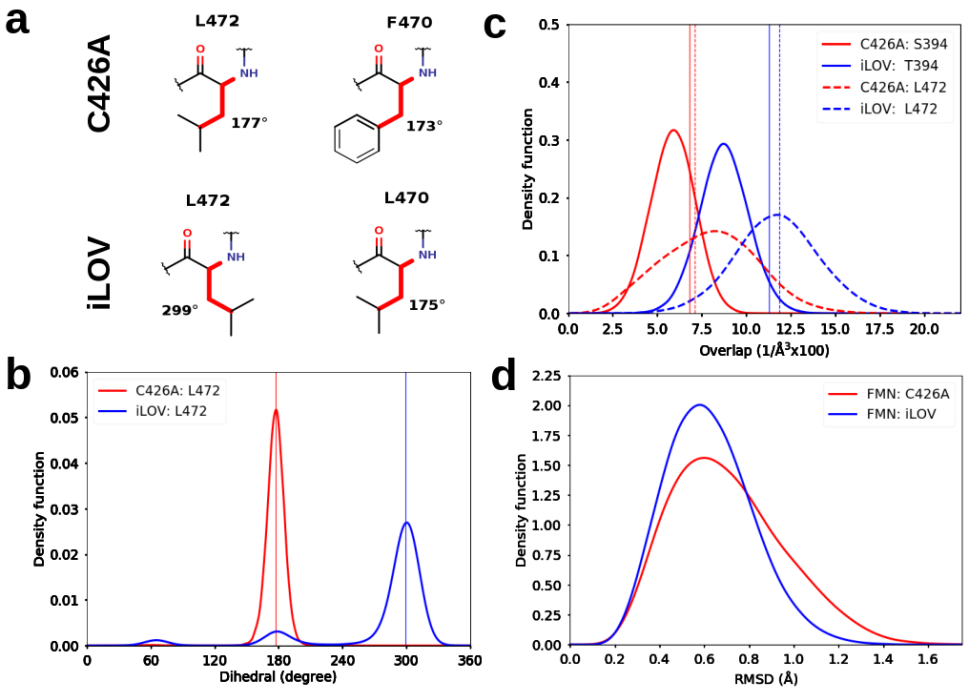

Figure 5.7: (a) Dihedral angles (red color) analyzed for C426A (top) and iLOV (bottom). The values in degree refers to the crystal structures. (b) L472 dihedral distributions for C426A (red) and iLOV (blue). The vertical lines indicate the values found in the crystal structures. The dihedral distributions for F470 (in C426A) and L470 (in iLOV) are available in Fig. B10 of the Appendix B. (c) Distributions of the overlap between the FMN and the indicated residues for both systems. The vertical lines indicate the ovelap values obtained from the crystal structures. (d) RMSD distribution for FMN isoalloxazine ring along DM simulations of C426A (red) and iLOV (blue). The backbone of the protein crystal structures was used as reference for the alignment (the N- and C-terminal residues were excluded). Copyright: `https://dx.doi.org/10.1021/acs.jpcb.0c10834`.

side chain of L472 undergoes a "flipping" as a consequence of the F470L mutation. We assessed the validity of the "flipping" mechanism in solution by monitoring the dihedral angles of the 470 and 472 apolar side chains present in the active-site of iLOV and C426A (Fig. 5.7). Leucine L472 remains essentially in the crystal conformation, with a rotation of about 120° in iLOV when compared to the single mutant (see Fig. 5.7, top right panel). Also the side chain of residue 470 (Phe in C426A and Leu in iLOV) reproduces the crystal structure geometry for both systems (see Fig. B10, of the Appendix B); however, L470 is somewhat more flexible than F470.

In summary, our MD simulations confirm that the F470L mutation stabilizes the flipping of the L472 residue in iLOV, because L470 has a smaller van der Waals volume compared to F470. This allows the L472 side chain to fit better in the flipped position, closer to the isoalloxazine ring.

To investigate the crowding effect, we calculated the van der Waals overlap (Eq. 5.1) between the FMN and the apolar residues in the positions 394 and 472 of both systems (Fig. 5.7 bottom left). The crowding of the active site in iLOV is induced cooperatively by S394T and F470L mutations, since both side chains of T394 and flipped L472 are closer to the chromophore, thus contributing to a better residue packing around the FMN. Such optimized packing could increase the stiffness of the chromophore, as indicated by the lower RMSD for the isoalloxazine ring obtained from C426A and iLOV trajectories (Fig. 5.7, bottom-right). In other words, the isoalloxazine ring in the active site of iLOV has less freedom of movement with respect to C426A. This supports the idea that in iLOV the presence of T349 and L472 contributes to an improved protein-chromophore interaction which reduces the mobility of the chromophore in the active site.

5.3.2 Spectroscopic analysis

The clusters obtained from the analysis of the H-bonding network were finally used in combination with multiscale QM/MMPol calculations, to evaluate the impact of H-bonding dynamics on absorption spectra and Stokes shifts. From each cluster, 10 configurations were extracted and used to com-

Table 5.1: Calculated vertical emission (EMI) and excitation (EXC) energies and stokes shifts (SS) for the crystal structure and the most populated clusters of C426A and iLOV. The weighted average (AVG) values are also indicated. For each cluster we also indicate its IN or OUT conformation and the corresponding population (%). EMI and EXC energies are expressed in eV and the stokes shifts (SS) in cm^{-1}.

				C426A				
	CRY	$C2_{OUT}$	$C5_{OUT}$	$C7_{IN}$	AVG			
Pop. (%)	-	5	17	78	-			
EXC	3.17	3.27	3.23	3.22	3.22			
EMI	2.78	2.88	2.84	2.82	2.83			
SS	3154	3203	3145	3192	3185			
				iLOV				
	CRY	$C0_{OUT}$	$C1_{IN}$	$C2_{OUT}$	$C3_{OUT}$	$C6_{OUT}$	$C7_{OUT}$	AVG
Pop. (%)	-	11	15	35	7	28	3	-
EXC	3.15	3.25	3.21	3.25	3.25	3.25	3.26	3.24
EMI	2.77	2.87	2.82	2.87	2.87	2.86	2.87	2.86
SS	3062	3059	3098	3090	3079	3089	3090	3087

pute the chromophore vertical excitation and emission energies employing the QM/MMPol method (see the Computational details section and Fig. 5.3).

The calculated vertical energies are reported in Table 5.1 for each cluster of both systems. In addition, for each system, we also report the the population-weighted average values. All clusters give similar results, but some configurations have a more blue-shifted excitation/emission energy. The most blue-shifted cluster presents $Q489_{out}$ and $N468_{in}$ side chain conformations, with $Q489_{out}$ in its solvent-exposed conformation and no water molecules bound to the chromophore (cluster C2 in C426A).

By comparing the different clusters, we estimate that the loss of Q489-FMN H-bond causes a blue-shift of about 400 cm^{-1} in the excitation energy. A parallel quantification for the N468-FMN H-bond was not possible because, as mentioned before, in the frames in which we observe the occurrence of $N468_{out}$, the N468 side chain is always replaced by water molecules.

The absorption lineshapes were computed separately for each cluster as detailed in Section 5.2.3. The spectra of each cluster are reported in Fig. 5.8,

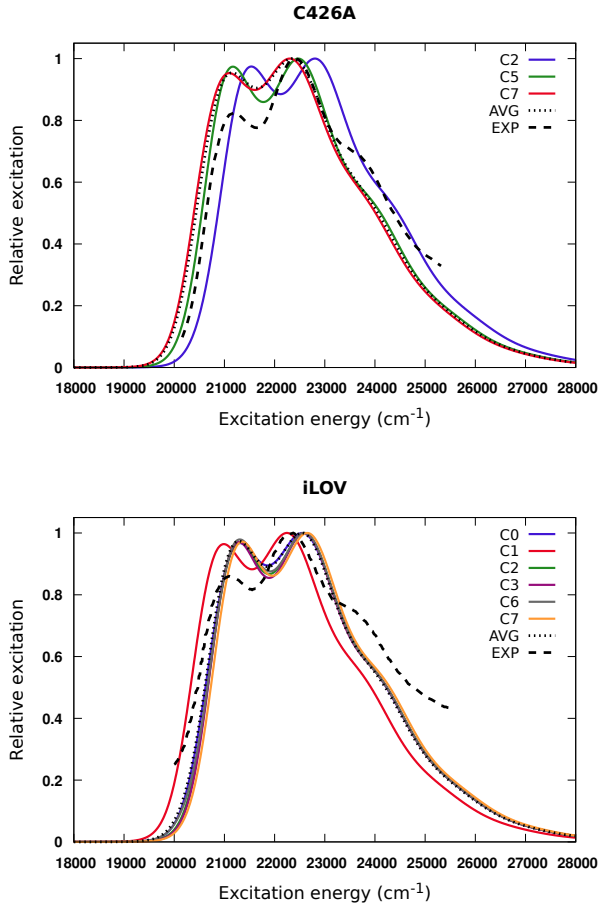

Figure 5.8: Absorption spectra for C426A (top) and iLOV (bottom) computed from MD using 10 frames for each cluster. The figures show the average spectra for each cluster (solid lines), the weighted average spectra (AVG) for all clusters (dotted line), as well as the experimental absorption spectra (EXP) obtained by Chapman and coworkers [121] (dashed line). All spectra intensities were noramlized so that the maximum is 1. All computed spectra were shifted by -3500 cm^{-1} to match the maximum absorption wavenumber of C426A. Copyright: `https://dx.doi.org/10.1021/acs.jpcb.0c10834`.

along with the population-weighted average and with the experiments. Owing to the large population of *in* conformations, the calculated average C426A

spectrum is essentially the same as the one of cluster C7. On the contrary, in iLOV the average spectrum is determined by the *out* clusters, which are all very similar, and account for ∼85% of the trajectory.

The absorption lineshapes compare well with experiments, reproducing the two main vibronic bands at ∼21000 and ∼23000 cm^{-1}, as well as the shoulder at higher energies. We can thus conclude that the broadening of the absorption band is mainly vibronic in nature, and that the modulation of pigment-protein interactions has a relatively small effect on the spectrum. While the experimental absorption spectra are very similar for the two proteins, our calculated spectra differ slightly by a small blue-shift for iLOV. We ascribe this difference to the different cluster populations observed in the two sets of MDs. Nonetheless, the spectra of different clusters are shifted by ∼300 cm^{-1}, less than the vibronic broadening, suggesting that the IN and OUT conformations could be populated in both C426A and iLOV, without influencing the absorption lineshape. Our results suggest that the absorption spectrum of C426A and iLOV is likely the result of several pocket conformations, comprising both *in* and *out* conformations for the Gln (Q) and Asn (N) side chains.

Regarding the Stokes shift (SS) values (Table 5.1), no significant difference was observed among the clusters of each system. However, the SS obtained in C426A clusters is systematically 80–100 cm^{-1} greater than that obtained for iLOV clusters. The Stokes shifts computed on the crystal structures present the same trend for the two systems, with values comparable to the MD ones. From these data, it is clear that the Stokes shift does not depend on the dynamic H-bonding network, but it is apparently controlled by the mutations S349T and and F470L discussed above. By hampering the mobility of the isoalloxazine ring, the "crowding" effect slightly reduces the extent of intermolecular reorganization experienced by the chromophore after excitation to the S_1 state. We speculate that the same mechanism also contributes to slow down the internal conversion of FMN towards the ground state.

Intriguingly, the measured fluorescence spectra of C426A and iLOV,

while very similar [121], show some non-negligible differences (Fig. B11). Indeed, the fluorescence spectrum of iLOV is narrower and slightly blue shifted, in agreement with the Stokes shifts calculated in this work.

5.4 Conclusions

In this study we analyzed the microsecond time-scale dynamics of two LOV-based fluorescent protein variants, namely C426A and iLOV, by means of MD simulations. We have investigated the dynamics of the H-bonding network of conserved residues in the FMN binding pocket of the two systems and found a significant flexibility especially for the glutamine (Q489) and asparagine (N486) side chains. These residues adopt different conformations, which interchange in time scales in the order of 200-300 ns. These findings stress the importance of a proper sampling for any study aimed at optimizing LOV-derived fluorescent proteins [134].

Our simulations have confirmed the "flipping and crowding" effect induced in iLOV by the additional mutations and revealed its mechanism of action. In particular, the crowding is cooperatively triggered by S394T and F470L mutations. The former leads to a direct packing effect, while the latter allows the conserved L472 side chain to change conformation and fit closer to the chromophore. On the other hand, we have shown that these mutations, and the resulting differences in the composition and flexibility of the binding pockets, are not reflected in significant shifts of the excitation and emission energies, in agreement with the similarity of the spectra measured for the two systems. However, a small but consistent reduction was found in the Stokes shift, moving from C426A and iLOV. This suggests that the reduction induced by F470 and S394T mutations in the mobility of the isoalloxazine ring of FMN also reduces the intermolecular reorganization experienced by the chromophore after excitation. Here, we hypothesize that the same mechanism could also contribute to slow down the internal conversion of FMN towards the ground state and to improve the fluorescence of iLOV.

