UNIVERSITÀ DEGLI STUDI DI TRIESTE

UNIVERSITÀ DEGLI STUDI DI TRIESTE

XXXVIII CICLO DEL DOTTORATO DI RICERCA IN FISICA

A Computational Study of Optical Properties of Nonaromatic Proteins: The Case of α_3C

Settore scientifico-disciplinare: FIS/03 FISICA DELLA MATERIA

DOTTORANDO / A
Germaine Neza Hozana

COORDINATORE
Prof. Francesco Longo

SUPERVISORE DI TESI
Prof. Ali Hassanali

CO-SUPERVISORE DI TESI
Prof. Nicola Seriani

ANNO ACCADEMICO 2024/2025

Abstract

The mechanisms by which ultraviolet (UV) and visible light interact with nonconjugated molecules remain poorly understood. Traditionally, proteins lacking aromatic residues and prosthetic groups were believed to be optically silent at wavelengths longer than 250 nm. However, growing experimental evidence over the past decade has challenged this assumption, showing that proteins devoid of aromatic and conjugated groups can absorb light in the near-UV beyond 300 nm and emit in the visible region. Understanding the origins of this phenomenon offers exciting opportunities for designing noninvasive spectroscopic probes to study local interactions in biological systems.

Among the emerging cases of nonaromatic and nonconjugated light absorption and emission, the synthetic protein α_3C stands out as an interesting example. Recent experiments demonstrated that α_3C exhibits a broad UV-visible absorption band between 250-800 nm [1], attributed to charge-transfer excitations between charged amino acids, and emission in the 310-550 nm range upon excitation at 295 nm [2].

In this work, we investigate the origins of this unconventional absorption and emission in α_3C. An unsupervised machine learning approach is used to automatically detect statistically significant structural motifs, that are then subjected to QM/MM simulations of the full protein in explicit solvent using the time dependent density-functional tight-binding (TD-DFTB) method. This integrated approach streamlines the identification of statistically significant structural motifs and their direct connection to the observed absorption and emission features.

Our simulated absorption spectra calculations reveal unconventional absorption features spanning 250-350 nm, with the arginine-glutamic acid interactions contributing to all transitions beyond 300 nm. However, the simulated absorption tail remains notably shorter than what is observed experimentally. In particular, our calculations do not predict any absorption between 400 and 800 nm. Transitions in this range appear only when environmental interactions are neglected - a simplification we consider physically inaccurate. To investigate the source of this discrepancy between simulated and experimental results, we examined the influence of nuclear quantum effects (NQEs) on the absorption spectra. Incorporating these effects significantly broaden the spectrum and redshift it by approximately 100 nm, extending the calculated absorption to about 450 nm.

Excited-state dynamics provide additional insight into the emission behavior. Although the arginine-glutamic acid interactions are more prone to low-energy electronic transitions, they contribute minimally to emission. Instead the backbone interactions appear to be the most important

for the emission of the protein, when its α-helical structure remains intact. The resulting emission profile considering contributions from all the three types of interactions (arginine-glutamic acid, other side-chain interactions, and backbone H-bond contacts) combined, are in better agreement with the experimentally reported emission spectrum. Nonradiative relaxation to the ground state primarily proceeds through secondary-structure distortions, proton transfer, and arginine deplanarization.

In summary, this study elucidates the molecular basis of unconventional nonaromatic fluorescence using α_3C as a model system. Hydrogen bonding and charge-transfer interactions, particularly between arginine and glutamic acid residues, drive near-UV absorption, while backbone interactions dominate emission. The results highlight challenges and opportunities from both computational and experimental perspectives. Discrepancies at longer wavelengths highlight the need to incorporate nuclear quantum effects for quantitative accuracy. Experimentally, α_3C variants with targeted arginine mutations and minimal peptide analogs could clarify the roles of hydrogen bonding and structural rigidity, guiding the design of new noninvasive fluorescent probes.

To my Mom and Dad.

Acknowledgements

I would like to express my sincere gratitude to my supervisor, Ali Hassanali, for his guidance, encouragement and patience throughout this journey. I am also very thankful to Gonzalo Diaz Miron, with whom we collaborated during these three years.

I am grateful to the ICTP for providing resources, and a good research environment during my PhD program and the Diploma program which preceded it.

My heartfelt thanks to all my colleagues in Ali's group for their help; to Marco and Kyrell who was always willing to assist with matters concerning PhD students at the University we attended; and to my office-mates for the time we spent together.

I would also like to take this opportunity to extend my deepest gratitude to my parents, N. M. Chantal and N. Germain. Thank you for your endless sacrifices, and for encouraging me to pursue sciences. To my brothers, Kami and Kaze, you are the best! And to my friends who are miles away, your care and encouragement made this journey much easier.

Finally, I would like to acknowlege the support and the wonderful company of the friends I met along this journey, especially, Anna, Diganta and Alessandro, you made this experience lighter and much more enjoyable.

Contents

List of Figures

Chapter 1

Introduction

Light-matter interactions are fundamental to many biological processes, as well as technological innovations [4, 5, 6, 7]. The way molecules absorb, emit, and transform energy upon exposure to light underpins a wide range of natural phenomena and engineered systems. By investigating the photophysical and photochemical properties of molecular systems, researchers can gain insights into how light drives essential biological functions such as vision, photosynthesis, and circadian regulation [5, 8]. These studies also shed light on energy transfer mechanisms, electron dynamics, and conformational changes that govern molecular behavior under illumination [4, 5].

In animals, vision is a crucial sense for survival, and although it involves highly complex biological mechanisms, it begins with light-driven events. Photons interact with specialized light-sensitive proteins known as Rhodopsins in the retina, triggering conformational changes in their chromophores [9, 10], initiating a series of biochemical and electrical processes that ultimately generate neural signals interpreted by the brain as vision.

The profound impact of these light-induced processes extends far beyond individual sensory perception, influencing behavior, survival strategies, and ecological dynamics. In birds, for example, vision plays a central role in flight [11], in detecting predators or locating prey [12], and in identifying flowers for nectar feeding which, in turn, facilitate pollination[13]. Another central example is photosynthesis, through which plants, algae, and certain bacteria convert light energy into chemical energy [14]. This process starts the global food chain, and maintain the planet's oxygen supply. Bioluminescence similarly illustrates the ecological significance of light, allowing organisms to produce their own light through luciferin–luciferase reactions. This self-emitted glow serves functions such as communication, camouflage, and attracting mates or prey, exemplified by fireflies on land and countless deep-sea species that rely on light in the absence of sunlight [15].

Beyond biological contexts, understanding light–matter interactions is crucial for developing advanced materials and technologies. The sun, the most abundant source of renewable energy, drives efforts to harness light efficiently through technologies such as dye-sensitized solar cells (DSSCs), inspired by photosynthesis [16]. DSSCs, employ light-harvesting dyes that can be from natural plants, offering a low-cost, flexible and environmental friendly alternative to conventional silicon-based photovoltaics [17, 18]. Similarly, in biomedicine, photodynamic therapy (PDT) ap-

plies the same principles by using a photosensitizer, light at specific photon energies and oxygen, to generate reactive oxygen species that selectively destroy diseased cells, offering a targeted and minimally invasive approach to cancer treatment [19, 20, 21].

Light–matter interactions also form the foundation of display technologies, where specialized molecules and materials control the generation, modulation, and emission of photons to create visual interfaces [22]. These capabilities underpin a wide range of applications, from high-definition televisions, smartphones, and flexible displays to virtual and augmented reality devices. Beyond commercial electronics, advances in light-matter engineering are also applied in medical imaging, optical sensors, and photonic circuits [23, 24].
Collectively, these examples illustrate how understanding and harnessing light–matter interactions has potential to drive innovation across disciplines, shaping sustainable technologies.

The growing demand for improved and novel light-based technologies underscores the importance of tailoring these interactions at the molecular and atomic scale [5]. In medicine, for example, the need for noninvasive cancer treatments highlights the importance of designing photosensitizing drugs suitable for various tissues and organs, with required penetration depth and biocompatibility [25, 26, 27]. Similarly in the context of DSSCs, which outperform other photovoltaic technologies indoor, and are well-suited for internet of things applications, there is need of designing dyes with improved efficiency and stability for organic DSSCs [28, 29, 30, 31]. This requires careful tailoring of molecular properties to match the optical characteristics of different biological/chemical environments. These examples illustrate a broader principle: understanding the interaction of light with matter at the molecular and atomic scale is essential for advancing sustainable technologies.

1.1 Photophysical Properties of Molecules

When light (treated as a stream of photons) interacts with a molecule, absorption occurs if a photon's energy matches the molecule's quantized energy gaps. Light of different frequencies can induce transitions between various types of quantum states. Ultraviolet (UV) and visible light primarily drive transitions between electronic energy levels, which are further split into vibrational and rotational sub-levels. These electronic states are described by potential energy surfaces (PESs), which represent how the energy of a molecule depends on the positions of its nuclei [32]. Each electronic state - such as the ground state (S_0) or an excited state - corresponds to a distinct PES with a characteristic minimum corresponding to a stable molecular geometry.

Absorption in the UV/visible range promotes the molecule from its stable electronic ground state to an excited electronic state, which is inherently unstable and short-lived. Excited-state molecules can undergo two general types of processes: photochemical and photophysical. In photochemical processes, the absorbed energy induces chemical transformations such as bond breaking, formation, or rearrangement, thereby altering the molecule's chemical identity. In contrast, during photophysical processes, the molecule relaxes without undergoing chemical change. These processes are described in Figure 1.1 below, know as Jablonski diagram [33, 34].

Following excitation, the molecule rapidly dissipates excess vibrational energy through vibrational relaxation, towards the lowest vibrational level within that excited electronic state. During this process, it may undergo internal conversion (IC), a nonradiative transition between electronic states of the same spin multiplicity, or intersystem crossing (ISC), a transition between states of different spin multiplicities, such as from a singlet (S_1) to a triplet (T_1) state. Once in the lowest excited state, the system can return to the ground state either nonradiatively or by emitting light. Nonradiative transitions are often facilitated by intersections of potential energy surfaces known as conical intersections. The latter process, known as radiative decay, manifests as fluorescence when emission occurs from a singlet state, or phosphorescence when it originates from a triplet state. Due to prior loss through vibrational relaxation, the emitted photon always has lower energy, and thus a longer wavelength, than the absorbed one. The difference between the absorbed and emitted photon energy is known as the Stokes shift. Following the fact that electrons are much lighter and move much faster than nuclei, Franck-Condon principle states that electronic transitions occur so rapidly that nuclear positions remain essentially unchanged. Consequently, these transitions are represented as vertical lines on Jablonski diagram [33].

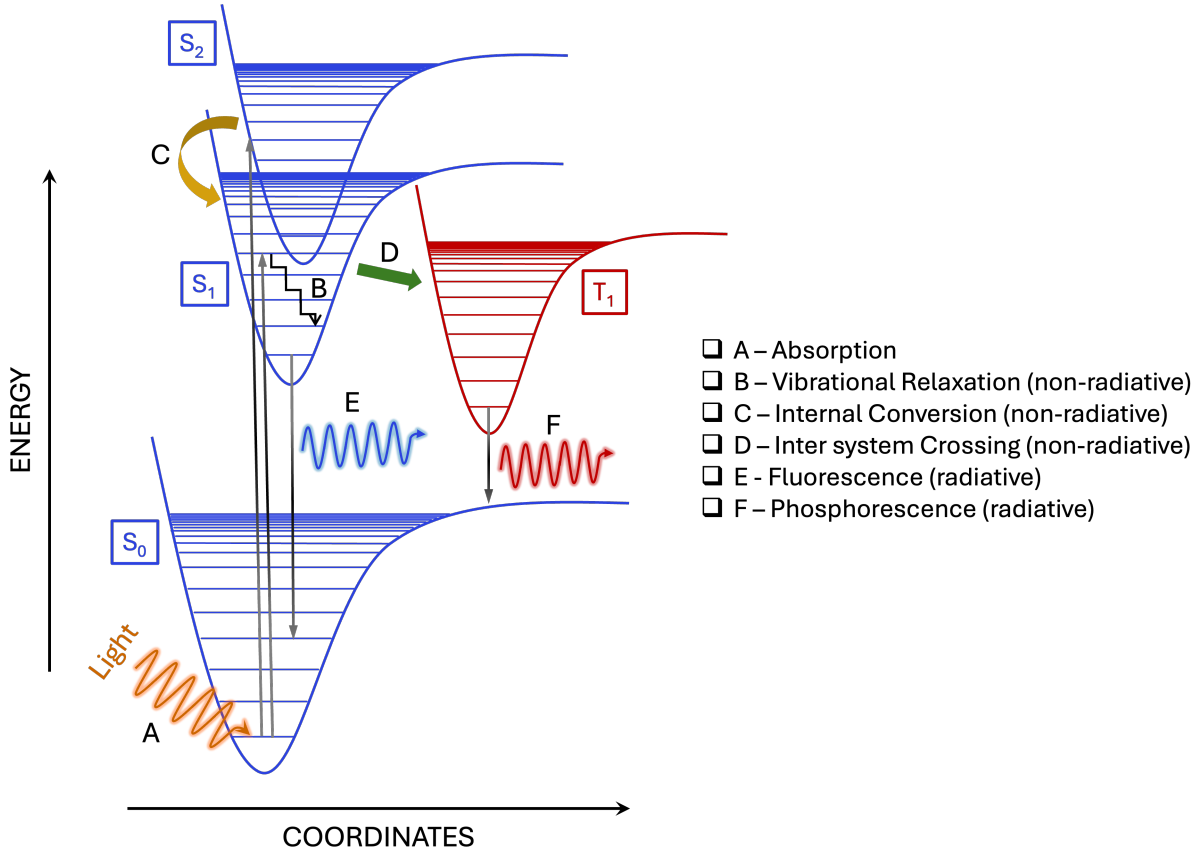


Figure 1.1: Jablonski diagram.

These photophysical processes occur across a wide range of timescales. Energy absorption is the fastest, taking place within femtoseconds, followed by internal conversion on the order of picosec-

onds. Fluorescence typically occurs within nanoseconds, whereas phosphorescence is much slower, lasting from microseconds to milliseconds [33, 34]. The extended lifetime of phosphorescence arises from its involvement of a spin-forbidden transition between states of different multiplicities. These timescales are influenced in part by the energy gaps between potential energy surfaces, which generally decrease at higher excitation energies. Consequently, Kasha's rule states that photon emission occurs predominantly from the lowest excited states, florescence from S_1 and phosphorescence from T_1 [35].

Beyond their description in terms of spin multiplicity, the excited states involved in these photophysical processes can also be understood in terms of the molecular orbitals participating in the electronic excitation, which provides a chemically intuitive framework for interpreting the transitions shown in a Jablonski diagram. In organic molecules, σ orbitals arise from head-on overlap of atomic orbitals and form the framework of single bonds, π orbitals result from side-on overlap and are characteristic of multiple bonds and conjugated systems, and nonbonding (n) orbitals correspond to localized lone pairs on heteroatoms. Electronic excitation maybe equivalently described involving the promotion of an electron from one of these occupied orbitals to a higher-energy antibonding orbital. Although transition probabilities are formally governed by symmetry-based selection rules, this orbital-based description provides clear insight into the energies and spectroscopic behavior of different excitations.

One of the highest-energy electronic excitations is the $\sigma \rightarrow \sigma^*$ transition, in which an electron is promoted from a σ bonding orbital to a σ^* antibonding orbital. Such transitions are characteristic of saturated molecules such as alkanes and occur in the far-ultraviolet region; consequently, they are rarely observed in conventional UV-visible spectroscopy. Transitions involving nonbonding (n) orbitals, including $n \rightarrow \sigma^*$ and $n \rightarrow \pi^*$ excitations, are commonly found in heteroatom-containing molecules such as alcohols, amines, and carbonyl compounds. Although these transitions fall within the UV range, they are typically weak due to limited orbital overlap and their partially symmetry-forbidden nature. In contrast, $\pi \rightarrow \pi^*$ transitions are characteristic of unsaturated and conjugated systems, including alkenes, dienes, and aromatic molecules [32]. These transitions occur at moderate energies and are usually intense, as they are symmetry-allowed and involve substantial redistribution of electron density, leading to large transition dipole moments.

On the top of these, when electronic excitation results in a very large redistribution of electron density, the process is classified as a charge-transfer (CT) transition. In such transitions, electron density is effectively shifted from an electron-rich (donor) region to an electron-poor (acceptor) region, either within a single molecule or between interacting molecular species.

1.2 Aromatic and Nonaromatic Photophysics

A molecule's photophysical behavior is inherently determined by its structure. Biomolecules, like proteins, feature intrinsic light absorption or emission, but only a small subset of their chemical groups contribute efficiently to absorption and fluorescence in the near UV range. Most amino acid

residues are optically silent in near UV, while a select group of molecular motifs act as endogenous chromophores (motifs that absorb light of a given wavelength) and, in rarer cases, fluorophores (motifs that fluoresce). Distinguishing between structural elements that merely absorb light and those that can re-emit it is essential for accurately interpreting spectra [36].

Previous studies have consistently shown that proteins exhibiting significant intrinsic fluorescence share common structural features that stabilize their electronic states, thereby enhancing absorption and emission in the UV-visible range [37, 38]. One of the fundamental intramolecular mechanisms for stabilizing organic systems is extended conjugation, which involves a sequence of alternating single and double bonds that allows delocalization of π electrons across a molecular segment. Conjugation lowers the energy gap between the highest occupied and lowest unoccupied molecular orbitals (HOMO-LUMO), enabling excitation by lower-energy photons [37]. This is illustrated in Figure 1.2.

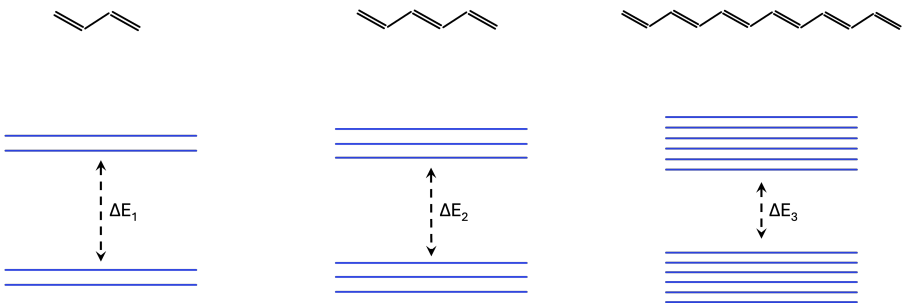


Figure 1.2: Conjugation: energy gap decreases with increase of degree of conjugation (from left to right).

In proteins devoid prosthetic groups, conjugation arises exclusively from presence of aromatic amino acids. Therefore, in the rest of this thesis, nonconjugated proteins are referred to as nonaromatic protein.

Aromatic residues, including tryptophan (maximum absorption/emission: $\approx$ 280 nm/350 nm), tyrosine (maximum absorption/emission: $\approx$ 275 nm/305 nm), and phenylalanine (maximum absorption/emission: $\approx$ 257 nm/280 nm) [33, 39, 40], possess planar and cyclic π systems that follow Hückel's (4n+2) rule [41]. This structural arrangement (illustrated in Figure 1.3, facilitates strong, symmetry-allowed $\pi \rightarrow \pi^*$ transitions, enabling the excited states to decay efficiently via radiative pathways. Consequently, over the years, these aromatic amino acids have been recognized as the primary contributors to the absorption of UV radiation at wavelengths longer than 250 nm (E $<$ 5 eV) and to emission in the near-UV region. As a result, they have been extensively studied and exploited in a wide range of applications [36, 42, 43, 44].

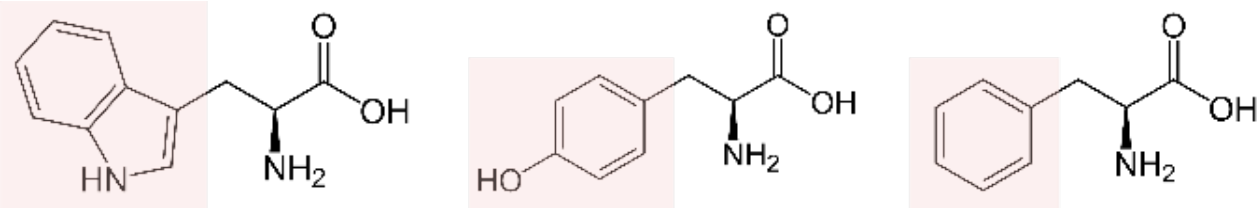


Figure 1.3: Aromatic amino acids: tryptophan on left panel, tyrosine in the middle and phenylalanine on the right panel.

Recent studies, nonetheless, challenge this view, showing that even nonaromatic systems with no π-conjugation, including proteins [1, 45, 46, 47, 48, 49, 50], supramolecular assemblies constituting amino acids[2, 51], polyamino acids [52, 53], amino acid crystals [54, 55], peptides [56, 57], and sugar polymers[58] can exhibit unconventional absorption and emission in the near-UV and visible region of the spectrum. There have also been reports of so-called cluster-triggered emission in a wide variety of supra-molecular assemblies [59, 60, 61, 62]. This unexpected behavior has prompted growing interest in reexamining the photophysical properties of biomolecules previously considered optically inactive in this range.

Several mechanisms have been proposed to explain the origin of these unconventional optical properties. Near-UV (340-380 nm) excitations in amyloid protein structures have been attributed to proton transfer between the N- and C-termini [47], as well as to the formation of stabilized $n \rightarrow \pi^*$ states resulting from distortions of amide and carbonyl (C=O) bonds within carboxyl groups [38]. In the latter case, luminescence becomes possible because rigid secondary structures hinder access to nonradiative decay pathways [38]. Furthermore, carbonyl bond elongation has been identified as a principal nonradiative decay mechanism in peptides [63]; when C=O stretching is artificially constrained, the excited state becomes more stable and its lifetime significantly increases. Enhanced fluorescence in protein-like structures has also been linked to short hydrogen bonds, which restrict C=O elongation and thus suppress nonradiative relaxation [55]. Additional studies suggest that polylysine and other polyamino acid systems exhibit intrinsic visible emission through clustering-triggered emission (CTE) [52], and that environment-dependent crystal packing can induce or modulate such emission [54]. Despite these findings, different systems appear to exhibit intrinsic absorption and emission arising from distinct molecular origins, which may not be shared across all materials. Therefore, the general mechanism underlying these unconventional photophysical properties remains unresolved.