Chapter 6

Conclusions

In this Thesis, we have presented a computational investigation of two different pigment-protein complexes with the aim of revealing the intriguing connection between the dynamics of the protein and the response to light of its chromophoric unit. The two investigated systems are examples of the two main classes of light-sensitive proteins, namely the ones based on a multichromophoric aggregate and those using a single chromophore.

As an example of multichromophoric systems, we have studied the light harvesting complex (LH2) of purple bacteria and tried to explain the observed change in the spectroscopic properties of the complex in response to high-light (HL) and low-light (LL) conditions. While, the available crystallographic data show that the HL-LH2 and its LL analog present exactly the same multichromophoric structure, a significant shift is observed in the corresponding absorption spectra. In all the different spectroscopic forms, the same two circular rings, made of 9 and 18 BChls respectively, characterize the multichromophoric aggregate with the only significant difference being the different local environment around the BChls of the 18-meric ring. By combining molecular dynamics with multiscale excitonic calculations we could show that the spectroscopic change is controlled by an interplay of protein-pigment interactions and dynamics. In particular, we found that the H-bond losses in the LL complexes (due to specific mutations) induces a

blue-shift in the excitation of the BChl and reduces the correlation in the BChls of the 18-meric ring. This impacts the interchromophore orientation and reduces the electronic coupling between. The combination of these two effects (local excitations and electronic couplings) explain the observed spectroscopic differences between HL and LL antenna complexes, showing the delicate connection between the protein dynamics and the excitonic properties of the multichromophoric aggregate.

As an example of single-chromophore system, we have analyzed the flavin-based system known as LOV (light, oxygen, and voltage) domains. In particular, we investigate two synthetic LOV systems obtained through protein engineering: one single-mutant (C426A) and another one with five additional mutations (iLOV). The latter has become one of the most popular fluorescent biomarkers in cell imaging studies thanks to its highly enhanced fluorescence and the increased photostability. In the literature, it was suggested that such fluoresce improvement observed in iLOV is manly related to optimized protein-chromophore interactions, through a mechanism that they called "crowding and flipping". The goal here was to combine a molecular-dynamics based exploration of the configurational space of the two mutants with multiscale calculations to evaluate the impact of the protein dynamics in the absorption spectra and Stokes shifts (SS) of each mutant. The obtained results confirm the proposed model showing that the mutations close to the active-site induce an improved protein-chromophore contact and contribute to the flavin stabilization.

From the two studies conducted in this Thesis, we can conclude that a realistic description of pigment-protein complexes cannot be achieved without accounting for the dynamics of the system. In this context, classical MD simulations become essential to sample the configurational space of such systems, which can be very large and contain hundreds of thousands of atoms. However, the simulation time required to achieve structural stability does not depend only on the size of the system, but also on the intrinsic characteristics of each photoresponsive protein and micro-second simulations are generally required for a proper sampling. On the other hand, we have shown that mul-

tiscale approaches based on the combination of a quantum chemical approach and a classical MM model can represent a very effective tool to describe the electronic processes undergone by the chromophoric unit within the protein complexes. However, this tool has to properly include the effects induced by the dynamics in the network of interactions of the chromophore(s) and the protein matrix.

Appendix A

Supporting information for Chapter 4

In this appendix are provided additional figures and tables as support information for Case study I: LH2 complexes from purple bacteria.

A.1 Additional figures

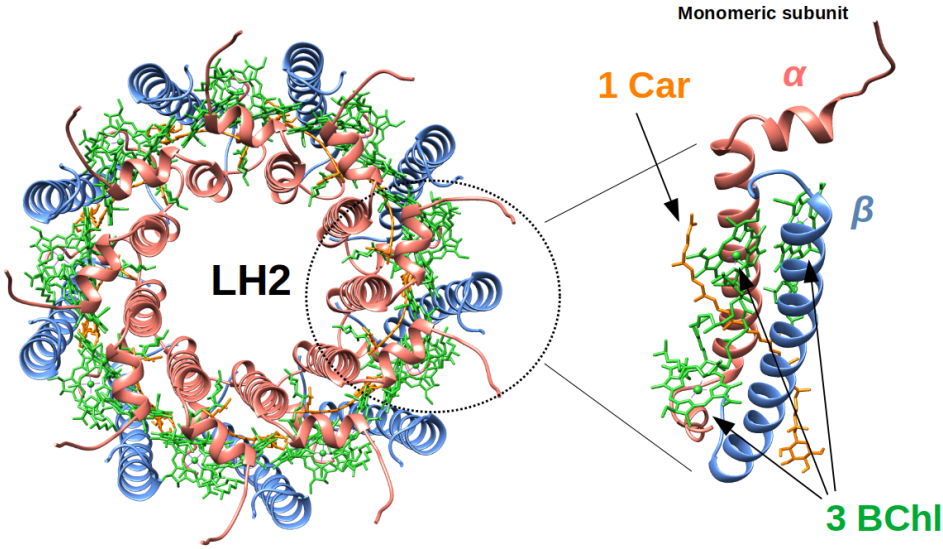

Figure A1: The structure of the purple bacteria LH2 complex. The whole protein has a circular shape and is formed by 9 subunits (nonameric structure). Each subunit (or monomer) is composed of two peptide chains (α and β) bound to one carotenoid (Car) and three BChls.

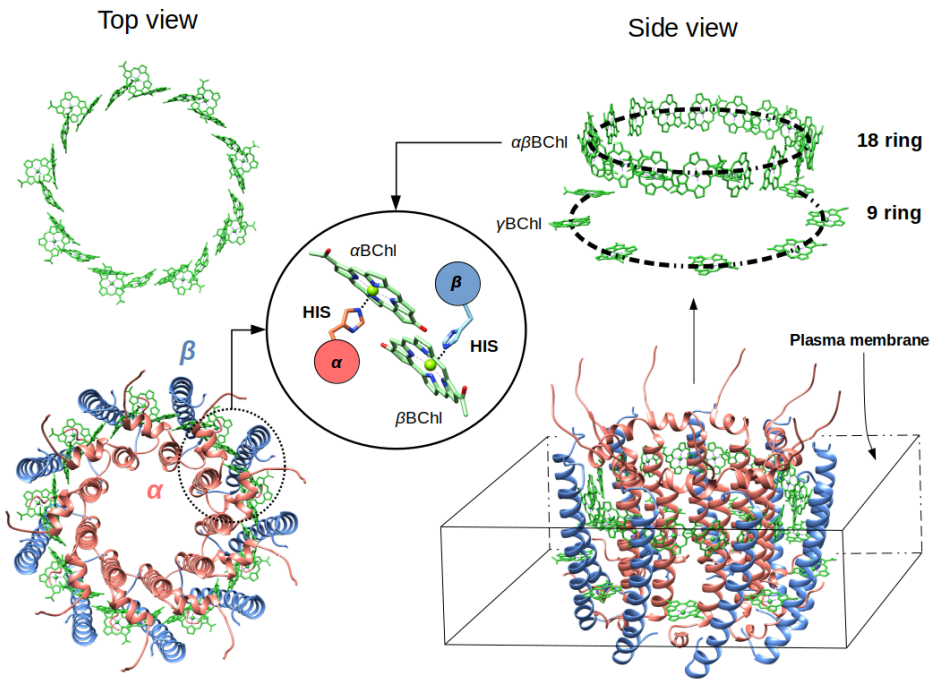

Figure A2: The BChl molecules form two rings in the nanomeric complex, an upper and a lower ring with 18 and 9 BChl molecules, respectively. In the 18 ring the Bchls are coordinated by a histidine residue (HIS) from the α chain or from the β chain, thus being called αBChl or βBChl (αBChls and βBChls are collectively named as $\alpha\beta$BChl).

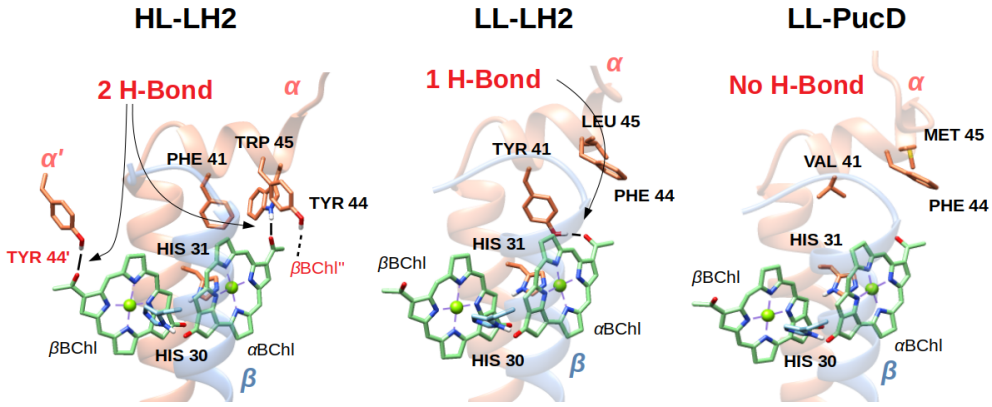

Figure A3: Protein environment of $\alpha\beta$BChl in the three complexes studied. From HL-LH2 to LL-LH2, we can see that BChl interacts with Tyr 41 instead of Trp 45 and βBChl loses a H-bond. In HL-LH2 complex the H-bond between Tyr 44 and BChl is an inter-monomer interaction so that the Tyr 44' residue belongs to the adjacent chain (α'). The LL-PucD complex has no protein-BChl H-bounds. The microenvironment of γBChl molecules is almost the same in the three complexes (not shown).

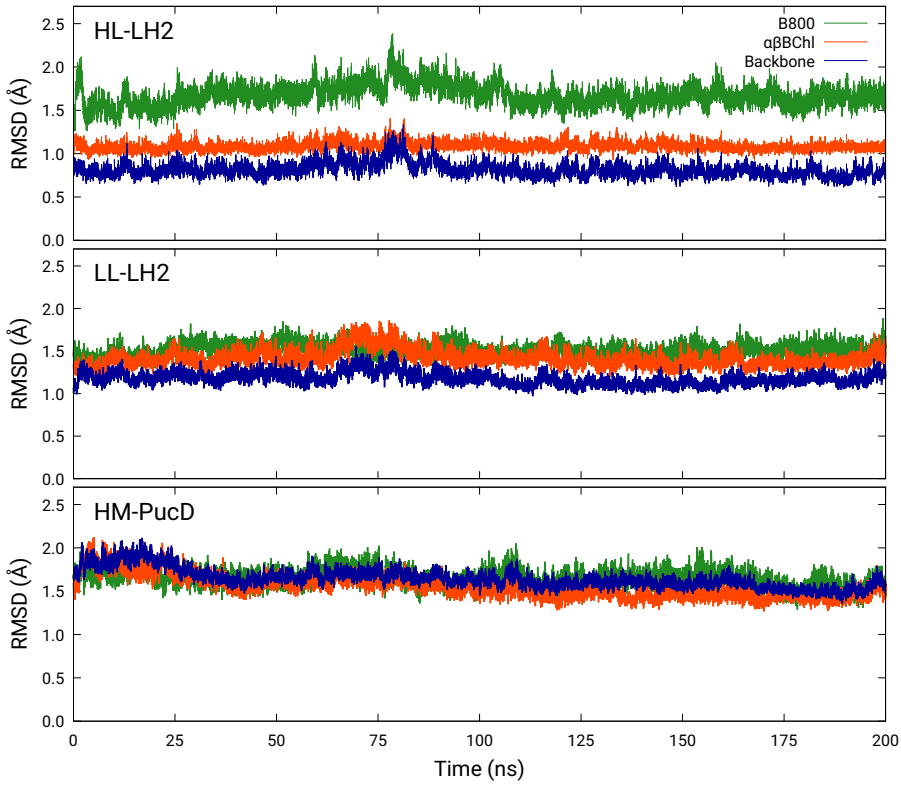

Figure A4: RMSD plot for the backbone of LH2 α-helices (blue), α and βBChl (orange) and B800 (green), with respect to the crystal structures (or Homology Modelling structure for HM-PucD). The selected atoms from BChls were the same as the QM part in the multiscale analysis (See Fig. 4.2 from the main text). Copyright: `https://doi.org/10.1039/c9sc02886b`.

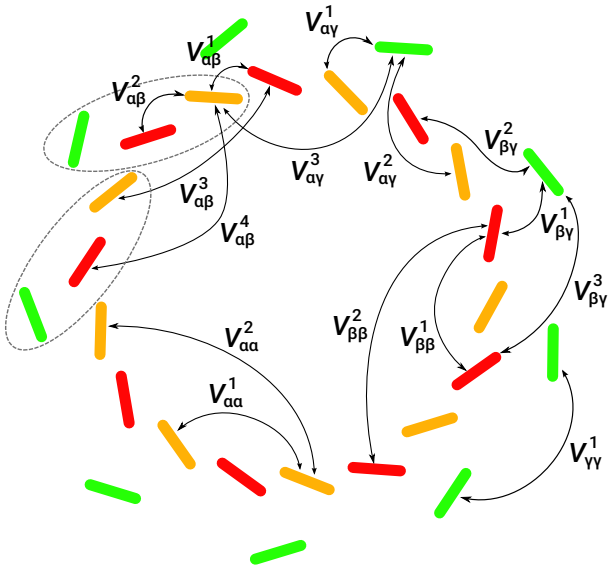

Figure A5: Schematic representation of the electronic coupling definitions. The BChls are highlighted using the following color code: yellow for αBChl, red for β and green for B800 (for brevity defined as γ). The superscript ranks the proximity according to inter-pigment distance. The dashed ellipses highlight two subunits of the protein. Copyright: `https://doi.org/10.1039/c9sc02886b`.