1.3 The Scope of the Study

In this study we investigate the origin of unconventional photophysical properties of nonaromatic proteins, using α_3C protein as a case study. The α_3C protein is a synthetic, monomeric three helix bundle protein, indicated in Figure 1.4, with over 50% of its constituent amino acid charged at neutral pH. It has a total of 67 residues, 36 of which are: 17 lysines (Lys), indicated in green, and two arginine (Arg) in blue, positively charges, and 17 glutamic acid (Glu) in orange, negatively

charged.

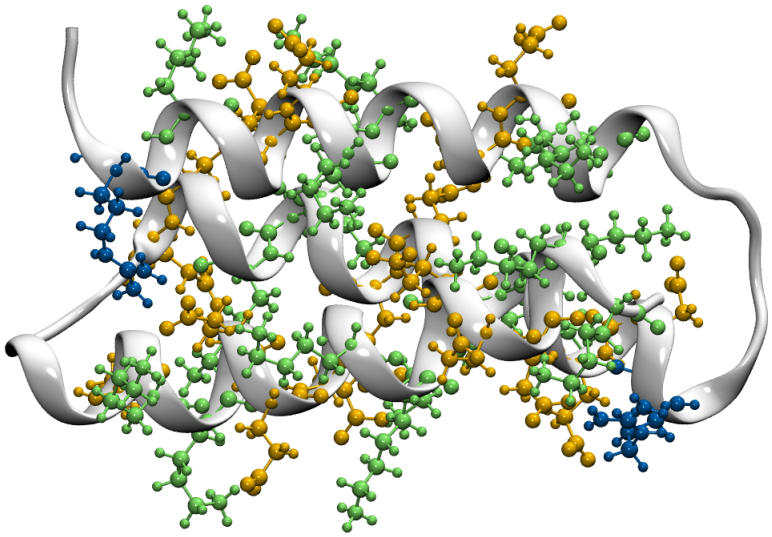


Figure 1.4: Structural representation of the α_3C protein illustrating its three-helix bundle secondary structure. The protein backbone is shown in cartoon style (light grey), while charged residues are displayed in ball-and-stick representation, with glutamic acid residues colored orange, lysine residues green, and arginine residues blue.

Given the preceding discussion, one would expect that the α_3C protein would not be optically active beyond 250 nm - in other words, none of its electronic states would be excited by a photon with a wavelength longer than 250 nm. However, recent experiments by Prasad et al. [1] revealed that this protein absorbs over a broad range of spectrum, from 250 nm to 800 nm, as indicated on the left panel of Figure 1.5.

UV/visible absorption spectra were measured at room temperature, for α_3C protein dissolved in ionized water at different concentrations (5-105 μM), with additional studies on the pH dependence at 85 μM. The the protein solution was found to absorb moderately in range 250-300 nm, with a tail extending to 800 nm. The molar extinction coefficient at 280 nm was found to be $\varepsilon = 4531 \pm 133$ $M^{-1}cm^{-1}$ which is around four fifths that of Tryptophan ($\varepsilon \approx 5500$ $M^{-1}cm^{-1}$). The results further revealed a linear relationship between absorbance and concentration at 270, 350, and 700 nm, indicating that the protein molecules remain in their monomeric form over the studied concentration range. The absorption spectrum is smooth, and potential scattering contributions were assessed by simulating Rayleigh scattering spectra, which scale inversely with the fourth power of the wavelength. Overlaying these simulations with the experimental absorption spectrum confirmed the absence of significant scattering effects.

Further experiments conducted five years later by Kumar et al. [2] have shown that α_3C emit weakly in near UV and visible region, between 310 nm and 550 nm when excited at 295 nm. In their

work, they investigate the effect of α_3C protein on Tryptophan fluorescence. This was achieved through a comparative analysis time-resolved fluorescence, of N-Acetyl-L-tryptophanamide (NATA) at 20 μM in the presence of either the α_3C or the α_3W mutant (a tryptophan-containing variant of α_3C), at concentrations up to 100 μM. It is worthy to note that the samples were excited exclusively at 295 nm. The resulting emission spectrum for α_3C is shown on the right panel of Figure 1.5.

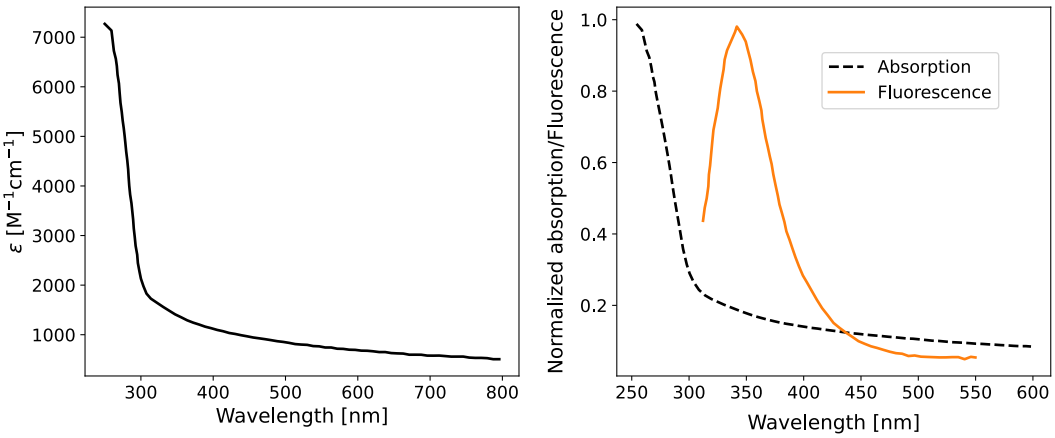


Figure 1.5: Experimental spectra of α_3C protein. The left panel shows absorption spectrum from 250 nm to 800 nm, adapted from reference [1]. The right panel shows absorption (black dashed line), and emission (orange line) spectrum in range 310-550 nm upon excitation at 295 nm, adapted from reference [2].

In their work, Prasad and collaborators proceeded with investigation of the origin of unconventional absorption in proteins through simulations of lysine and glutamic acid dimers in vacuum. The choice of using lysine and glutamic acid was motivated by the abundance of these residues in proteins and their propensity to form salt bridges or spatially proximal interactions within approximately 6 Å. The authors extracted lysine-glutamic acid dimer structures from molecular dynamics simulations of the α_3C protein in aqueous solution, at varying inter-residue separations, and capped the dangling bonds with glycine. Time-dependent density functional theory (TD-DFT) calculations were then performed on these isolated dimers in vacuum. Their results revealed that dimers separated by less than 3 Å exhibited electronic transitions only below 300 nm, while those separated by 3-4 Å showed transitions extending beyond 300 nm, reaching up to 400 nm. Dimers separated by 5-6 Å displayed transitions in the whole visible range. With these findings, summarized in Figure 1.6 (adapted from [1]), Prasad and collaborators proposed that this absorption arises from charge transfer between glutamic acid and lysine residues that exhibit spacial interactions between 5-6 Å.

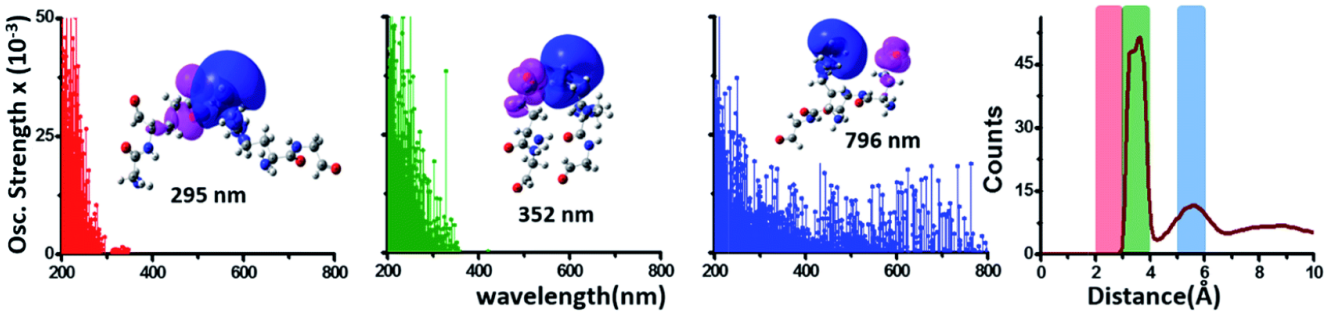


Figure 1.6: Simulated absorption spectra from reference [1], for distally separated lysine-glutamic acid dimers, chosen from shadded region of RDF (on the right panel): 2-3 Å in red, 3-4 Å in green and 5-6 Å in blue. Blue and pink lobes show regions with increase and decrease in electron density respectively.

While these results are intriguing, the computational methodology raises some concerns. First, the study assumes a priori that lysine-glutamic acid interactions are the relevant contributors to the observed absorption, potentially overlooking other possible relevant structures. Second, the electronic structure calculations were performed on dimers extracted from the protein and placed in vacuum, thereby neglecting the complex environmental effects present in the native protein and solvent. This limitation becomes increasingly significant as the separations distance between the dimers increases. At larger separations, such as around 6 Å, the neglected environment includes molecules (either other protein residues or water) that are directly bonded to the each of the two groups which could play an essential role in mediating their electronic interactions. Therefore we think these structures used do not represent the real interactions in the system.

In this work, we aim to investigate the optical properties of the α_3C protein using a more agnostic and data-driven approach. Specifically, we address two central questions: (i) what is the origin of the unconventional absorption observed in the α_3C protein, and (ii) which structural factors stabilize its excited state sufficiently to permit light emission, and, related to this, what the primary nonradiative vibrational relaxation pathways are. To this end, we employ unsupervised machine learning techniques to identify the most relevant intermolecular interactions within the system. Electronic structure calculations are then performed on these identified local environments, explicitly accounting for their respective environment. While this inclusion of environmental effects increases the computational cost, it is mitigated through the use of a validated semi-empirical electronic structure method. The theoretical foundations of these methods are presented in Chapter 2, while Chapter 3 details the procedures used to estimate the absorption properties of different local structure in α_3C and elucidates the origins of its near-UV absorption. Chapter 4 focuses on the characterization of excited states, and the final chapter presents the final conclusions and future perspectives of this work.

Chapter 2

Methods

In this chapter, we detail the theoretical foundations of the computational methodologies employed in this thesis. We start with classical molecular dynamics simulations, a method that allows us to generate atomic positions of the protein in its aqueous environment. Then we discuss machine learning methods used to describe the local atomic environments in an agnostic fashion, and to identify the relevant structural patterns. This is followed by an overview of the quantum mechanics/molecular mechanics (QM/MM) method, which was applied to selected configurations of interest, on which electronic structure calculations were done, enabling inclusion of environmental effects. Finally, we describe the methods used to conduct the ground and excited state electronic structure calculations and dynamics.

2.1 Classical Molecular Dynamics

Molecular dynamics (MD) is a computational method used to study the motion of atoms within molecular systems. It provides detailed structural and dynamical information that is valuable for a wide range of applications, including explaining the mechanisms underlying biological processes like protein folding and enzyme catalysis [64, 65], guiding the design and discovery of new materials and drugs, and complementing experimental observations through atomistic insights. In principle, atoms (comprising of protons, neutrons and electrons) are quantum mechanical particles, and their motion should be ideally modeled quantum mechanically. However the computational cost of such a quantum treatment becomes prohibitive for large systems such as biological macromolecules. In the context of molecular dynamics, therefore, several approximations are made, to enable the study of such systems. The nuclei are modeled as classical particles that move according to Newton's law of motion, under the influence of an effective potential energy which is represented by mathematical equations known as a force field. Over the years, various force fields have been built AMBER [66], GROMOS [67] and CHARMM [68] for molecules and SPC [69], TIP3P [70] and TIP4P [71] specifically designed to describe the properties of water. These models differ in their specific combinations of parameters, but the basic building blocks are similar. A force field considers two main types of interactions: bonded and nonbonded.

$$U = U_{\text{bonded}} + U_{\text{nonbonded}} \tag{2.1}$$

Bonded interactions include bond stretching and angle bending, typically modeled as harmonic potentials, as well as dihedral torsions, which are often represented by periodic functions. Some force fields also incorporate improper dihedrals to maintain the planarity of specific molecular groups [72, 73].

$$U_{\text{bonded}} = \sum_{\text{bonds}} k_b(r - r_o)^2 + \sum_{\text{angles}} k_\theta(\theta - \theta_0)^2 + \sum_{\text{dihedrals}} V_n[1 + cos(n\phi - \gamma)] \tag{2.2}$$

Non-bonded interactions consist of electrostatic forces described by Coulomb's law, and van der Waals forces, commonly modeled using the Lennard-Jones potential:

$$U_{\text{nonbonded}} = \sum_{i<j} \left[4\epsilon_{ij} \left(\frac{\sigma_{ij}^{12}}{r_{ij}^{12}} + \frac{\sigma_{ij}^{6}}{r_{ij}^{12}} \right) + \frac{q_i q_j}{4\pi\epsilon_0 r_{ij}} \right] \tag{2.3}$$

The force field parameters, such as bond force constants, equilibrium geometries, atomic charges, and van der Waals coefficients, are carefully derived from experimental data and/or high-level quantum mechanical calculations to ensure accuracy of MD simulations.

Essentially, each atom move in response to the force arising from interactions with all other atoms in the system. The force on the i^{th} atom in a system of N atoms is computed from the potential energy of the system with respect to position of atoms, according to the equation below:

$$\vec{F_i} = -\frac{\partial U(\vec{r_1}, \ldots, \vec{r_N})}{\partial \vec{r_i}} \tag{2.4}$$

which gives it an acceleration of

$$a_i = \frac{\vec{F_i}}{m_i} \tag{2.5}$$

where m_i is the atom's mass. Numerically, the equations of motion are solved using finite difference methods by discretizing time into small intervals, Δt, typically on the order of 1-2 femtoseconds to accurately capture the fastest atomic motions, such as bond vibrations. One commonly used integration method is the Verlet [74] and velocity Verlet algorithm[75], which integrate Newton's law using Taylor expansion of position. For velocity Verlet, the position of atom i is updated as follow:

$$r_i(t + \Delta t) = r_i(t) + v_i(t)\Delta t + \frac{1}{2}a_i(t)\Delta t^2 \tag{2.6}$$

forces for $r(t + \Delta t)$ are updated, and acceleration at $t + \Delta t$ is calculated as

$$\vec{a_i}(t + \Delta t) = \frac{\vec{F_i}(t + \Delta t)}{m_i} \tag{2.7}$$

and the velocity can be calculated in the following way:

$$v_i(t + \Delta t) = v_i(t) + \frac{1}{2}[a_i(t + \Delta t) - a_i(t + \Delta t)]\Delta t \tag{2.8}$$

To have a control on thermodynamic conditions of the system during MD simulations, thermostats for temperature control, like Nose-Hoover [76] and velocity-rescale [77], and barostat for pressure control like Parrinello-Rahman [78] are used.

As the MD simulation progresses, the system evolves over time, generating a trajectory - a sequence of time-ordered atomic positions and velocities. Each time step in the simulation provides a complete set of atomic coordinates, representing one specific configuration of the system in phase space. By sampling configurations that are sufficiently separated in time along the trajectory, statistical correlations between successive frames are minimized. These uncorrelated configurations serve as the basis for subsequent analyses, facilitating the computation of the system's structural and dynamical properties.

2.2 Unsupervised Learning Techniques

MD simulations generate vast amounts of data, capturing the time evolution of atomic positions. For a system containing N atoms, the configurational space is $3N$-dimensional. However, it is often seen that the most relevant structural and dynamical characteristics reside on a low dimensional manifold within this high-dimensional configurational space [79, 80, 81, 82, 83, 84]. Direct analysis of such high-dimensional data is challenging, as it is difficult to discern which degrees of freedom are most relevant to the system's behavior [85, 86].

Traditionally, researchers extract chemically meaningful information via the construction of low dimensional descriptors, referred to as collective variables (CVs), based on the features of atomic configurations, such as interatomic distances, bond angles, or hydrogen bond networks, and monitor their evolution throughout the trajectory. Sets of such CVs have been shown to succeed at capturing the effects of local, pair-wise interactions [85, 87, 88, 89] and allow projecting the system's probability density onto reduced dimensional support. The equilibrium probability distribution projected onto the CV space is given by :

$$P(s) = \frac{1}{Z} \int dr \delta(s - s(r)) e^{-\beta U(r)} \tag{2.9}$$

where $U(r)$ is the potential energy and Z the partition function. One can build an effective free energy surface (FES) from this through the Boltzmann relation as :

$$F(s) = -k_B T ln[P(s)] + C \tag{2.10}$$

where k_B is the Boltzmann constant, T is the temperature, and C is an arbitrary reference constant.

The FES provides an interpretable thermodynamic landscape, highlighting metastable states (minima), transition barriers, and connectivity between basins [90]. From a data-analysis perspective, constructing an FES and estimating densities are equivalent tasks - both recover the underlying probability distribution on the reduced manifold, where clusters correspond to high-density (low-free-energy) regions.

However, traditional CVs are often limited in their ability to describe collective phenomena governed by many-body correlations, such as conformational rearrangements and phase transitions. To overcome these limitations, machine learning-based descriptors have emerged as powerful alternatives capable of systematically capturing complex, many-body interactions [91, 92, 93]. Descriptors built on symmetry-preserving bases [94, 95] and cluster expansions [96]) have become increasingly popular [84, 97]. These descriptors are often themselves inherently high-dimensional, containing thousands of components per atomic site. While this expressiveness allows the encoding of complex correlations, it often limits their use as CVs themselves. Consequently, additional dimensionality reduction is frequently performed to extract a small number of low-dimensional, physically meaningful variables that capture the essential slow modes of the system [98].

Building on these foundations, we employed an unsupervised machine learning approach to systematically uncover structural features relevant to the photophysics under investigation. This method enabled the identification of recurring and representative structural motifs, which were subsequently used for further analysis and quantum mechanical calculations. This workflow consisted of three key steps: (1) the application of a local atomic descriptor to quantitatively characterize the atomic environments, (2) estimation of the dataset's intrinsic dimensionality (ID) and the implementation of dimensionality reduction strategies, and (3) the use of clustering algorithms to group structurally similar environments to get a low-dimensional projection of the configurational probability density.

2.2.1 Smooth Overlap of Atomic Positions

Smooth overlap of atomic positions (SOAP) descriptor has emerged as a powerful technique for encoding information about local environments in a wide variety of molecular systems, including organic molecules [99], biological systems [100, 101], solid-state systems [102, 103, 104], and also in liquid water [79, 85, 105, 106, 107, 108]. The SOAP descriptor captures the spatial distribution of atoms surrounding a central atom by representing the positions of neighboring atoms within a defined cutoff radius. Its key strength lies in its ability to construct a smooth, continuous representation of the local atomic environment that is both physically meaningful and robust.

Being physically meaningful in this context requires preserving the fundamental symmetries and invariances dictated by the underlying physics. These ensure that equivalent physical configurations are represented identically, independent of arbitrary choices such as coordinate origin (translational invariance), molecular orientation (rotational invariance), or atom indexing (permutational invariance). How this is brought to effect, is rooted in the mathematical construction of the descriptor:

Given an atomic environment χ around a central atom, one determines the local density as a sum of Gaussian functions with variance σ^2 centered on each of its neighbors including the central atom itself,

$$\rho_\chi(r) = \sum_{i \in \chi} \exp\left(-\frac{|r_i - r|^2}{2\sigma^2}\right) \tag{2.11}$$

here, r_i are the positions of neighboring atoms. The use of relative position naturally brings translational invariance. Moreover, representing the local atomic density as a sum over neighbors guarantees that swapping the indices of atoms does not affect the calculation, thereby providing permutational invariance.
The atomic neighbor density can be expanded in terms of radial basis functions and spherical harmonics Y_{lm} such that,

$$\rho_\chi(r) \approx \sum_{n=0}^{n_{max}} \sum_{l=0}^{l_{max}} \sum_{m=-l}^{l} c_{nlm} g_n(r) Y_{lm}(\theta, \phi) \tag{2.12}$$

where the c_{nlm} are the expansion coefficients. From the atomic density representation, one can construct a rotationally invariant power spectrum vector $\mathbf{p}$ whose elements are defined as:

$$p_{nn'l} = \pi \sqrt{\frac{8}{2l+1}} \sum_m (c_{nlm})^\dagger c_{n'lm} \tag{2.13}$$

However, this power spectrum is high-dimensional, resulting in large and complex datasets that hinder straightforward interpretation. Consequently, dimensionality reduction techniques are necessary to facilitate meaningful analysis.

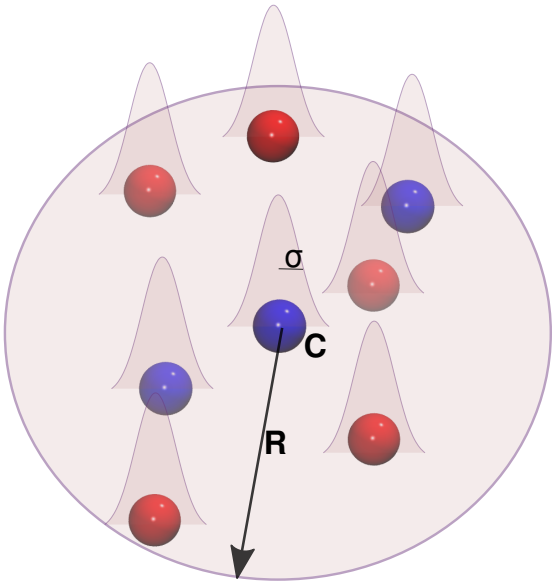


Figure 2.1: A description of an example of environment χ, with two atom types (red and blue), centered at the center of atom c, with radius R, to be described by SOAP using Gaussian functions with half-width σ.

2.2.2 Intrinsic Dimension

In many high-dimensional datasets, correlations among degrees of freedom constrain the data to a lower-dimensional manifold, meaning that the system's essential features can be captured by a reduced set of variables. A similar behavior is expected for the SOAP spectrum (Eqn. 2.13) defined in the previous section, where correlated variations among atomic environments confine the data to a lower-dimensional manifold. Identifying this manifold enables accurate estimation of the probability density in SOAP space and the detection of dominant structural motifs.