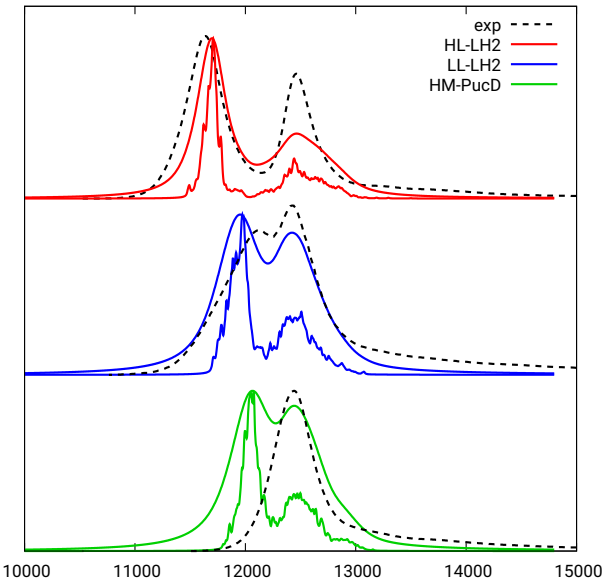

Figure A6: Plot showing the sum of stick spectra computed on individual frames of molecular dynamics, their convolution obtained with Lorentzian functions and the experimental spectra (as black dashed lines). The computed spectra are shifted by -1247 cm^{-1} to match the experimental B800 band. Moreover, we used the procedure described in Ref. 86 to remove the inhomogeneous broadening from the internal degrees of freedom of the BChls. Copyright: `https://doi.org/10.1039/c9sc02886b`.

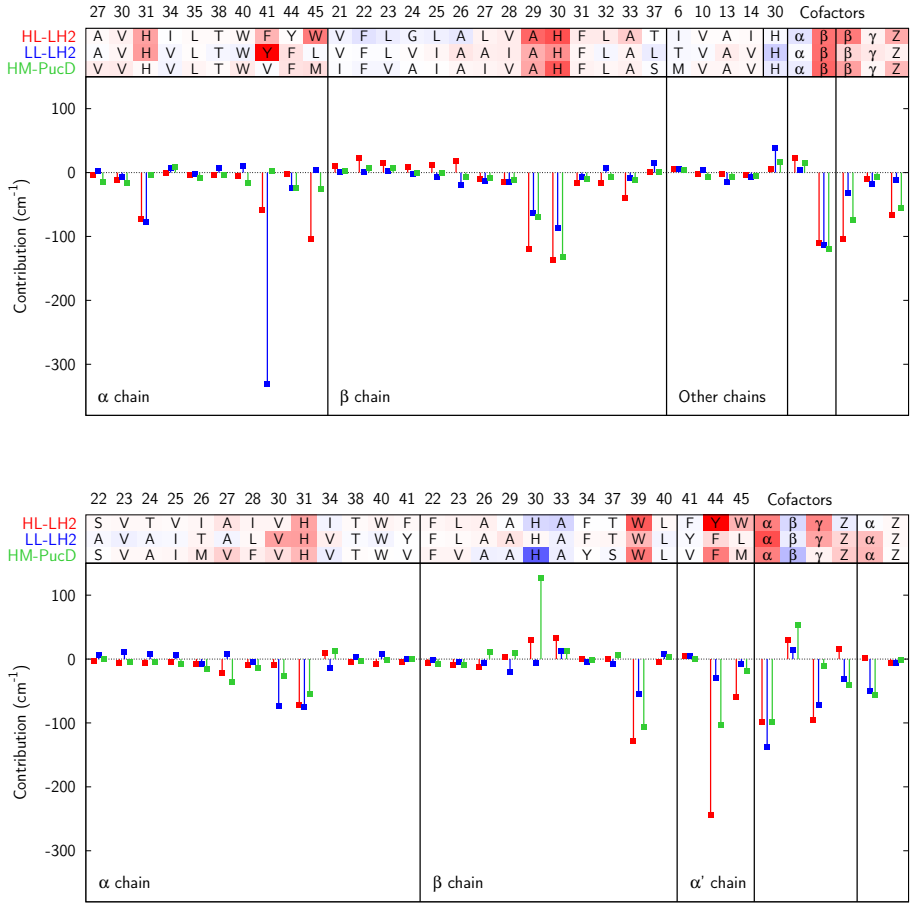

Figure A7: Contributions to the BChl transitions from all the residues within 6 Å from them (top: αBChl, bottom: βBChl). Copyright: `https://doi.org/10.1039/c9sc02886b`.

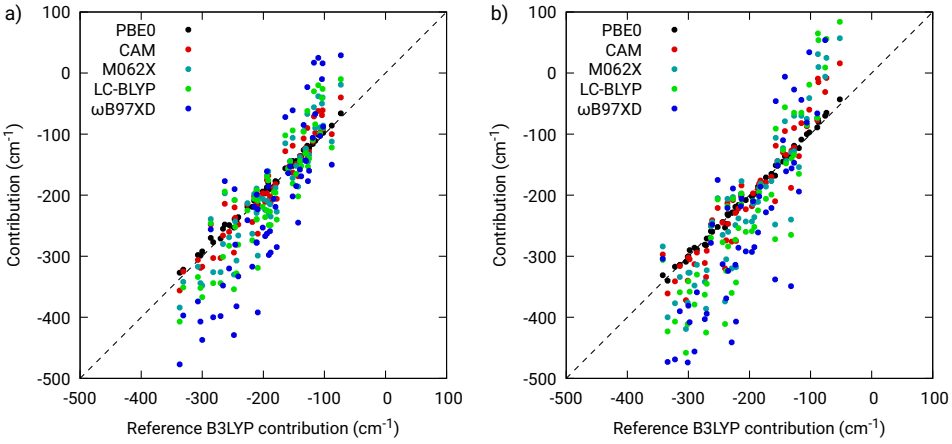

Figure A8: Benchmark of several different functionals against the B3LYP results. The measured property is the effect of the H-bonding residue on the BChl Q_y excitation energy. The results are computed on 20 structures extracted from the MD trajectory. The plot a) refers to calculations in which the H-bonded residue is treated at the MMPol level whereas the plot b) refers to calculations in which the residue is treated at QM level. Copyright: `https://doi.org/10.1039/c9sc02886b`.

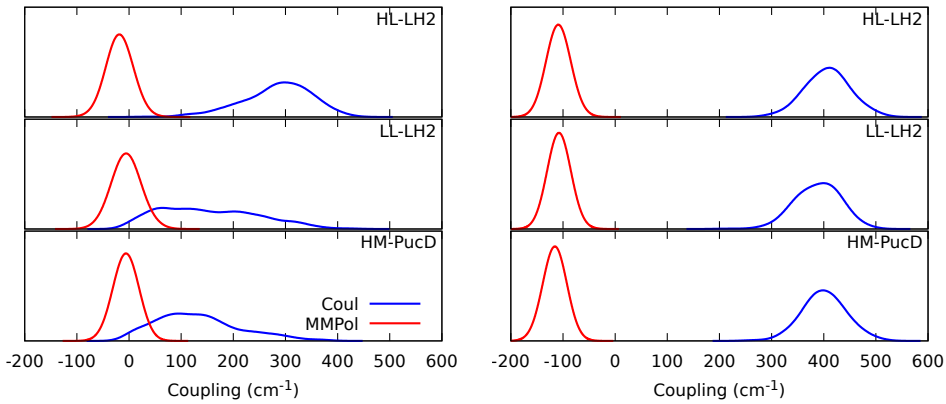

Figure A9: Distributions of the coupling components along the MD trajectory, left: $V^1_{\alpha\beta}$, right: $V^2_{\alpha\beta}$. The distribution of the Coulomb component (V_{Coul}) is drawn in blue, that of the environmental component (V_{MMPol}) in red. Copyright: `https://doi.org/10.1039/c9sc02886b`.

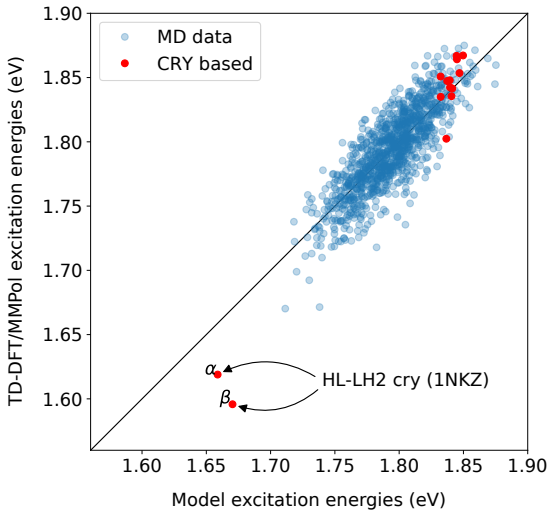

Figure A10: Results of the regression of Q_y excitation energies using the bond lengths of the macrocycle ring ($R^2_{\text{adj.}} = 0.68$). Blue dots represent excitation energies calculated/predicted on the MD structures, and used to fit the model. Red dots represent all the crystal-based structures (including the full and the constrained optimization structures), which were not included in the fit, and are here used as a test sample. The outliers indicated by the arrows correspond to the unrelaxed crystal structures of HL-LH2. All excitaton energies have been computed from the isolated BChls with geometries coming from the MD and the crystal structures. Copyright: `https://doi.org/10.1039/c9sc02886b`.

A.2 Additional tables

Table A1: Average of the equivalent couplings and standard deviations computed along 50 frames of MD. A schematic representation of the nomenclature is shown in Fig. A5. All the missing couplings are between distant BChls and are treated as 0 cm^{-1}.

Type	HL-LH2		LL-LH2		HM-PucD	
	V (cm^{-1})	σ (cm^{-1})	V (cm^{-1})	σ (cm^{-1})	V (cm^{-1})	σ (cm^{-1})
$V^1_{\alpha\beta}$	266	55	149	89	127	74
$V^2_{\alpha\beta}$	298	35	281	37	285	35
$V^3_{\alpha\beta}$	18	2	14	4	12	3
$V^4_{\alpha\beta}$	17	1	15	2	15	2
$V^1_{\alpha\alpha}$	-62	5	-66	7	-66	6
$V^2_{\alpha\alpha}$	-9	1	-9	1	-9	1
$V^1_{\beta\beta}$	-48	6	-30	12	-22	10
$V^2_{\beta\beta}$	–	–	-5	2	-3	1
$V^1_{\alpha\gamma}$	38	21	36	4	35	4
$V^2_{\alpha\gamma}$	-15	13	-13	2	-13	2
$V^3_{\alpha\gamma}$	-2	2	-6	2	-5	2
$V^1_{\beta\gamma}$	-8	4	-4	6	-4	5
$V^2_{\beta\gamma}$	-5	8	-1	3	1	3
$V^3_{\beta\gamma}$	9	7	8	1	8	1
$V^1_{\gamma\gamma}$	-34	25	-29	4	-28	4

Table A2: Top: Average of the equivalent charge transfer state energies and standard deviations. Bottom: Average of the equivalent charge transfer couplings and standard deviations. These data are computed along 10 frames of MD

Type	HL-LH2		LL-LH2	
	E (cm^{-1})	σ (cm^{-1})	E (cm^{-1})	σ (cm^{-1})
E^1_{CT} $(\alpha \rightarrow \beta)$	20846	1460	23538	2189
E^1_{CT} $(\beta \rightarrow \alpha)$	21790	1539	20724	1958
E^2_{CT} $(\alpha \rightarrow \beta)$	22412	1606	22783	2063
E^2_{CT} $(\beta \rightarrow \alpha)$	23752	1401	22059	1924
Type	V (cm^{-1})	σ (cm^{-1})	V (cm^{-1})	σ (cm^{-1})
$V^1_{CT}(\alpha \rightarrow \beta, \alpha^*)$	399	165	146	198
$V^1_{CT}(\alpha \rightarrow \beta, \beta^*)$	363	233	190	206
$V^1_{CT}(\beta \rightarrow \alpha, \alpha^*)$	420	236	236	240
$V^1_{CT}(\beta \rightarrow \alpha, \beta^*)$	475	171	197	219
$V^2_{CT}(\alpha \rightarrow \beta, \alpha^*)$	170	119	369	165
$V^2_{CT}(\alpha \rightarrow \beta, \beta^*)$	108	156	191	215
$V^2_{CT}(\beta \rightarrow \alpha, \alpha^*)$	107	150	188	216
$V^2_{CT}(\beta \rightarrow \alpha, \beta^*)$	162	113	380	169

Table A3: Bright excitonic states (cm^{-1}) computed from the average excitonic Hamiltonian and from average over spectra (Fig. A6). The values reported in parenthesis are computed with the inclusion of CT states.