The dimensionality of this manifold - the intrinsic dimension (ID) - quantifies the minimum number of coordinates required to represent the dataset without significant loss of information [109, 110, 111]. It provides a judicious estimate of the number of degrees of freedom governing the structural variability of the system and is crucial crucial for estimating probability densities on the correct low-dimensional support. However, estimating the ID is a challenging task because data density typically varies across the manifold, making it necessary to jointly estimate both the intrinsic dimension and the local data density for accurate characterization [84].

Several methods exist for estimating the dimensionality of a manifold, often leveraging projection techniques. One of the common approaches is Principal Component Analysis (PCA) [112], which projects the data onto a lower-dimensional subspace by minimizing the reconstruction error. PCA identifies directions (principal components) that capture the greatest variance in the data, allowing for an estimate of the intrinsic dimensionality based on the number of components needed to retain a significant portion of the variance. Another widely used approach is Uniform Manifold Approximation and Projection (UMAP) [113], employs a nonlinear projection that maintains the topological relationships among data points. Other methods include Isomap [114], which preserves geodesic distances along the manifold, and t-SNE [115], which emphasizes local neighborhood relationships.

For the purpose of this work, the recent Two-Nearest Neighbor (Two-NN) method, developed by Facco and the collaborators [116] is used. Here the ID is computed based on information from the first and second nearest neighbors of each data point. Considering just the first and second neighbor, reduces the effect of non homogeneous density. It was shown in their work that the ratio of the second to the first nearest neighbor distances ($\mu_i = r_i^2/r_i^1$) follows a specific distribution, which is not density dependent namely,

$$P(\mu_i) = \frac{d}{\mu_i^{d+1}} \tag{2.14}$$

where d is the ID. Assuming independence of sampled ratios μ_i, the ID can be estimated through a maximum likelihood technique as:

$$d = \frac{N}{\sum_{i=1}^{N} \log(\mu_i)} \tag{2.15}$$

where N is the total number of samples in the dataset.

2.2.3 Density Peak Advanced Clustering

Once the ID is determined, the effective free energy surface (FES) with the relevant states may be recovered in the high-dimensional SOAP space by estimating the probability density on the corresponding 'ID'-dimensional manifold. This can be achieved through clustering methods, which partition the configurational space based on similarities in local structure or density [117]. Among these, density-based clustering algorithms are particularly suited for molecular data, as they identify clusters as peaks in the underlying probability distribution without requiring explicit dimensionality reduction, thereby preserving the full structural information. In this work, we employ the Density Peak Advanced (DPA) algorithm [3], which integrates intrinsic dimension estimation with adaptive nearest-neighbor density estimation to robustly identify clusters and their boundaries in high-dimensional spaces.

A. Density Estimation

Density estimation methods fall into two broad categories: parametric and nonparametric. Parametric methods assume a predefined functional form for the distribution, which can be limiting for complex, unknown distributions [118]. Nonparametric methods, such as the classic k-nearest neighbor (k-NN) estimator [119], do not impose such assumptions, making them more flexible for heterogeneous data.

However, traditional k-NN methods have a fixed neighborhood size k, which make them unsuitable for data with highly variable local densities. To overcome this, we utilize the Point Adaptive k-nearest neighbor (PAk) estimator developed by [120]. Unlike standard k-NN, PAk determines an adaptive neighborhood size k_i for each data point i, selecting the largest k for which the local density remains statistically constant. This is determined via a likelihood ratio test comparing the density estimates at k and $k+1$ with a high confidence threshold (p=10^{-6}). Having obtained k_i, it is then used to calculate the density of point i in low dimension d equal to the intrinsic dimension in the following way:

$$\rho_i = \frac{k_i}{r_{k_i}^d} \tag{2.16}$$

B. Clustering

From these density estimates, the algorithm constructs a density topography to identify clusters as local maxima in the transformed density function:

$$g_i = log(\rho_i) - \epsilon_i \tag{2.17}$$

where ϵ_i accounts for the variance in $\log(\rho_i)$, ensuring that only stable, high-density peaks are selected. Each non-peak point is assigned to the cluster of its nearest neighbor with a higher g-value. Then from the boundary data points between clusters, saddle points are identified. These are points with locally maximal g values along cluster borders. Genuine clusters are validated by requiring the difference in log-densities between cluster centers and saddle points to exceed a

threshold Z times the combined density fluctuations, where Z controls statistical confidence and is the sole adjustable parameter in DPA.

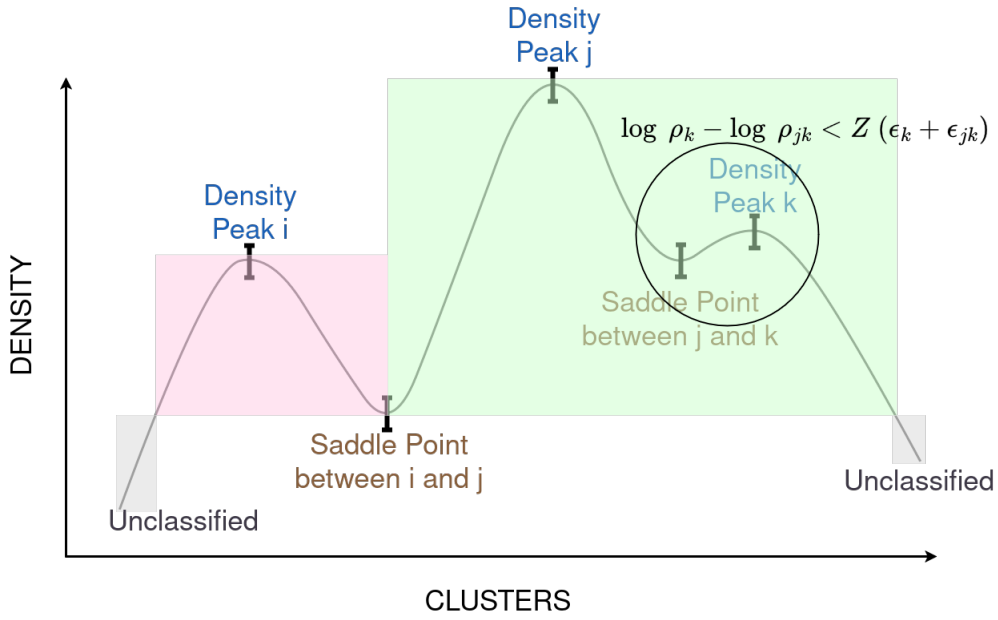


Figure 2.2: A 2D description of an example of density topography and clusters selection in DPA scheme, adapted from reference[3].

2.3 Quantum Mechanics/Molecular Mechanics (QM/MM)

From the methods described above, classical MD simulations provided representative configurations of the system, from which unsupervised learning protocol reveled relevant local interactions within the system. However, the information accessible from classical MD alone is insufficient to investigate the optical properties, specifically, the absorption and emission characteristics of the α_3C protein, which is the ultimate objective of this study. As discussed earlier in section 2.1, classical MD does not treat electrons explicitly; it provides only the positions and velocities of atomic nuclei evolving on a predefined potential energy surface. Consequently, electronic distributions and their rearrangements cannot be captured within this framework. To explore the electronic structure and related spectroscopic properties of the system, it is therefore necessary to employ quantum mechanical (QM) methods beyond the scope of classical MD.

Available quantum mechanical methods are computationally demanding and thus impractical for investigating large systems [121, 122, 123]. Also, isolating local interactions from their environment and modeling them in vacuum can alter their intrinsic properties significantly [124, 125, 126]. To overcome these limitations and accurately characterize the electronic structure of our system, we employ a hybrid quantum mechanics/molecular mechanics (QM/MM) approach. This method allows the region of interest to be treated at the quantum level while simultaneously incorporating the effects of its surrounding environment with computationally efficient molecular mechanics. The

portion treated quantum mechanically is referred to as the QM region, whereas the surrounding environment described using classical mechanics constitutes the MM region.

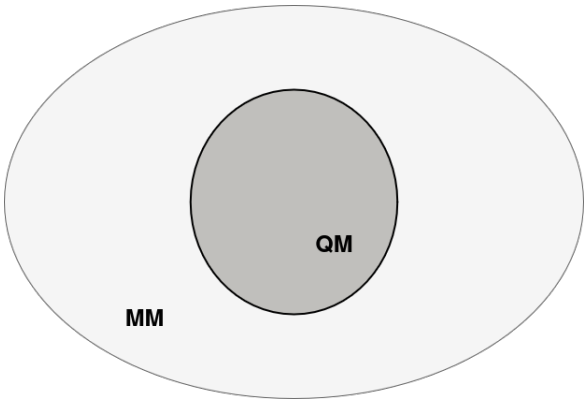


Figure 2.3: A QM/MM system representation.

Several strategies have been developed to compute the total energy of the QM/MM system. The first of these is subtractive scheme [127]. In this method, three calculations are done: 1) molecular mechanical energy of the entire system, 2) quantum mechanical energy of the QM region and 3) the molecular mechanical energy of the QM region. The total energy of the QM/MM system is then obtained by subtracting the MM energy of the QM region from the sum of the first two terms, to avoid double consideration of QM energy, as expressed by the following equation:

$$E_{\text{QM/MM}}(S) = E_{\text{MM}}(S) + E_{\text{QM}}(QM) - E_{\text{MM}}(QM) \tag{2.18}$$

This is a simple way for finding the total energy of the system, easy to implement, with no artificial QM/MM interaction problems at the interface and can be easily generalized to multiple layers QM/MM systems, like in the ONIOM method [128, 129]. However, it treats the QM/MM boundary classically. This means, the quantum calculations on the QM region are performed with no consideration to existence of MM region. In other words, the resulting electronic density is similar to that in vacuum. Therefore, this method is not commonly used. In recent improvements, the quantum calculations in QM region are done in presence of MM charges [130]. This way they polarize the QM region, and the method is more accurate than where the interactions are simply classical.

The second way of computing the QM/MM system's total energy is by additive scheme. With this method, the interactions between QM and MM regions are explicitly computed and the total energy of the system is obtained by taking the sum of MM energy, QM energy and that of QM/MM interactions:

$$E_{\text{QM/MM}}(S) = E_{\text{MM}}(MM) + E_{\text{QM}}(QM) + E_{\text{QM-MM}} \tag{2.19}$$

This QM-MM coupling term, $E_{\text{QM-MM}}$, constitutes bonded interactions (when present), electrostatic and van der Waals interactions between QM and MM atoms. If at the boundary there is a

covalent bond between an atom in QM region and another atom in MM region, then the bonded interaction is necessary and is computed classically. For the cases where the QM region - MM region are like solute - solvent, which does not involve any covalent bond between them at the boundary, the bonded energy is not necessary, and the QM-MM interactions is given by just the sum of non-bonded contributions. The Van der Waals interactions between QM and MM atoms are also computed at MM level. In this method, the most important part of the QM/MM interactions is electrostatics, and the name of the scheme is derived from how it is treated.