	HL-LH2	LL-LH2	HM-pucD
Average Hamiltonian			
k= ±1	12906 (12821)	13183 (13133)	13281 (13230)
k= ±8	14004 (14000)	14039 (14032)	14091 (14086)
B800	13734 (13733)	13649 (13649)	13724 (13724)
Average over spectra			
k= ±1	12908	13151	13262
B800	13660	13619	13641

Table A4: Values of the acetyl dihedral angle for different structures (crystal structure Cry, full optimization fOpt and MM optimization MM-Opt).

	BChl	Cry	fOpt	MM-Opt
HL-LH2	α	20	15	8
	β	-25	-20	-5
LL-LH2	α	-30	-13	-6
	β	-36	-32	-6

Table A5: Site energies (cm^{-1}) from constrained geometry optimization. The values in parentheses refer to calculations without the effect of the MMPol.

	α	β
Structure	HL	
Cry	12253 (13057)	12010 (12871)
cOpt	14050 (14833)	13865 (14658)
fOpt	13978 (14800)	13954 (14853)
Structure	LL	
Cry	14325 (15036)	14127 (14804)
cOpt	14123 (14874)	14034 (14426)
fOpt	14032 (14859)	14154 (14899)

Appendix B

Supporting information for Chapter 5

In this appendix are provided additional figures as support information for Case study II: LOV-based fluorescent proteins.

B.1 Additional figures

```
WT          IEKNFVISDPRLPDNPIIFASDSFLELTEYSREEILGRNCRFLQGPETDQATVQKIRDAI 60
C426A       IEKNFVISDPRLPDNPIIFASDSFLELTEYSREEILGRNARFLQGPETDQATVQKIRDAI 60
iLOV        IEKNFVITDPRLPDNPIIFASDGFLELTEYSREEILGRNARFLQGPETDQATVQKIRDAI 60
            *******:**************.***************.********************

WT          RDQREITVQLINYTKSGKKFWNLFHLQPMRDQKGELQYFIGVQLDGSDHV   110
C426A       RDQREITVQLINYTKSGKKFWNLFHLQPMRDQKGELQYFIGVQLDGSDHV   110
iLOV        RDQRETTVQLINYTKSGKKFWNLLHLQPVRDQKGELQYFIGVQLDGSDHV   110
            ***** *****************:****:********************
```

Figure B1: Multiple alignment between the protein sequences of the wild-type LOV2 domain (WT), C426A and iLOV. For clarity, residues are numbered according to its position in the simulated protein sequence (considering I387 as the first residue), the same sequence synthesized and expressed by Christie and coworkers. Copyright: `https://dx.doi.org/10.1021/acs.jpcb.0c10834`.

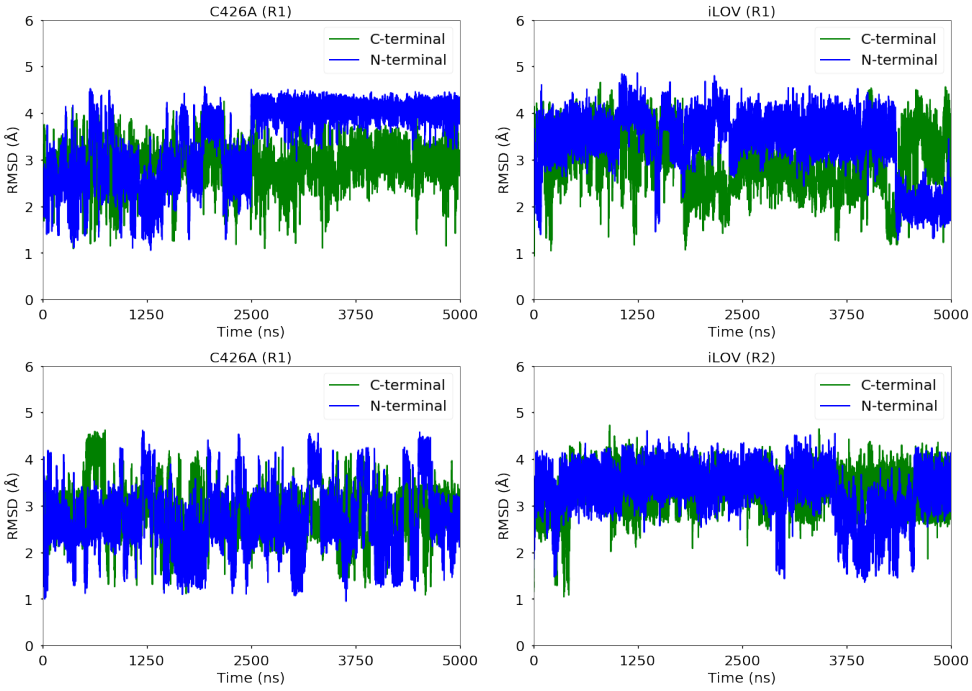

Figure B2: RMSD plots for the C-terminal (residues 1-4) and N-terminal (residues 106-110) extremities of C426A (right) and iLOV (left). R1 and R2 indicates the two MD replicas performed for both systems. This analysis was performed by using the protein crystal structures as reference for the structural fitting (just residues 5-105 were considered for the fitting). Copyright: `https://dx.doi.org/10.1021/acs.jpcb.0c10834`.

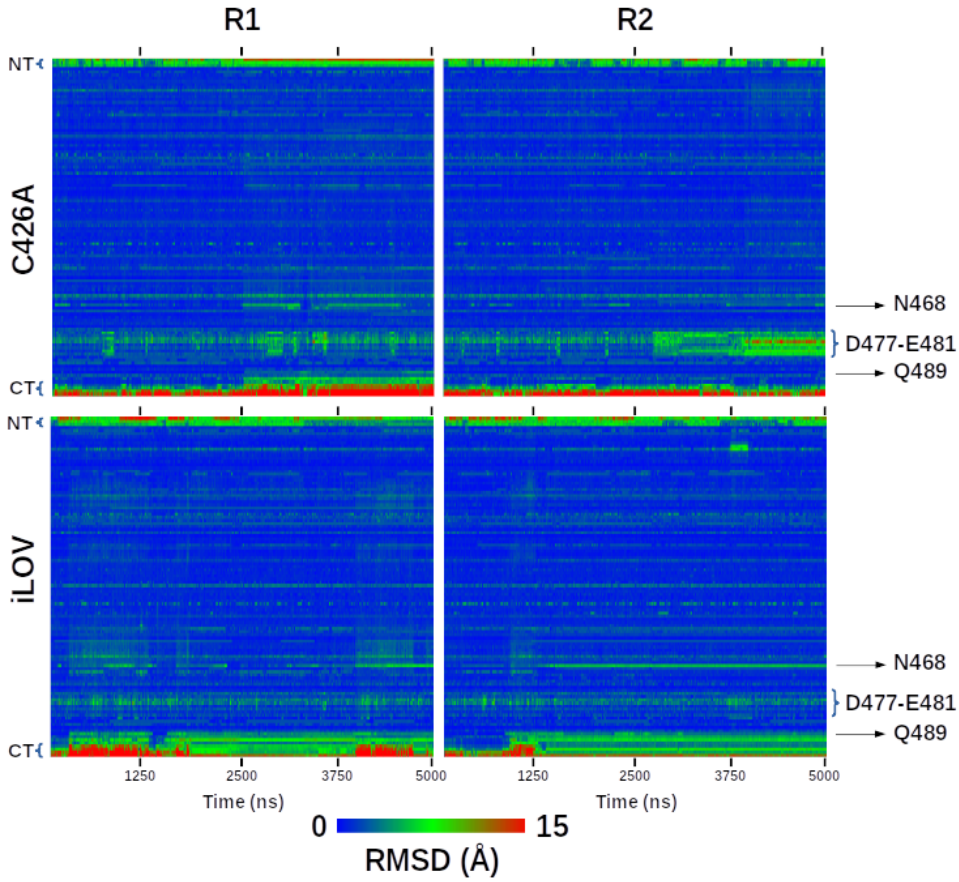

Figure B3: Per-residue RMSD calculated for C426A (top) and iLOV (bottom) employing both replicas R1 (left) and R2 (right). The crystal structures were used as reference for the fitting. Each plot shows the obtained results for 5000 ns (5 μs) of simulation (2500 frames were employed in this analysis). Are indicted the position of residues N468 (N82) and Q489 (Q103), as well as the flexible the loop between residues D477 and E481 (D91 and E95). NT and CT indicates, respectively, the N- and C-terminal ends. The regions in red have RMSD values equal to or greater than 15 Å. Copyright: `https://dx.doi.org/10.1021/acs.jpcb.0c10834`.

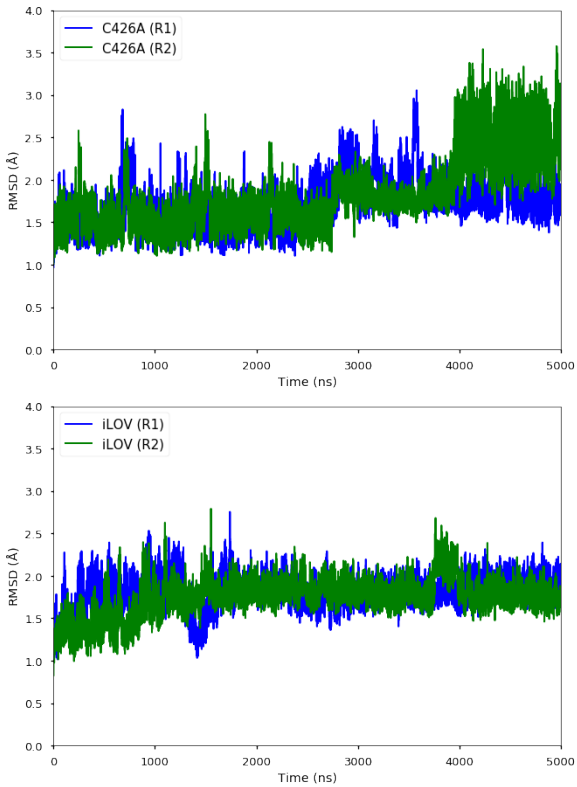

Figure B4: Protein backbone RMSD analysis of C426A (top) and iLOV (bottom) along each MD replica (R1 and R2). Only residues in positions from 5 to 105 in the simulated protein sequence were considered so that the C- end N-terminal ends were excluded from this analysis. Copyright: `https://dx.doi.org/10.1021/acs.jpcb.0c10834`.

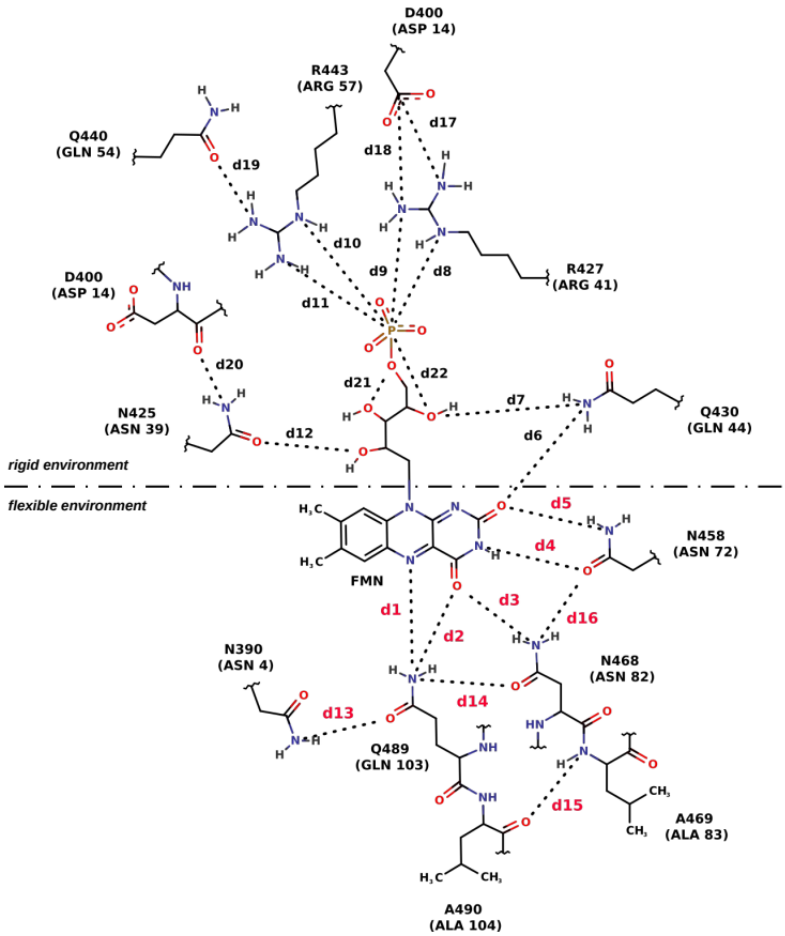

Figure B5: Map of interactions for some important H-bonds present in both C426A and iLOV systems. We monitored 22 distances (d1-22) to study the stability of protein-FMN (d1-12), protein-protein (d12-20), as well as intra-FMN interactions (d21,22). Residue numbering is based on phot2 sequence, while the numbers in parentheses refer to the residue position in the synthetic protein sequences. This map can be divided into two parts: one part represents the rigid environment around the ribityl tail of FMN (upper part) and the other part represents the flexible environment of the FMN-binding site (lower part). In the rigid environment, the indicated H-bond showed to be quite persistent so that the network of interactions observed in the crystal structures were well reproduced in the MD simulations. The flexible environment, in turn, presented a complex H-bonding dynamics. Distances d1-5,d13-16, in red color, were employing for further analysis which resulted in obtaining of different clusters for C426A and iLOV, as discussed in the main text. Copyright: `https://dx.doi.org/10.1021/acs.jpcb.0c10834`.

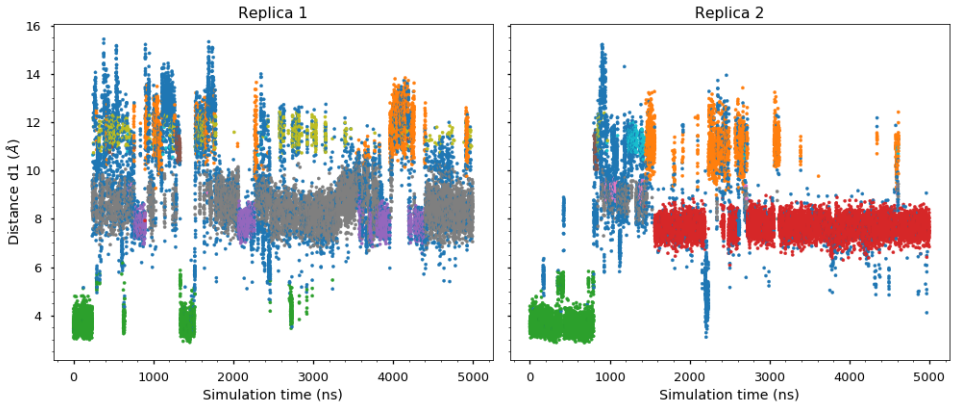

Figure B6: Distance d1 along the two iLOV replicas. The colors refer to the different clusters identified by HDBSCAN. Copyright: `https://dx.doi.org/10.1021/acs.jpcb.0c10834`.

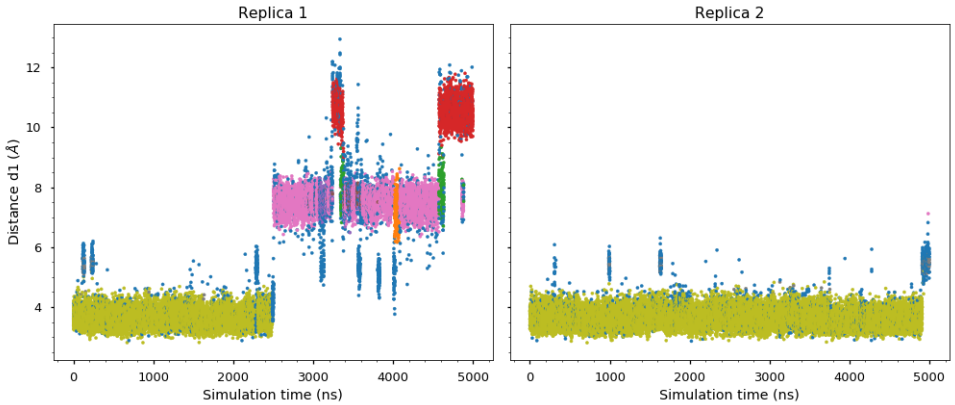

Figure B7: Distance d1 along the two C426A replicas. The colors refer to the different clusters identified by HDBSCAN. Copyright: `https://dx.doi.org/10.1021/acs.jpcb.0c10834`.

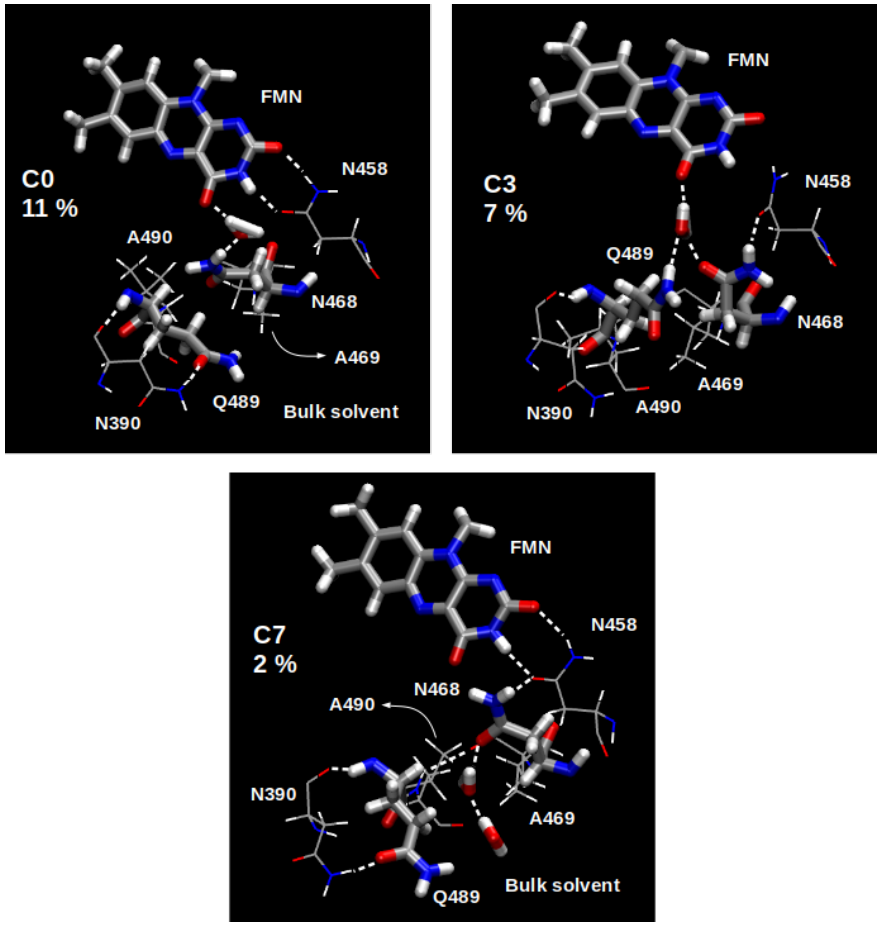

Figure B8: Less populated clusters obtained for iLOV (C0, C3 and C7) also employed for the multiscale calculations. Copyright: `https://dx.doi.org/10.1021/acs.jpcb.0c10834`.

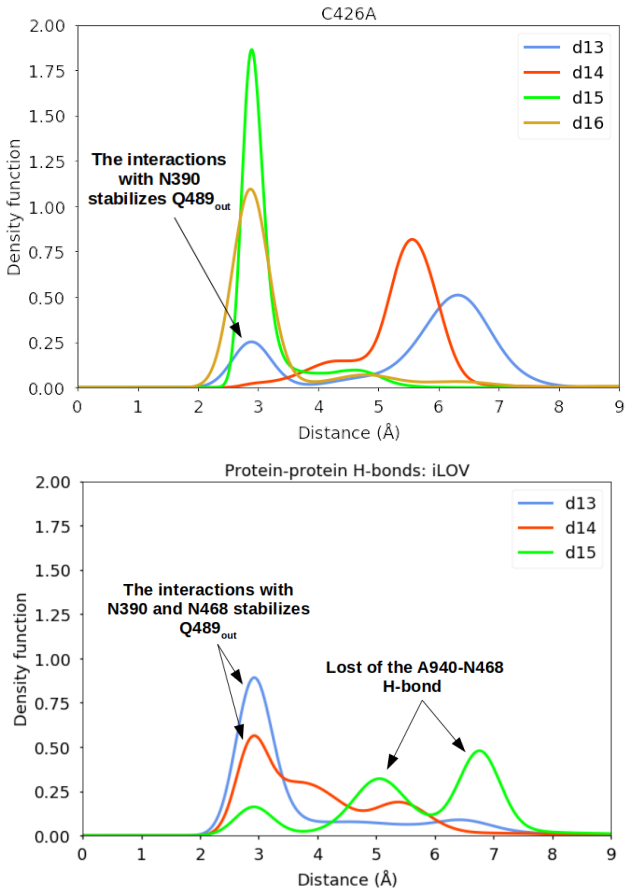

Figure B9: Distributions obtained for d13-15 distances. Copyright: `https://dx.doi.org/10.1021/acs.jpcb.0c10834`.

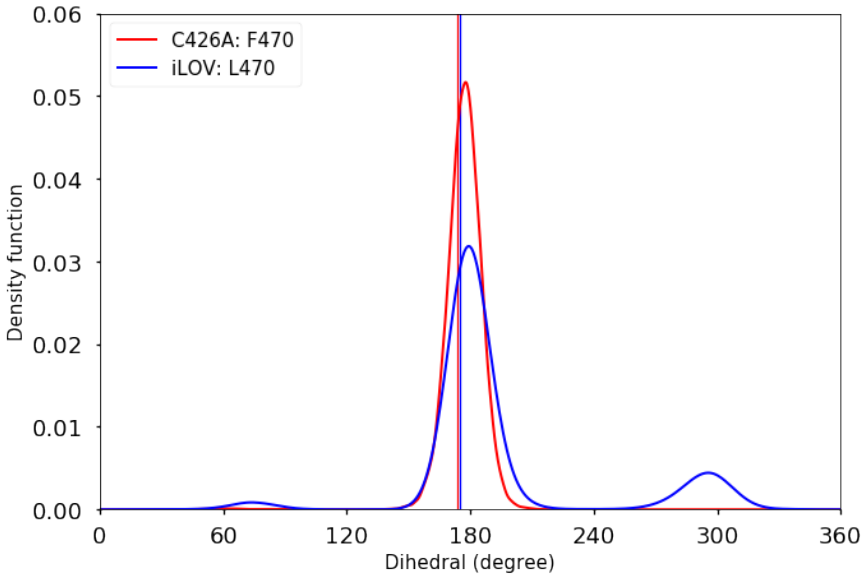

Figure B10: Dihedral distributions for F470 in C426A (red) and L470 in iLOV (blue). Vertical lines indicate the values obtained for crystal structures. Copyright: `https://dx.doi.org/10.1021/acs.jpcb.0c10834`.

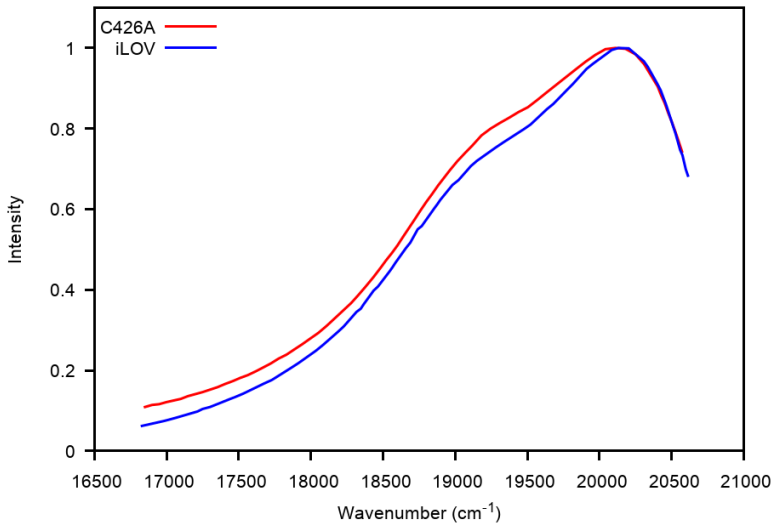

Figure B11: Comparison of the C426A and iLOV fluorescence spectra measured by Chapman et al. [121]. Copyright: `https://dx.doi.org/10.1021/acs.jpcb.0c10834`.

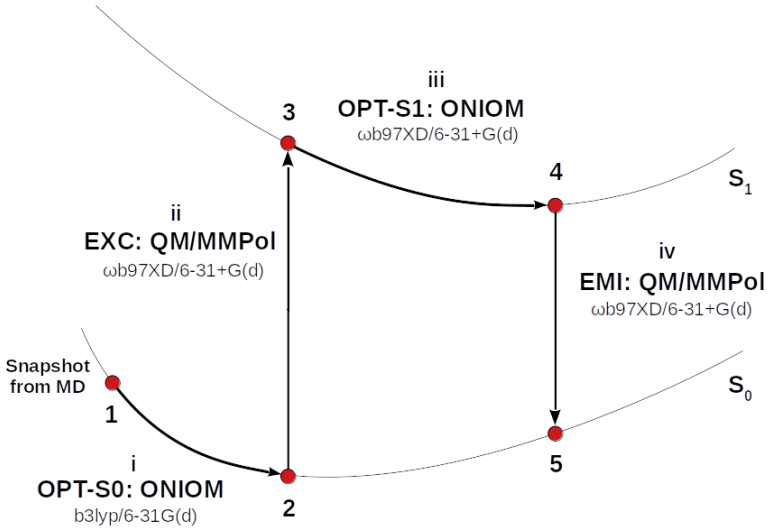

Figure B12: Schematic diagram representing the multiscale steps employed to obtain the excitation and emission vertical energies.