The electrostatic interactions between the two regions, MM point charges and QM charge density can be treated at three deferent level. The lower level is mechanical embedding where the electrostatic interactions are modeled as MM-MM electrostatic interactions. The QM part is not polarized by the MM environment. The second level is electrostatic embedding scheme [131], where the QM-MM electrostatic interaction is treated at quantum level. The QM calculations are done in presence of MM point charges, that way the QM charge distribution reacts to point charge distribution and is polarized by it. However, the QM region does not polarize the MM region. The third level which is the highest, and in principle the best approximation, is using polarized embedding schemes [132]. This makes the rigid MM charges model flexible that there is mutual polarization between QM and MM regions. This is however substantially more expensive, as the mutual polarization is done in a self-consistent manner. Also this method requires using a polarizable force field to describe the MM region, which are not available yet for all kinds of system. Therefore, among the three methods, electrostatic embedding is the most widely used and it is the method used for this study.

In QM/MM calculations, overpolarization of the QM region by its environment is unavoidable, as QM electron density can penetrate the MM region due to the inadequate description of Pauli repulsion in MM models, such as Lennard-Jones potentials. This effect is particularly pronounced when MM atoms are close to the QM region or when diffuse basis sets are employed [133, 134], and is commonly mitigated by screening short-range electrostatic interactions [135].

A further challenge occurs at covalent bonds crossing the QM/MM boundary, which create unsaturated valence orbitals in the QM region. This is typically resolved by introducing a link atom, usually hydrogen to saturate the orbital. To minimize artificial interactions, the charge of the boundary atom is redistributed among its neighboring atoms, preserving the total charge of the system [135, 136, 137].

2.4 Density Functional Tight Binding

The QM/MM framework enables the investigation of the electronic structure of the quantum mechanical (QM) region embedded within a classical molecular mechanics (MM) environment. The MM environment is modeled using molecular dynamics (MD) simulations, as described above, while the QM region is treated fully quantum mechanically, allowing for an accurate description of electronic properties in the context of its surrounding environment.

One of the common approaches to calculate the systems' electronic structure is through density functional theory (DFT). DFT determines the ground-state electronic structure of a system using the electron density as the fundamental variable [138]. The method is founded on two key theorems formulated by Hohenberg and Kohn [139]. The first theorem establishes that all ground-state properties of a many-electron system are uniquely determined by the electron density. The second theorem introduces a variational principle, stating that for any trial electron density that integrates to the correct number of electrons, the corresponding total energy will be greater than or equal to the true ground-state energy. Consequently, in DFT, the total energy of the system is expressed as a functional of the electron density, which is then minimized self-consistently to obtain the ground-state energy and density.

$$E_{\mathrm{DFT}} = E[\rho(r)] \tag{2.20}$$

Though widely used in materials science [140, 141, 142], catalysis [143, 144, 145, 146], spectroscopy and photochemistry [147, 148], and in studying water and aqueous systems [149], DFT is limited by the size of systems it can be applied to, as noted previously. For the large QM regions considered here, full DFT calculations quickly become computationally prohibitive [123].

To address this limitation, we employ Density Functional Tight Binding (DFTB), a DFT-based tight-binding approach. The tight binding method approximates electronic band structures in condensed-phase systems by assuming that electrons are primarily localized around their parent atoms but can hop to neighboring atoms [150]. This assumption makes the method very useful in modeling systems where electronic states are relatively localized, and significantly reduces computational cost. DFTB combines the computational efficiency of the the tight binding method with the accuracy of DFT, enabling the study of large systems at speeds at two orders of magnitude faster than conventional DFT while maintaining excellent agreement with DFT accuracy [151, 152, 153]. This make it particularly useful in modeling complex biomolecular systems [154, 155].

Like in DFT, density is a fundamental quantity for the DFTB method and it is decomposed into a fixed reference density and a small fluctuating term:

$$\rho(r) = \rho_0(r) + \delta\rho(r) \tag{2.21}$$

Here, the reference density, $\rho_0(r)$, represents the sum of atomic densities in a confined potential. The DFTB total energy is then obtained by taking the truncated Taylor expansion of DFT energy, around that reference density ρ_0, up to a chosen order: first order (DFTB1) [156], the second (DFTB2) [157] or third order (DFTB3) [158]. The third order approximation is mathematically expressed as:

$$E_{\mathrm{DFTB3}}[\rho_0 + \delta\rho] \approx E_0[\rho_0] + E_1[\rho_0, \delta\rho] + E_2[\rho_0, (\delta\rho)^2] + E_3[\rho_0, (\delta\rho)^3] \tag{2.22}$$

The $E_0[\rho_0]$ in equation 2.22 is the repulsive term, and can be expressed as a sum of pairwise repulsive potential energy:

$$E_0[\rho_0] = \sum_{A>B} E_{AB}^{\text{rep}} \tag{2.23}$$

and the first order term, $E_1[\rho_0]$ is the band structure energy:

$$E_1[\rho_0] = \sum_{i}^{\text{occ.}} n_i \langle \psi_i | H[\rho_0] | \psi_i \rangle \tag{2.24}$$

with $H[\rho] = T + V_{eff}[\rho]$ the Kohn-Sham Hamiltonian that includes the external potential, Hartree and the exchange-correlation contribution. It is important to note that these first two terms of equation 2.22, the repulsive term and the band energy, depend only on the reference density.

Using a linear combination of atomic orbital basis, molecular orbitals can be described as:

$$\psi_i = \sum_{\mu} c_{\mu i} \phi_\mu (r - R_A) \tag{2.25}$$

where R_A is a position vector of atom A. The elements of the zero-order Hamiltonian, $H^0 = H[\rho_0]$, are approximated as follow:

$$H^0_{\mu\nu} = \langle \phi_\mu | \hat{T} + V_{eff}[\rho_A + \rho_B] | \phi_\nu \rangle, \quad \mu \in A, \nu \in B \tag{2.26}$$

These pairwise terms in equation 2.26 and the overlap matrix terms: $S_{\mu\nu} = \langle \phi_\mu | \phi_\nu \rangle$, along with the repulsive terms in equation 2.23 are the semi-empirical terms which are precomputed from higher ab-intio methods and saved in the so called Slater Koster files [152, 159].

The second order term in the Taylor series, E_2, accounts for charge redistribution effects. Assuming that, the electronic density fluctuations on atom A can be represented by Mulliken charge [160], $\Delta q_A = q_A - Z_A$, the second term become:

$$E_2 = \frac{1}{2} \sum_{AB} \gamma_{AB} \Delta q_A \Delta q_B \tag{2.27}$$

where γ_{AB} represents the electron interactions of the Slater-type spherical charge densities on atom A and B. Since Mulliken charges depends on the molecular orbitals, evaluation of this second term require a self-consistent procedure. This term offers more accurate description of the charge transfer interactions [157].

The third order term, E_3, includes the fluctuations in Mulliken charges, capturing higher order polarization effects. It is expressed as:

$$E_3 = \frac{1}{3} \sum_{AB} \Gamma_{AB} (\Delta q_A)^2 \Delta q_B \tag{2.28}$$

where $\Gamma_{AB} = \left. \frac{\partial \gamma_{AB}}{\partial q_A} \right|_{q_A^0}$. Substituting relations all the terms in equation 2.22 gives:

$$E_{\mathrm{DFTB3}}[\rho_0+\delta\rho] = \sum_{A>B} E_{AB}^{\mathrm{rep}} + \sum_i \sum_{\mu\nu} n_i c_{\mu i} c_{\nu i} H^0_{\mu\nu} + \frac{1}{2}\sum_{AB}\gamma_{AB}\Delta q_A \Delta q_B + \frac{1}{3}\sum_{AB}\Gamma_{AB}(\Delta q_A)^2 \Delta q_B \tag{2.29}$$

The forces are computed by taking the derivative of the total energy with respect to atomic coordinates, R, subject to the Kohn-Sham orbital normalization constraint: $\int \psi_i \psi_i d^3r = 1$ as indicated in the following equation

$$F_{kx} = -\frac{\partial}{\partial R_{kx}}\left[E_{\mathrm{DFTB3}} - \sum_i n_i \epsilon_i \left(\sum_{AB}\sum_{\mu\nu} c_{\mu i} c_{\nu i} S_{\mu\nu} - 1\right)\right] \quad \forall k, x \tag{2.30}$$

where k is an atom index, and $x \in \{1, 2, 3\}$ a cartesian coordinate index.

The accuracy of the DFTB method strongly depends on the quality of the underlying parameter sets. Considerable effort has therefore been devoted to the development and validation of suitable parameters, as well as benchmarking DFTB against DFT and higher level methods, to assess its accuracy and applicability for ground state calculations across a wide range of systems.

Gauss and co-workers [152] introduced the OB3 parameter set for DFTB3, optimized for molecules containing C, H, N, and O. Benchmarks on small organic molecules indicated that it particularly describes well geometry for the nonbonded interactions like hydrogen bonding. While conventional DFT employing generalized gradient approximation (GGA) functionals (e.g., BLYP and PBE) generally yields higher accuracy when used with sufficiently large basis sets, DFTB3/OB3 often outperforms GGA-DFT calculations based on small basis sets, offering an advantageous balance between accuracy and computational efficiency. Similarly, Vuong et al. [159] developed parameters and benchmarked long-range-corrected DFTB2 (LC-DFTB2) and evaluated their performance on biologically relevant molecules. Their results demonstrated that LC-DFTB2 mitigates self-interaction errors and overpolarization effects inherent to standard DFT. Moreover, the method improves the description of charge-transfer excited states, while yielding geometries and vibrational frequencies comparable to those obtained with DFTB3/OB3.

2.5 Time Dependent Density Functional Tight Binding

DFTB, that is just described above, is a ground state method, that provide only information at the ground state. To be able to study spectral properties, therefore, we needed a method that provides information about excited state properties. We employed the time-dependent density functional tight-binding (TD-DFTB) method, developed analogously to the original TD-DFT approach [161, 162]. As in DFTB, the central quantity is the electron density; however, in TD-DFTB, this density is time-dependent rather than stationary. Moreover, it evolves within the framework of a time-dependent local Kohn–Sham potential, allowing the method to capture the dynamics of electronic excitations efficiently.

$$\rho(r,t) = \sum_{i=1}^{N} \phi_i(r,t) \tag{2.31}$$

A natural approach to solve a DFT/DFTB time dependent problem, would be by going for its numerical solution, where DFT/DFTB calculations are done at $t = 0$, and then propagated it in time to find the electronic excited state structure. By applying a Fourier transform to the time-dependent response, one can obtain the entire excitation spectrum in a single calculation [162]. However, this approach does not provide detailed information about individual electronic transitions, such as their oscillator strengths or specific contributions to the spectrum.

To address this limitation, the widely used Casida formalism provides an alternative framework [163, 164]. The approach described here represents an extension of the original time-dependent density functional tight-binding method incorporating long-range correction, commonly referred to as long-range corrected TD-DFTB (LC-TD-DFTB).

In this approach, the excitation energies and excited states are accessed through solving the following eigenvalue problem:

$$\begin{pmatrix} \mathrm{A} & \mathrm{B} \\ \mathrm{B} & \mathrm{A} \end{pmatrix} \begin{pmatrix} \mathrm{X} \\ \mathrm{Y} \end{pmatrix} = \Omega \begin{pmatrix} \mathbb{1} & 0 \\ 0 & -\mathbb{1} \end{pmatrix} \begin{pmatrix} \mathrm{X} \\ \mathrm{Y} \end{pmatrix} \tag{2.32}$$

where Ω is the excitation energy and X, Y provide access to the state properties. Using indices $\{i, j, ...\}$ to label occupied orbitals, and $\{a, b, ...\}$ for unoccupied orbitals, and σ and σ' spin indices, matrices A and B can be expressed as:

$$\mathrm{A}_{ia\sigma,jb\sigma'} = \frac{\delta_{ij}\delta_{ab}\delta_{\sigma\sigma'}\omega_{jb\sigma'}}{n_{j\sigma'} - n_{b\sigma}} + K_{ia\sigma,bj\sigma'} \tag{2.33}$$

$$\mathrm{B}_{ia\sigma,jb\sigma'} = K_{ia\sigma,bj\sigma'} \tag{2.34}$$

where $\omega_{jb\sigma'} = \epsilon b\sigma' - \epsilon_{j\sigma'}$ with $n_{i\sigma}n_{a\sigma}$ and $n_{j\sigma'}n_{b\sigma'}$, and K the coupling matrix. And the forces are obtained from the derivative of Ω with respect to the perturbation, s as follow:

$$\frac{d\Omega}{ds} = \sum_{\mu\nu\sigma} \left(\frac{\partial H^0_{\mu\nu}}{\partial s} P_{\mu\nu\sigma} - \frac{\partial S_{\mu\nu}}{\partial s} W_{\mu\nu\sigma} \right) \tag{2.35}$$

$$+ \sum_{\mu\nu\sigma\kappa\lambda\sigma'} \frac{\partial(\mu\nu|\nu_C + f^{XC,W}_{\sigma\sigma}|\kappa\lambda)}{\partial s} \tilde{\Gamma}_{\mu\nu\sigma,\kappa\lambda\sigma'} \tag{2.36}$$

$$+ \sum_{\mu\nu\sigma\kappa\lambda\sigma'} \frac{\partial(\mu\nu|\nu^{lr,\omega}_C|\kappa\lambda)}{\partial s} \tilde{\Gamma}_{\mu\nu\sigma,\kappa\lambda\sigma'} \tag{2.37}$$

where Ω is the excitation energy, H^0 the zero-order Hamiltonian, and S the overlap matrix. P, W and Γ are the relaxed one particle difference, the weighted energy and the two particle density

matrices respectively.

The TD-DFTB method has also been benchmarked against first-principles approaches to assess its accuracy and applicability for excited-state calculations across various systems. Using a benchmark set of 100 molecules, Sokolov et al. evaluated the accuracy of LC-TD-DFT potential energy surfaces by comparison with second-order approximate coupled cluster (CC2) calculations [165, 166]. A mean deviation of 0.31 eV was reported, indicating that LC-TD-DFT is an efficient and reliable method for obtaining excited-state geometries, optical properties of large biomolecules, and excited-state dynamics.

The applicability of (TD-)DFTB has further been demonstrated in QM/MM studies. Bold and co-workers applied LC-DFTB2 to chromophores in large light-harvesting complexes and rhodopsins, showing that the method is efficient and reliable for sampling absorption energies [167]. Consistent conclusions were reported by Miron and collaborators in their investigation of non-aromatic fluorescence in L-pyroglutamine–ammonium crystals [153]. Their study demonstrated that DFTB correctly captures the origin of fluorescence in this system, as well as the nonradiative decay pathways in non-fluorescent L-glutamine. Collectively, these studies, among many others, encourage the use of DFTB methods in combination with classical force fields within QM/MM frameworks, to enable simulations of realistic systems with improved sampling efficiency.

In light of its demonstrated accuracy and efficiency for excited-state properties and dynamics, this LC-TD-DFTB method is employed for the excited-state simulations in this work; for simplicity, it is referred to as TD-DFTB throughout the rest of the manuscript.

Chapter 3

The Origins of UV Absorption in the α_3C Protein

The work presented in this Chapter was carried out in collaboration with Gonzalo Diaz Miron, and has been published in Reference [168].

3.1 Introduction

Building on the foundations established in the earlier chapters, we next focus to exploring the origins of unusual UV-visible absorption properties of α_3C. Recently, Prasad et al. showed using UV-vis absorption spectroscopy that α_3C [1], which is devoid of any aromatic or conjugated groups, shows a broad absorption between 250-800 nm. Using excited-state quantum chemistry calculations of amino acids extracted from the protein and modeled in vacuum, they found that low energy transitions between 250-800 nm could be attributed to charge-transfer transitions between negatively charged carboxylate groups and backbone groups as well as positively charged lysines. These calculations were shown to quantitatively reproduce the experimental spectrum.

In this work we avoid choosing hydrogen-bonding patters a priori by employing an unsupervised learning approach to identify relevant structural motifs of the α_3C protein that subsequently serve as input for QM/MM simulations from which TD-DFTB simulations are performed. Our procedure automatically identifies three structural motifs involving the amide-peptide backbone, lysine and finally arginine amino acid all of which form thermodynamically stable clusters involving hydrogen bonding interactions with other protein chemical groups as well as water molecules. These three clusters are then used to perform QM/MM simulations and subsequently excited state calculations. Interestingly, we find that hydrogen bonding interactions involving arginine and carboxylate groups seem to be the key interactions leading to charge-transfer excitations extending from 250-350 nm. The extent of this red-edge absorption is highly sensitive to the local chemistry and solvation included into the QM region. For some of the systems studied, we also examine the role of nuclear quantum effects (NQEs).

The following sections elaborate on the computational details, results and discussions and

finally the conclusion of the investigation.

3.2 Computational Details

In this section, we describe the computational protocols employed to study the optical properties of the α_3C protein in aqueous solution. Our methodology involves several key steps, summarized in Figure 3.1.

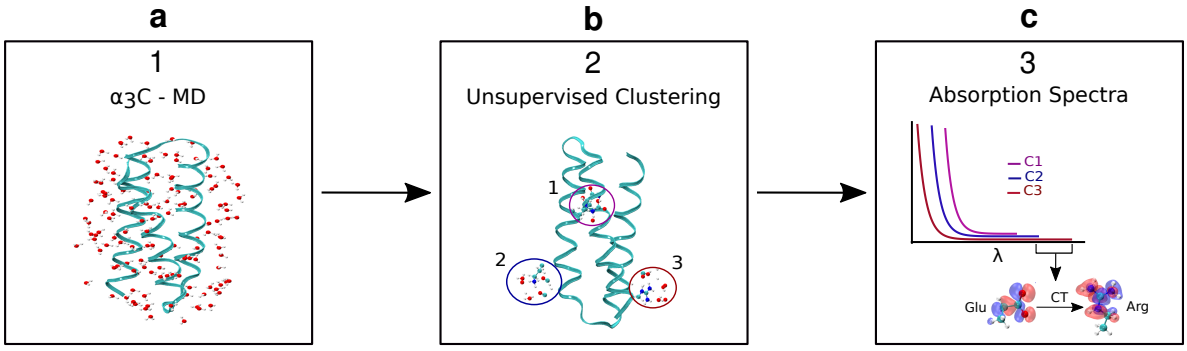


Figure 3.1: The three steps in our data-driven approach are summarized in the scheme. The first step (1) involves performing MD simulations of the α_3C protein. Configurations obtained from the MD are then used to build local-atomic descriptors followed by dimensionality reduction and subsequently clustering (2). Finally, the identified clusters are then used perform calculations using TD-DFTB to obtain absorption spectra.

Initially, we conducted classical molecular dynamics (MD) simulations of α_3C in water (1). The trajectory was then pipelined using machine learning techniques to automatically identify the relevant structural patterns (2). From these identified clusters, we performed Quantum Mechanics/Molecular Mechanics (QM/MM) simulations on the ground electronic state (3). Finally, we computed the absorption spectra for each relevant cluster, providing insights into the protein's optical properties. In the following, we will describe the specific details for each of these steps.

3.2.1 Classical Molecular Dynamics Simulations

We began by constructing the system using the conformation of the α_3C protein obtained through nuclear magnetic resonance (NMR) from the Protein Data Bank (PDB code 2LXY) [169]. The protein was then solvated in a rectangular box of water molecules with dimensions 6.7 × 5.6 × 6.0 nm^3, and two chloride ions were added to neutralize the system. For all classical simulations, we employed the GROMACS software suite [170], using the CHARMM27 force field [72] for the protein and the TIP3P model [70] for water molecules as used in a previous study [1]. Our simulation protocol began with 10000 steps of energy minimization to relax the system and remove any steric clashes or unfavorable contacts. This was followed by an initial 100 ps equilibration process involving a gradual rise in temperature from 0 K to 300 K using the NVT ensemble. Subsequently, a 200 ps equilibration at constant pressure and temperature (1 atm and 300 K) was performed using the NPT ensemble to equilibrate the density. The final production run consisted of a 1 µs simulation in the NPT ensemble, maintaining the same temperature and pressure. The V-rescale

thermostat [77] was used to control the temperature, while the Parrinello-Rahman barostat [78] was employed for pressure coupling. The protein maintained its three-helix bundle structure over the simulation period (see Figure A.1 in the Appendix). A subset of the MD trajectory that consists of configurations sampled at every 100 ps (a dataset with 10000 frames) was used for subsequent calculations.

3.2.2 Identification of Relevant Local Structures

Using configurations extracted from the MD simulation, we applied machine learning algorithms to identify the most relevant structural motifs that could serve as input for the optical calculations. This process involves three critical steps: firstly the use of a local atomic descriptor to characterize the local environments, secondly estimating the intrinsic dimension (ID) of the dataset and finally, a clustering technique to group similar environments together.

We used SOAP to describe local environments in our MD simulations chosen in this way: we put SOAP centers on all nitrogen atoms of the protein, and include in the environment, nitrogen and oxygen atoms regardless of whether they come from the protein or water molecule type. We employed a cutoff radius of 4 Å, chosen based on the radial distribution functions between nitrogen and neighboring oxygen atoms (see Figure A.3 in the Appendix). Our motivation for using this setup to build the chemical environments is dictated by the fact that the strong polar interactions in α_3C involve N-H hydrogen bond donors either along the protein backbone or coming from the side-chains. Using the oxygens of the protein to build SOAP centers led to the identification of very similar statistically significant clusters. We have found that increasing the cutoff radius does not change the outcome of our results (see Figure A.4 in the Appendix). The width of the Gaussian function was set to 0.25 Å, consistent with previous studies from our group [85]. The maximum number of radial (n_{max}) and angular basis (l_{max}) functions was set to 8 and 6, respectively. Using the DScribe software package [171], the SOAP power spectra were computed for 90×10000 local environments (90 SOAP centers each of the 10000 frames).

With the SOAP features in hand we employed the Two-NN estimator [116] to estimate its ID, we then employed the Point Adaptive k-nearest neighbor estimator (PAk) to construct the high-dimensional free energies. The combination of these unsupervised techniques has been successfully used in our group for several other applications involving liquid water [79, 106] and concentrated acids [105]. In our analysis, we randomly sampled 100000 data points from the SOAP dataset and performed DPA clustering using $Z = 14$. Note that the Z-parameter in DPA determines the statistical confidence of the cluster. The sensitivity of our results to the choice of this parameter are discussed in the Appendix, see Figure A.4. We use distance-based analysis of data-manifolds in Python (DADApy)[172], a Python software library, for all the analysis. For visualization purposes, we employed the Uniform Manifold Approximation and Projection (UMAP) method [113], to project the high-dimensional dataset into a 2D space. Additionally, we also performed the same analysis using the HDBSCAN clustering method [173] for comparison and validation obtaining consistent results.