Bibliography

[1] Albert, B., Johnson, A., Lewis, J., Raff, M., Roberts, K. & Walter, P. Molecular Biology of the Cell 5th edition. Garland Science, 2008.

[2] Nelson, D. L. & Cox, M. M. Lehninger principles of biochemistry 7th edition, 2017.

[3] Berman, H. M. The Protein Data Bank: A historical perspective. *Acta Crystallogr., Sect. A: Found. Crystallogr.*, 64(1):88–95, 2008.

[4] PDB Statistics. `https://www.rcsb.org/stats`, 2020.

[5] Orozco, M. A theoretical view of protein dynamics. *Chem. Soc. Rev.*, 43(14):5051–5066, 2014.

[6] Janeway Jr, C. A., Travers, P., Walport, M. & Shlomchik, M. J. The structure of a typical antibody molecule. *Immunobiology: The Immune System in Health and Disease. 5th edition*. Garland Science, 2001.

[7] Sandin, S., Öfverstedt, L.-G., Wikström, A.-C., Wrange, Ö. & Skoglund, U. Structure and flexibility of individual immunoglobulin G molecules in solution. *Structure*, 12(3):409–415, 2004.

[8] Cogdell, R. J., Gall, A. & Köhler, J. The architecture and function of the light-harvesting apparatus of purple bacteria: from single molecules to in vivo membranes. *Q. Rev. Biophys.*, 39(03):227–98, 2006.

[9] Sakai, T., Kagawa, T., Kasahara, M., Swartz, T. E., Christie, J. M., Briggs, W. R., Wada, M. & Okada, K. Arabidopsis nph1 and npl1:

blue light receptors that mediate both phototropism and chloroplast relocation. *PNAS*, 98(12):6969–6974, 2001.

[10] Herrou, J. & Crosson, S. Function, structure and mechanism of bacterial photosensory LOV proteins. *Nat. Rev. Microbiol.*, 9(10):713–723, 2011.

[11] Swartz, T. E., Tseng, T. S., Frederickson, M. A., Paris, G., Comerci, D. J., Rajashekara, G., Kim, J. G., Mudgett, M. B., Splitter, G. A., Ugalde, R. A., Goldbaum, F. A., Briggs, W. R. & Bogomolni, R. A. Blue-light-activated histidine kinases: Two-component sensors in bacteria. *Science*, 317(5841):1090–1093, 2007.

[12] Karplus, M. & Kuriyan, J. Molecular dynamics and protein function. *Proc. Natl. Acad. Sci. U. S. A.*, 102(19):6679–6685, 2005.

[13] Adcock, S. A. & McCammon, J. A. Molecular Dynamics: Survey of Methods for Simulating the Activity of Proteins. *Chem. Rev.*, 106(5):1589–1615, 2006.

[14] Brini, E., Simmerling, C. & Dill, K. Protein storytelling through physics. *Science*, 370(6520):eaaz3041, 2020.

[15] Brunk, E. & Rothlisberger, U. Mixed Quantum Mechanical/Molecular Mechanical Molecular Dynamics Simulations of Biological Systems in Ground and Electronically Excited States. *Chem. Rev.*, 115(12):6217–6263, 2015.

[16] Curutchet, C. & Mennucci, B. Quantum Chemical Studies of Light Harvesting. *Chem. Rev.*, 117:294–343, 2017.

[17] Leach, A. R. *Molecular Modelling: Principles and Applications*. Pearson Education Limited, Harlow, second edi ed., 2001.

[18] Martínez, L., Borin, I. A. & Skaf, M. S. Fundamentos de Simulação por Dinâmica Molecular. N. H. Morgon & K. Coutinho, eds., *Métodos de*

Química Teórica e Modelagem Molecular, 413–452. Livravia da Física, São Paulo, 2007.

[19] Brooks, B. R., Bruccoleri, R. E., Olafson, B. D., States, D. J., Swaminathan, S. a. & Karplus, M. CHARMM: a program for macromolecular energy, minimization, and dynamics calculations. *J. Comput. Chem.*, 4(2):187–217, 1983.

[20] Jorgensen, W. L. & Tirado-Rives, J. The OPLS [optimized potentials for liquid simulations] potential functions for proteins, energy minimizations for crystals of cyclic peptides and crambin. *J. Am. Chem. Soc.*, 110(6):1657–1666, 1988.

[21] Case, D. A., Ben-Shalom, I. Y., Brozell, S. R., Cerutti, D. S., Cheatham, T. E., III, Cruzeiro, V. W. D., Darden, T. A., Duke, R., Ghoreishi, D., Gilson, M. K., Gohlke, H., Goetz, A. W., Greene, D., Harris, R., Homeyer, N., Izadi, S., Kovalenko, A., Kurtzman, T., Lee, T. S., LeGrand, S., Li, P., Lin, C., Liu, J., Luchko, T., Luo, R., Mermelstein, D. J., Merz, K. M., Miao, Y., Monard, G., Nguyen, C., Nguyen, H., Omelyan, I., Onufriev, A., Pan, F., Qi, R., Roe, D. R., Roitberg, A., Sagui, C., Schott-Verdugo, S., Shen, J., Simmerling, C. L., Smith, J., Salomon-Ferrer, R., Swails, J., Walker, R. C., Wang, J., Wei, H., Wolf, R. M., Wu, X., Xiao, L., York, D. M. & Kollman, P. A. AMBER 2018, 2018. University of California, San Francisco.

[22] Scott, W. R., Hünenberger, P. H., Tironi, I. G., Mark, A. E., Billeter, S. R., Fennen, J., Torda, A. E., Huber, T., Krüger, P. & van Gunsteren, W. F. The GROMOS biomolecular simulation program package. *J. Phys. Chem. A*, 103(19):3596–3607, 1999.

[23] Gordon, J. C., Myers, J. B., Folta, T., Shoja, V., Heath, L. S. & Onufriev, A. H++: a server for estimating pKas and adding missing hydrogens to macromolecules. *Nucl. Acids Res.*, 33(Web Server Issue):W368–W371, 2005.

[24] Kaufmann, K. W., Lemmon, G. H., DeLuca, S. L., Sheehan, J. H. & Meiler, J. Practically useful: what the Rosetta protein modeling suite can do for you. *Biochemistry*, 49(14):2987–2998, 2010.

[25] Eswar, N., Webb, B., Marti-Renom, M. A., Madhusudhan, M., Eramian, D., Shen, M.-y., Pieper, U. & Sali, A. Comparative protein structure modeling using Modeller. *Curr. Protoc. Bioinformatics*, 15(1):5–6, 2006.

[26] Braun, E., Gilmer, J., Mayes, H. B., Mobley, D. L., Monroe, J. I., Prasad, S. & Zuckerman, D. M. Best Practices for Foundations in Molecular Simulations [Article v1.0]. *LiveCoMS*, 1(1):1–28, 2019.

[27] Humphrey, W., Dalke, A., Schulten, K. *et al.*. VMD: visual molecular dynamics. *J. Mol. Graph.*, 14(1):33–38, 1996.

[28] Warshel, A. & Levitt, M. Theoretical studies of enzymic reactions: Dielectric, electrostatic and steric stabilization of the carbonium ion in the reaction of lysozyme. *J. Mol. Biol.*, 103(2):227–249, 1976.

[29] Gao, J. Hybrid Quantum and Molecular Mechanical Simulations: An Alternative Avenue to Solvent Effects in Organic Chemistry. *Acc. Chem. Res.*, 29(6):298–305, 1996.

[30] Senn, H. M. & Thiel, W. QM/MM methods for biomolecular systems. *Angewandte Chemie - International Edition*, 48(7):1198–1229, 2009.

[31] Lin, H. & Truhlar, D. G. QM/MM: what have we learned, where are we, and where do we go from here? *Theor. Chem. Acc.*, 117:185, 2006.

[32] Brunk, E. & Rothlisberger, U. Mixed Quantum Mechanical/Molecular Mechanical Molecular Dynamics Simulations of Biological Systems in Ground and Electronically Excited States. *Chem. Rev.*, 115(12):6217–6263, 2015.

[33] Morzan, U. N., de Armiño, D. J. A., Foglia, N. O., Lebrero, M. C. G., Scherlis, D. A. & Estrin, D. A. Spectroscopy in Complex Environments from QM–MM Simulations. *Chem. Rev.*, 118(7):4071–4113, 2018.

[34] Chung, L. W., Hirao, H., Li, X. & Morokuma, K. The ONIOM method: its foundation and applications to metalloenzymes and photobiology. *Wiley Interdiscip. Rev. Comput. Mol. Sci.*, 2(2):327–350, 2011.

[35] Chung, L. W., Sameera, W. M. C., Ramozzi, R., Page, A. J., Hatanaka, M., Petrova, G. P., Harris, T. V., Li, X., Ke, Z., Liu, F., Li, H. B., Ding, L. & Morokuma, K. The ONIOM Method and Its Applications. *Chem. Rev.*, 115(12):5678–5796, 2015.

[36] Bondanza, M., Nottoli, M., Cupellini, L., Lipparini, F. & Mennucci, B. Polarizable embedding QM/MM: the future gold standard for complex (bio)systems? *Phys. Chem. Chem. Phys.*, 22(26):14433–14448, 2020.

[37] Gordon, M. S., Fedorov, D. G., Pruitt, S. R. & Slipchenko, L. V. Fragmentation Methods: A Route to Accurate Calculations on Large Systems. *Chem. Rev.*, 112(1):632–672, 2012.

[38] Gao, J., Truhlar, D. G., Wang, Y., Mazack, M. J. M., Löffler, P., Provorse, M. R. & Rehak, P. Explicit Polarization: A Quantum Mechanical Framework for Developing Next Generation Force Fields. *Acc. Chem. Res.*, 47(9):2837–2845, 2014.

[39] Cieplak, P., Dupradeau, F.-Y., Duan, Y. & Wang, J. Polarization effects in molecular mechanical force fields. *J. Phys. Condens. Matter*, 21(33):333102–22, 2009.

[40] Boulanger, E. & Thiel, W. Solvent Boundary Potentials for Hybrid QM/MM Computations Using Classical Drude Oscillators: A Fully Polarizable Model. *J. Chem. Theory Comput.*, 8(11):4527–4538, 2012.

[41] Lu, Z. & Zhang, Y. Interfacing ab Initio Quantum Mechanical Method with Classical Drude Osillator Polarizable Model for Molecular Dy-

namics Simulation of Chemical Reactions. *J. Chem. Theory Comput.*, 4(8):1237–1248, 2008.

[42] Lemkul, J. A., Huang, J., Roux, B. & MacKerell Jr., A. D. An Empirical Polarizable Force Field Based on the Classical Drude Oscillator Model: Development History and Recent Applications. *Chem. Rev.*, 116(9):4983–5013, 2016.

[43] Rappe, A. & Goddard, W. Charge Equilibration for Molecular Dynamics Simulations. *J. Phys. Chem.*, 95(8):3358–3363, 1991.

[44] Bryce, R. A., Buesnel, R., Hillier, I. H. & Burton, N. A. A solvation model using a hybrid quantum mechanical/molecular mechanical potential with fluctuating solvent charges. *Chem. Phys. Lett.*, 279(5–6):367 – 371, 1997.

[45] Lipparini, F. & Barone, V. Polarizable Force Fields and Polarizable Continuum Model: A Fluctuating Charges/PCM Approach. 1. Theory and Implementation. *J. Chem. Theory Comput.*, 7(11):3711–3724, 2011.

[46] Thompson, M. A. & Schenter, G. K. Excited states of the bacteriochlorophyll b dimer of *Rhodopseudomonas viridis*: a QM/MM study of the photosynthetic reaction center that includes MM polarization. *J. Phys. Chem.*, 99(17):6374–6386, 1995.

[47] Gao, J. Energy components of aqueous solution: Insight from hybrid QM/MM simulations using a polarizable solvent model. *J. Comput. Chem.*, 18(8):1061–1071, 1997.

[48] Curutchet, C., Muñoz-Losa, A., Monti, S., Kongsted, J., Scholes, G. D. & Mennucci, B. Electronic energy transfer in condensed phase studied by a polarizable QM/MM model. *J. Chem. Theory Comput.*, 5(7):1838–1848, 2009.

[49] Olsen, J. M. H. & Kongsted, J. Molecular Properties through Polarizable Embedding. J. R. Sabin & E. Brändas, eds., *Advances in Quantum Chemistry*, 107–143. Elsevier, 2011.

[50] Ponder, J. W., Wu, C., Ren, P., Pande, V. S., Chodera, J. D., Schnieders, M. J., Haque, I., Mobley, D. L., Lambrecht, D. S., DiStasio Jr, R. A., Head-Gordon, M., Clark, G. N. I., Johnson, M. E. & Head-Gordon, T. Current Status of the AMOEBA Polarizable Force Field. *J. Phys. Chem. B*, 114(8):2549–2564, 2010.

[51] Lipparini, F., Lagardère, L., Stamm, B., Cancès, E., Schnieders, M., Ren, P., Maday, Y. & Piquemal, J.-P. Scalable Evaluation of Polarization Energy and Associated Forces in Polarizable Molecular Dynamics: I. Toward Massively Parallel Direct Space Computations. *J. Chem. Theory Comput.*, 10(4):1638–1651, 2014.

[52] Wang, J., Cieplak, P., Li, J., Hou, T., Luo, R. & Duan, Y. Development of polarizable models for molecular mechanical calculations I: parameterization of atomic polarizability. *J. Phys. Chem. B*, 115(12):3091–9, 2011.

[53] Thole, B. T. Molecular polarizabilities calculated with a modified dipole interaction. *Chem. Phys.*, 59(3):341, 1981.

[54] Georg, H. C. & Canuto, S. Métodos Híbridos para Modelagem do Ambiente Molecular. N. H. Morgon & K. Coutinho, eds., *Métodos de Química Teórica e Modelagem Molecular*, 453–488. Livravia da Física, São Paulo, 2007.

[55] Mennucci, B. Modeling absorption and fluorescence solvatochromism with QM/Classical approaches. *Int. J. Quantum Chem.*, 115(18):1202–1208, 2015.

[56] Zuehlsdorff, T. J. & Isborn, C. M. Modeling absorption spectra of molecules in solution. *Int. J. Quantum Chem.*, 119(1):e25719–18, 2018.

[57] Santoro, F. & Jacquemin, D. Going beyond the vertical approximation with time-dependent density functional theory. *Wires Comput. Mol. Sci.*, 6(5):460–486, 2016.

[58] Ramos, F. C., Nottoli, M., Cupellini, L. & Mennucci, B. The molecular mechanisms of light adaption in light- harvesting complexes of purple bacteria revealed by a multiscale modeling. *Chem. Sci.*, 10(42):9615–9894, 2019.

[59] Robert, B., Cogdell, R. J. & van Grondelle, R. *The Light-Harvesting System of Purple Bacteria*, 169–194. Springer Netherlands, Dordrecht, 2003.

[60] Law, C. J., Roszak, A. W., Southall, J., Gardiner, A. T., Isaacs, N. W. & Cogdell, R. J. The structure and function of bacterial light-harvesting complexes. *Mol. Membr. Biol.*, 21(3):183–191, 2009.

[61] Mirkovic, T., Ostroumov, E. E., Anna, J. M., van Grondelle, R., Govindjee & Scholes, G. D. Light Absorption and Energy Transfer in the Antenna Complexes of Photosynthetic Organisms. *Chem. Rev.*, 117(2):249–293, 2017.

[62] Gardiner, A. T., Cogdell, R. J. & Takaichi, S. The effect of growth conditions on the light-harvesting apparatus in *Rhodopseudomonas acidophila*. *Photosynth. Res.*, 38(2):159–167, 1993.

[63] Scheuring, S. & Sturgis, J. N. Chromatic adaptation of photosynthetic membranes. *Science*, 309(5733):484–487, 2005.

[64] Moulisová, V., Lüer, L., Hoseinkhani, S., Brotosudarmo, T. H. P., Collins, A. M., Lanzani, G., Blankenship, R. E. & Cogdell, R. J. Low Light Adaptation: Energy Transfer Processes in Different Types of Light Harvesting Complexes from *Rhodopseudomonas palustris*. *Biophys. J.*, 97(11):3019–3028, 2009.

[65] Gardiner, A. T., Niedzwiedzki, D. M. & Cogdell, R. J. Adaptation of *Rhodopseudomonas acidophila* strain 7050 to growth at different light intensities: what are the benefits to changing the type of LH2? *Faraday Discuss.*, 207:471–489, 2018.

[66] Henry, S. L. & Cogdell, R. J. The Evolution of the Purple Photosynthetic Bacterial Light-Harvesting System. *Adv. Bot. Res.*, 66:205 – 226, 2013.

[67] Prince, S. M., Papiz, M. Z., Freer, A. A., McDermott, G., Hawthornthwaite-Lawless, A. M., Cogdell, R. J. & Isaacs, N. W. Apoprotein structure in the LH2 complex from *Rhodopseudomonas acidophila* strain 10050: modular assembly and protein pigment interactions. *J. Mol. Biol*, 268(2):412–423, 1997.

[68] Papiz, M. Z., Prince, S. M., Howard, T., Cogdell, R. J. & Isaacs, N. W. The Structure and Thermal Motion of the B800–850 LH2 Complex from Rps.acidophila at 2.0Å Resolution and 100K: New Structural Features and Functionally Relevant Motions. *J. Mol. Biol*, 326(5):1523–1538, 2003.

[69] McLuskey, K., Prince, S. M., Cogdell, R. J. & Isaacs, N. W. The Crystallographic Structure of the B800-820 LH3 Light-Harvesting Complex from the Purple Bacteria *Rhodopseudomonas acidophila* Strain 7050. *Biochemistry*, 40(30):8783–8789, 2001.

[70] Hartigan, N., Tharia, H. A., Sweeney, F., Lawless, A. M. & Papiz, M. Z. The 7.5-Å electron density and spectroscopic properties of a novel low-light B800 LH2 from *Rhodopseudomonas palustris. Biophys. J.*, 82(2):963–977, 2002.

[71] Brotosudarmo, T. H. P., Kunz, R., Böhm, P., Gardiner, A. T., Moulisová, V., Cogdell, R. J. & Köhler, J. Single-Molecule Spectroscopy Reveals that Individual Low-Light LH2 Complexes from

Rhodopseudomonas palustris 2.1.6. Have a Heterogeneous Polypeptide Composition. *Biophys. J.*, 97(5):1491–1500, 2009.

[72] Southall, J., Henry, S. L., Gardiner, A. T., Roszak, A. W., Mullen, W., Carey, A.-M., Kelly, S. M., de Percin Northumberland, C. O. & Cogdell, R. J. Characterisation of a pucBA deletion mutant from *Rhodopseudomonas palustris* lacking all but the pucBAd genes. *Photosynth. Res.*, 135(1-3):9–21, 2018.

[73] Scholes, G. D., Gould, I. R., Cogdell, R. J. & Fleming, G. R. Ab Initio Molecular Orbital Calculations of Electronic Couplings in the LH2 Bacterial Light-Harvesting Complex of Rps. Acidophila. *J. Phys. Chem. B*, 103(13):2543–2553, 1999.

[74] Montemayor, D., Rivera, E. & Jang, S. J. Computational Modeling of Exciton-Bath Hamiltonians for Light Harvesting 2 and Light Harvesting 3 Complexes of Purple Photosynthetic Bacteria at Room Temperature. *J. Phys. Chem. B*, 122(14):3815–3825, 2018.

[75] Gudowska-Nowak, E., Newton, M. & Fajer, J. Conformational and environmental effects on bacteriochlorophyll optical spectra: correlations of calculated spectra with structural results. *J. Chem. Phys.*, 94:5795–5801, 1990.

[76] Ketelaars, M., Segura, J.-M., Oellerich, S., de Ruijter, W. P. F., Magis, G., Aartsma, T. J., Matsushita, M., Schmidt, J., Cogdell, R. J. & Köhler, J. Probing the Electronic Structure and Conformational Flexibility of Individual Light-Harvesting 3 Complexes by Optical Single-Molecule Spectroscopy. *J. Phys. Chem. B*, 110(37):18710–18717, 2006.

[77] De Vico, L., Anda, A., Osipov, V. A., Madsen, A. Ø. & Hansen, T. Macrocycle ring deformation as the secondary design principle for light-harvesting complexes. *Proc. Natl. Acad. Sci. U.S.A.*, 6(39):E9051–E9057, 2018.

[78] Beekman, L. M. P., Frese, R. N., Fowler, G. J. S., Picorel, R., Cogdell, R. J., van Stokkum, I. H. M., Hunter, C. N. & van Grondelle, R. Characterization of the Light-Harvesting Antennas of Photosynthetic Purple Bacteria by Stark Spectroscopy. 2. LH2 Complexes: Influence of the Protein Environment. *J. Phys. Chem. B*, 101(37):7293–7301, 1997.

[79] Chmeliov, J., Songaila, E., Rancova, O., Gall, A., Robert, B., Abramavicius, D. & Valkunas, L. Excitons in the LH3 Complexes from Purple Bacteria. *J. Phys. Chem. B*, 117(38):11058–11068, 2013.

[80] Li, X., Parrish, R. M., Liu, F., Kokkila Schumacher, S. I. L. & Martínez, T. J. An Ab Initio Exciton Model Including Charge-Transfer Excited States. *J. Chem. Theory Comput.*, 13(8):3493–3504, 2017.

[81] Alden, R. G., Johnson, E., Nagarajan, V., Parson, W. W., Law, C. J. & Cogdell, R. J. Calculations of Spectroscopic Properties of the LH2 Bacteriochlorophyll | Protein Antenna Complex from *Rhodopseudomonas acidophila*. *J. Phys. Chem. B*, 101(23):4667–4680, 1997.

[82] Nottoli, M., Jurinovich, S., Cupellini, L., Gardiner, A. T., Cogdell, R. & Mennucci, B. The role of charge-transfer states in the spectral tuning of antenna complexes of purple bacteria. *Photosynth. Res.*, 137(2):215–226, 2018.

[83] Eswar, N., Webb, B., Marti-Renom, M. a., Madhusudhan, M. S., Eramian, D., Shen, M.-Y., Pieper, U. & Sali, A. Comparative protein structure modeling using Modeller. *Curr. Protoc. Bioinformatics*, 2006.

[84] Wu, E. L., Cheng, X., Jo, S., Rui, H., Song, K. C., Dávila-Contreras, E. M., Qi, Y., Lee, J., Monje-Galvan, V., Venable, R. M., Klauda, J. B. & Im, W. CHARMM-GUI membrane builder toward realistic biological membrane simulations. *J. Comput. Chem.*, 35(27):1997–2004, 2014.

[85] Dickson, C. J., Madej, B. D., Skjevik, Å. A., Betz, R. M., Teigen, K., Gould, I. R. & Walker, R. C. Lipid14: The amber lipid force field. *J. Chem. Theory Comput.*, 10(2):865–879, 2014.

[86] Cupellini, L., Jurinovich, S., Campetella, M., Caprasecca, S., Guido, C. A., Kelly, S. M., Gardiner, A. T., Cogdell, R. & Mennucci, B. An ab initio description of the excitonic properties of lh2 and their temperature dependence. *J. Phys. Chem. B*, 120(44):11348–11359, 2016.

[87] Darden, T., Perera, L., Li, L. & Lee, P. New tricks for modelers from the crystallography toolkit: The particle mesh Ewald algorithm and its use in nucleic acid simulations. *Structure*, 7(3):55–60, 1999.

[88] Roe, D. R. & Cheatham, T. E. PTRAJ and CPPTRAJ: Software for Processing and Analysis of Molecular Dynamics Trajectory Data. *J. Chem. Theory Comput.*, 9(7):3084–95, 2013.

[89] Humphrey, W., Dalke, A. & Schulten, K. VMD – Visual Molecular Dynamics. *J. Mol. Graphic.*, 14:33–38, 1996.

[90] Maier, J. A., Martinez, C., Kasavajhala, K., Wickstrom, L., Hauser, K. E. & Simmerling, C. ff14SB: Improving the Accuracy of Protein Side Chain and Backbone Parameters from ff99SB. *J. Chem. Theory Comput.*, 11(8):3696–3713, 2015.

[91] Ceccarelli, M., Procacci, P., Marchi, M. & Cedex, G.-s.-y. An Ab initio Force Field for the Cofactors. *J. Comput. Chem.*, 983308:129–142, 2002.

[92] Prandi, I. G. *Development of classical models for the description of QM properties in photoinduced processes*. Ph.D. thesis, University of Pisa, 2016.

[93] Jurinovich, S., Cupellini, L., Guido, C. A. & Mennucci, B. EXAT: EXcitonic analysis tool. *J. Comput. Chem.*, 39(5):279–286, 2018.

[94] Becke, A. D. Density-functional thermochemistry. III. The role of exact exchange. *J. Chem. Phys.*, 98(7):5648, 1993.

[95] Lee, C., Yang, W. & Parr, R. G. Development of the Colle-Salvetti correlation-energy formula into a functional of the electron density. *Phys. Rev. B*, 37(2):785–789, 1988.

[96] Iozzi, M., Mennucci, B., Tomasi, J. & Cammi, R. Excitation energy transfer (EET) between molecules in condensed matter: A novel application of the polarizable continuum model (PCM). *J. Chem. Phys.*, 120:7029–40, 2004.

[97] Scholes, G. D., Curutchet, C., Mennucci, B., Cammi, R. & Tomasi, J. How Solvent Controls Electronic Energy Transfer and Light Harvesting. *J. Phys. Chem. B*, 111(25):6978–6982, 2007.

[98] Cupellini, L., Caprasecca, S., Guido, C. A., Müh, F., Renger, T. & Mennucci, B. Coupling to Charge Transfer States is the Key to Modulate the Optical Bands for Efficient Light Harvesting in Purple Bacteria. *J. Phys. Chem. Lett.*, 9(23):6892–6899, 2018.

[99] Hsu, C. P., You, Z. Q. & Chen, H. C. Characterization of the Short-Range Couplings in Excitation Energy Transfer. *J. Phys. Chem. C*, 112(4):1204–1212, 2008.

[100] Voityuk, A. A. & Rösch, N. Fragment charge difference method for estimating donor–acceptor electronic coupling: Application to DNA π-stacks. *J. Chem. Phys.*, 117(12):5607–11, 2002.

[101] Yang, C.-H. & Hsu, C.-P. A multi-state fragment charge difference approach for diabatic states in electron transfer: Extension and automation. *J. Chem. Phys.*, 139(15):154104–16, 2013.

[102] Higashi, M., Kosugi, T., Hayashi, S. & Saito, S. Theoretical Study on Excited States of Bacteriochlorophyll a in Solutions with Density Functional Assessment. *J. Phys. Chem. B*, 118(37):10906–10918, 2014.

[103] Freiberg, A., Timpmann, K. & Trinkunas, G. Spectral fine-tuning in excitonically coupled cyclic photosynthetic antennas. *Chem. Phys. Lett.*, 500(1-3):111–115, 2010.

[104] Pajusalu, M., Rätsep, M., Trinkunas, G. & Freiberg, A. Davydov Splitting of Excitons in Cyclic Bacteriochlorophyll a Nanoaggregates of Bacterial Light-Harvesting Complexes between 4.5 and 263 K. *ChemPhysChem*, 12(3):634–644, 2011.

[105] Kunz, R., Timpmann, K., Southall, J., Cogdell, R. J., Köhler, J. & Freiberg, A. Fluorescence-Excitation and Emission Spectra from LH2 Antenna Complexes of *Rhodopseudomonas acidophila* as a Function of the Sample Preparation Conditions. *J. Phys. Chem. B*, 117(40):12020–12029, 2013.

[106] Magdaong, N. M., LaFountain, A. M., Greco, J. A., Gardiner, A. T., Carey, A.-M., Cogdell, R. J., Gibson, G. N., Birge, R. R. & Frank, H. A. High Efficiency Light Harvesting by Carotenoids in the LH2 Complex from Photosynthetic Bacteria: Unique Adaptation to Growth under Low-Light Conditions. *J. Phys. Chem. B*, 118(38):11172–11189, 2014. PMID: 25171303.

[107] Fowler, G. J., Visschers, R. W., Grief, G. G., van Grondelle, R. & Hunter, C. N. Genetically modified photosynthetic antenna complexes with blueshifted absorbance bands. *Nature*, 355(6363):848–850, 1992.

[108] Olsen, J. D., Sockalingum, G. D., Robert, B. & Hunter, C. N. Modification of a hydrogen bond to a bacteriochlorophyll a molecule in the light-harvesting 1 antenna of *Rhodobacter sphaeroides*. *Proc. Natl. Acad. Sci. U.S.A*, 91(15):7124–7128, 1994.

[109] Jang, S., Rivera, E. & Montemayor, D. Molecular Level Design Principle behind Optimal Sizes of Photosynthetic LH2 Complex: Taming Disorder through Cooperation of Hydrogen Bonding and Quantum Delocalization. *J. Phys. Chem. Lett.*, 6(6):928–934, 2015.

[110] Anda, A., De Vico, L. & Hansen, T. Intermolecular Modes between LH2 Bacteriochlorophylls and Protein Residues: The Effect on the Excitation Energies. *J. Phys. Chem. B*, 121(22):5499–5508, 2017.

[111] Ferretti, M., Hendrikx, R., Romero, E., Southall, J., Cogdell, R. J., Novoderezhkin, V. I., Scholes, G. D. & van Grondelle, R. Dark States in the Light-Harvesting complex 2 Revealed by Two-dimensional Electronic Spectroscopy. *Sci. Rep.*, 6(1):20834, 2016.

[112] Ramos, F. C., Cupellini, L. & Mennucci, B. Computational Investigation of Structural and Spectroscopic Properties of LOV-Based Proteins with Improved Fluorescence. *J. Phys. Chem. B*, 125(7):1768, 2021.

[113] Chalfie, M., Tu, Y., Euskirchen, G., Ward, W. W. & Prasher, D. C. Green fluorescent protein as a marker for gene expression. *Science*, 263(5148):802–805, 1994.

[114] Chia, H. E., Marsh, E. N. G. & Biteen, J. S. Extending fluorescence microscopy into anaerobic environments. *Curr. Opin. Chem. Biol.*, 51:98–104, 2019.

[115] Ozbakir, H. F., Anderson, N. T., Fan, K.-C. & Mukherjee, A. Beyond the Green Fluorescent Protein: Biomolecular Reporters for Anaerobic and Deep-Tissue Imaging. *Bioconjugate Chem.*, 31(2):293–302, 2020.

[116] Drepper, T., Eggert, T., Circolone, F., Heck, A., Krauß, U., Guterl, J.-K., Wendorff, M., Losi, A., Gärtner, W. & Jaeger, K.-E. Reporter proteins for in vivo fluorescence without oxygen. *Nat. Biotechnol*, 25(4):443–445, 2007.

[117] Mukherjee, A. & Schroeder, C. M. Flavin-based fluorescent proteins: Emerging paradigms in biological imaging. *Curr. Opin. Biotechnol.*, 31:16–23, 2015.

[118] Losi, A. & Viappiani, C. *Flavin Mononucleotide-Binding Fluorescent Proteins*, 1–9. Springer Berlin Heidelberg, Berlin, Heidelberg, 2020.

[119] Anderson, N. T., Weyant, K. B. & Mukherjee, A. Characterization of flavin binding in oxygen-independent fluorescent reporters. *AIChE J.*, 66(12):4784–9, 2020.

[120] Kopka, B., Magerl, K., Savitsky, A., Davari, M. D., Röllen, K., Bocola, M., Dick, B., Schwaneberg, U., Jaeger, K. E. & Krauss, U. Electron transfer pathways in a light, oxygen, voltage (LOV) protein devoid of the photoactive cysteine. *Sci. Rep.*, 7(1):1–17, 2017.

[121] Chapman, S., Faulkner, C., Kaiserli, E., Garcia-Mata, C., Savenkov, E. I., Roberts, A. G., Oparka, K. J. & Christie, J. M. The photoreversible fluorescent protein iLOV outperforms GFP as a reporter of plant virus infection. *Proc. Natl. Acad. Sci. U.S.A.*, 105(50):20038–20043, 2008.

[122] Christie, J. M., Hitomi, K., Arvai, A. S., Hartfield, K. A., Mettlen, M., Pratt, A. J., Tainer, J. A. & Getzoff, E. D. Structural tuning of the fluorescent protein iLOV for improved photostability. *J. Biol. Chem.*, 287(26):22295–22304, 2012.

[123] Shu, X., Lev-Ram, V., Deerinck, T. J., Qi, Y., Ramko, E. B., Davidson, M. W., Jin, Y., Ellisman, M. H. & Tsien, R. Y. A genetically encoded tag for correlated light and electron microscopy of intact cells, tissues, and organisms. *PLoS Biol.*, 2011.

[124] Mukherjee, A., Weyant, K. B., Walker, J. & Schroeder, C. M. Directed evolution of bright mutants of an oxygen-independent flavin-binding fluorescent protein from Pseudomonas putida. *J. Biol. Eng*, 6(1):20, 2012.

[125] Wingen, M., Potzkei, J., Endres, S., Casini, G., Rupprecht, C., Fahlke, C., Krauss, U., Jaeger, K. E., Drepper, T. & Gensch, T. The photophysics of LOV-based fluorescent proteins-new tools for cell biology. *Photochem. Photobiol. Sci*, 13(6):875–883, 2014.

[126] Mukherjee, A., Walker, J., Weyant, K. B. & Schroeder, C. M. Characterization of flavin-based fluorescent proteins: an emerging class of fluorescent reporters. *PloS one*, 8(5), 2013.

[127] Davari, M. D., Kopka, B., Wingen, M., Bocola, M., Drepper, T., Jaeger, K. E., Schwaneberg, U. & Krauss, U. Photophysics of the LOV-Based Fluorescent Protein Variant iLOV-Q489K Determined by Simulation and Experiment. *J. Phys. Chem. B*, 120(13):3344–3352, 2016.

[128] Losi, A., Gardner, K. H. & Möglich, A. Blue-light receptors for optogenetics. *Chem. Rev.*, 118(21):10659–10709, 2018.

[129] Shcherbakova, D. M., Shemetov, A. A., Kaberniuk, A. A. & Verkhusha, V. V. Natural photoreceptors as a source of fluorescent proteins, biosensors, and optogenetic tools. *Annu. Rev. Biochem*, 84:519–550, 2015.

[130] Kobayashi, H., Ogawa, M., Alford, R., Choyke, P. L. & Urano, Y. New strategies for fluorescent probe design in medical diagnostic imaging. *Chem. Rev.*, 110(5):2620–2640, 2010.

[131] Khrenova, M. G., Nemukhin, A. V. & Domratcheva, T. Theoretical characterization of the flavin-based fluorescent protein iLOV and its Q489K mutant. *J. Phys. Chem. B*, 119(16):5176–5183, 2015.

[132] Khrenova, M. G., Meteleshko, Y. I. & Nemukhin, A. V. Mutants of the Flavoprotein iLOV as Prospective Red-Shifted Fluorescent Markers. *J. Phys. Chem. B*, 121(43):10018–10025, 2017.

[133] Meteleshko, Y. I., Nemukhin, A. V. & Khrenova, M. G. Novel flavin-based fluorescent proteins with red-shifted emission bands: a computational study. *Photochem. Photobiol. Sci*, 18(1):177–189, 2019.

[134] Nemukhin, A. V., Grigorenko, B. L., Khrenova, M. G. & Krylov, A. I. Computational Challenges in Modeling of Representative Bioimaging

Proteins: GFP-Like Proteins, Flavoproteins, and Phytochromes. *J. Phys. Chem. B*, 123(29):6133–6149, 2019.

[135] Schneider, C. & Sühnel, J. A molecular dynamics simulation of the flavin mononucleotide–RNA aptamer complex. *Biopolymers*, 50(3):287–302, 1999.

[136] Campello, R. J., Moulavi, D. & Sander, J. Density-based clustering based on hierarchical density estimates. *Lecture Notes in Comput. Sci.*, 7819 LNAI(PART 2):160–172, 2013.

[137] McInnes, L. & Healy, J. Accelerated Hierarchical Density Based Clustering. *IEEE International Conference on Data Mining Workshops, ICDMW*, 2017-Novem:33–42, 2017.

[138] Frisch, M., Trucks, G., Schlegel, H., Scuseria, G., Robb, M., Cheeseman, J., Scalmani, G., Barone, V., Mennucci, B., Petersson, G. *et al.*. Gaussian 09; Gaussian, Inc. *Wallingford, CT*, 32:5648–5652, 2009.

[139] Wang, J., Cieplak, P., Li, J., Wang, J., Cai, Q., Hsieh, M., Lei, H., Luo, R. & Duan, Y. Development of Polarizable Models for Molecular Mechanical Calculations II: Induced Dipole Models Significantly Improve Accuracy of Intermolecular Interaction Energies. *J. Phys. Chem. B*, 115(12):3100–3111, 2011.

[140] Mukamel, S. *Principles of Nonlinear Optical Spectroscopy*. Oxford University Press, New York, 1995.

[141] Freddolino, P. L., Dittrich, M. & Schulten, K. Dynamic switching mechanisms in LOV1 and LOV2 domains of plant phototropins. *Biophys. J.*, 91(10):3630–3639, 2006.

[142] Lokhandwala, J., Hopkins, H. C., Britton, C. W., Rodriguez-Iglesias, A., Bogomolni, R., Schmoll, M., Zoltowski, B. D. *et al.*. A native threonine coordinates ordered water to tune Light-Oxygen-Voltage (LOV)

domain photocycle kinetics and osmotic stress signaling in *Trichoderma reesei* ENVOY. *J. Biol. Chem.*, 291(28):14839–14850, 2016.

[143] Yee, E. F., Diensthuber, R. P., Vaidya, A. T., Borbat, P. P., Engelhard, C., Freed, J. H., Bittl, R., Möglich, A. & Crane, B. R. Signal transduction in light–oxygen–voltage receptors lacking the adduct-forming cysteine residue. *Nat. Commun.*, 6(1):1–10, 2015.