3.2.3 Ground State QM/MM Dynamics

Following the identification of the different clusters in our system, 10 distinct configurations were randomly selected from each cluster to run the QM/MM ground state electronic simulations. Before delving into the details of these simulations, we will describe our protocol for partitioning the system into the QM and MM regions.

The output provided by our clustering technique identifies various environments around the hydrogen bonds involving the nitrogen atom and its bound proton. Therefore for the QM region, we included all atoms that are approximately within a radius of 4 Å from the center of mass of these hydrogen bonds. In some cases, if we identified a charged amino acid at the boundary, it was also included into the QM region. Each QM region comprises of up a maximum of 100 atoms including both solvent and protein atoms. Hydrogen atoms were added as needed at the boundaries of the QM/MM to serve as link atoms [131, 174]. The remainder of the system was assigned to the MM region. To assess the sensitivity of our analysis to the role of including QM waters, we also repeated some of our simulations removing all the waters from the QM region (see Results and Discussion for details).

The QM/MM simulations were run using the electrostatic embedding scheme [131], in which the electrostatic potential of the MM region influences the QM region, providing a more accurate representation of the interactions between the QM and MM regions. This approach ensures that the electronic structure of the QM region is properly polarized by the surrounding MM environment. All the simulations were performed using the GENESIS software [175, 176], which handles the propagation of nuclei and all MM calculations. The force field employed in this simulations are CHARMM36 [177] and TIP3P water model [70]. Each simulation was run for 5 ps with a timestep of 0.5 fs, using the canonical-sampling velocity-scaling thermostat [178] to maintain a constant temperature at 300 K. For the QM calculations, we employed the Density Functional Tight Binding (DFTB) theory as implemented in the DFTB+ package [179], utilizing third-order corrections with the 3ob Slater-Koster parameters [152]. For more methodological details, the reader is referred to previous literature on the topic [133, 180].

3.2.4 Absorption Spectra Calculations

For each of the previously run QM/MM ground state simulations, we selected 100 snapshots at 50 fs intervals and performed excited state calculations on these configurations. We used two different setups for these calculations: the QM/MM absorption spectra, where the MM region was included as point charges, and the QM vacuum absorption spectra, where all MM atoms were excluded from the calculation.

We utilized the DFTB+ software for these calculations, employing Tight-Binding Time-Dependent Density Functional Theory (TD-DFTB) [162] with Long-Range corrections using the ob2 Slater-Koster parameters [159]. For each conformation, a total of 30 excited states were calculated. When computing and displaying the spectra, each electronic transition was broadened using a Gaussian

function with a width of 1 nm.

3.2.5 Path-Integral Simulations

In order to assess the role of nuclear quantum effects (NQEs) on the absorption spectra, we conducted path-integral molecular dynamics simulations on the electronic ground-state for a selected cluster (see Results for more details). We selected a QM system from our protein and ran a simulation with classical nuclei in a vacuum using DFTB+ software employing the same parameters as previously described. Path integral simulations for the same cluster in vacuum were simulated using DFTB+[179] and i-Pi [181, 182] software, with the PIGLET thermostat [183] and four beads. In a previous work we have shown that the effect on the absorption spectra comparing classical and quantum simulations is qualitatively captured using both 2 and 6 beads [184]. Both classical and quantum simulations were performed at a constant temperature of 300 K for 30 ps with a timestep of 0.5 fs. After completing the ground state simulations, we selected 100 equidistant frames and determined the absorption spectra using TD-DFTB [162] using the same conditions as described earlier. All 4 beads of the path integral were used to compute the absorption spectra and therefore a total of 400 frames were used.

3.3 Results and Discussion

In this section, we begin by first presenting the results that emerge from our unsupervised clustering of the local environments in the protein and their subsequent role on determining the absorption spectra.

3.3.1 Hydrogen-Bond Network Motifs

Our unsupervised clustering procedure in SOAP space yields three statistically dominant clusters. The left panel of Figure 3.2 displays the UMAP projections colored based on the three dominant clusters obtained using the DPA method for a SOAP cut-off radii of 4 Å. The first cluster encompasses approximately 77% of the regions, while the second and third clusters contain 16% and 7%, respectively. Less than 0.2% of the total data-points are unclassified and are illustrated in the projection as black points. We have found that the existence of these three clusters is preserved for radial cutoffs that extend to larger values of 8 Å as well as changing the Z parameter (see Figure A.4 in the Appendix). Overall, the three clusters consistently occupy over 99% of the population. The right panel of Figure 3.2 shows the outcome using another clustering method namely, HDBSCAN which yields a total of 5 clusters. However, three of the clusters essentially occupy 83% of the population similar to DPA.

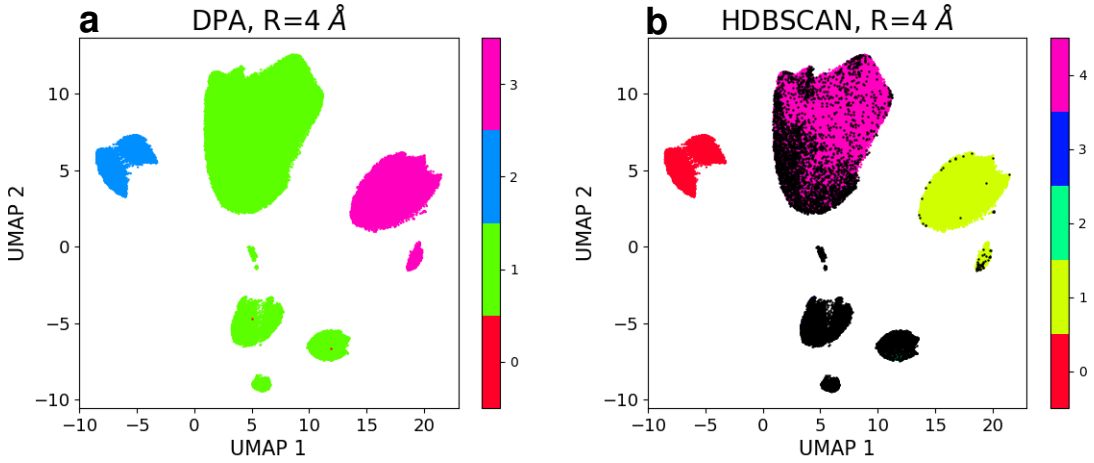


Figure 3.2: UMAP projections of the dominant clusters obtained using the DPA method (left) and HDBSCAN (right). The color codes correspond to the different clusters (4 in the case of DPA and 5 in the case of HDBSCAN).

Having identified three statistically dominant clusters in the protein, we turned next to interpreting their chemical origin. As already eluded to earlier, since α_3C consists of three helices, one might expect to see strong hydrogen bonds along the helix. Indeed, the largest cluster referred to in the rest of the manuscript as C1, consists of regions centered on nitrogen atoms of the protein backbone. These nitrogen atoms form intra-chain hydrogen bonds with oxygen atoms on the protein backbone (see leftmost panel of Figure 3.3) and occasionally also with water oxygens.

Cluster-2 (C2) the second largest cluster, is primarily composed of regions centered on nitrogen atoms on the side chains of lysine residues (middle panel of Figure 3.3). A small but non-insignificant proportion of this cluster consist of nitrogen atoms at the N-terminus of the protein, particularly on glycine, the first residue. These nitrogen atoms are highly exposed to the solvent and form hydrogen bonds with the side chains of glutamic acid residues and water molecules. Finally cluster-3 (C3) is dominated by the side-chains of arginine (rightmost panel of Figure 3.3). These nitrogen atoms are also exposed to the solvent forming hydrogen bonds with the side chains of glutamic acid and water. The chemical origin of the three dominant clusters obtained with HDBSCAN are fully consistent with these preceding findings.

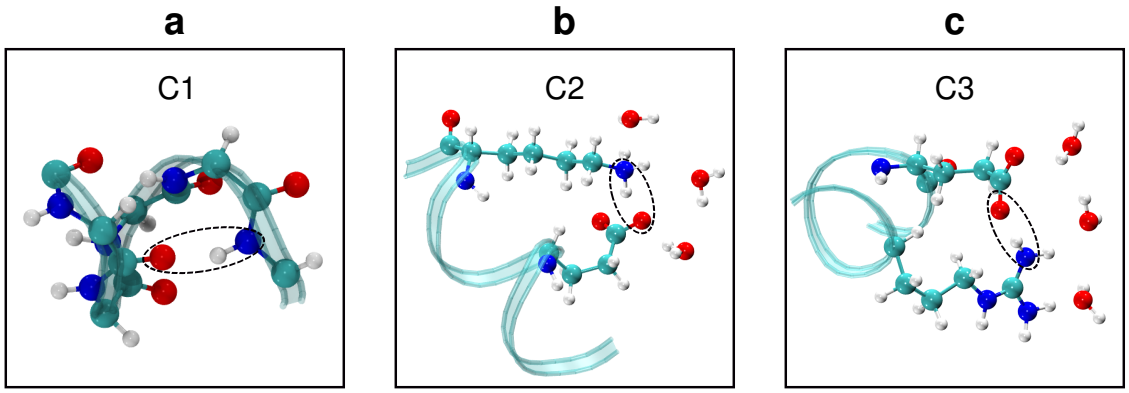


Figure 3.3: Schematic snapshots of the three clusters obtained from our analysis. Panel a) shows the backbone of the protein which primarily involves hydrogen-bonding of the amide-bonds along the helix (cluster-1, C1) . Panel b) shows lysine and glutamic acid hydrogen bonded to each other including some water molecules (cluster-2, C2). Panel c) shows arginine and glutamic acid side chains hydrogen bonded to each other also including some solvation water (cluster-3, C3).

3.3.2 Absorption Spectra

As eluded to earlier in the Introduction, Prasad and co-workers recently examined using UV absorption spectroscopy the optical properties of the α_3C protein in solution where a broad UV absorption between 250-800 nm was identified [1]. In order to interpret the physical origins of this long-tail absorption, they conducted time-dependent density functional theory (TD-DFT) calculations in vacuum using clusters extracted from the protein. Focusing specifically on hydrogen-bonding interactions involving the charged amino (NH_3^+) and carboxylate (COO^-) groups coming from the lysine and glutamic-acid side chains, they found charge-transfer excitations between these moieties that ranged between approximately 300-800 nm. The magnitude and intensity of these transitions was found to be rather sensitive to specific geometrical and environmental effects. For example, they found that pairs of charged amino acids that were positioned further away from each other ($\sim$5-6Å) were more likely to lead to lower energy excitations.

The preceding results were conducted using TD-DFT with the range-separated functional CAM-B3LYP [185]. While this in principle provides a more accurate treatment of the electronic structure, it is computationally expensive and therefore limited to small system sizes. Our approach of using TD-DFTB overcomes these challenges due to its more computationally tractable semi-empirical nature. Figure 3.4 shows the absorption spectra of the systems selected with our unsupervised algorithm described in the previous section in vacuum.

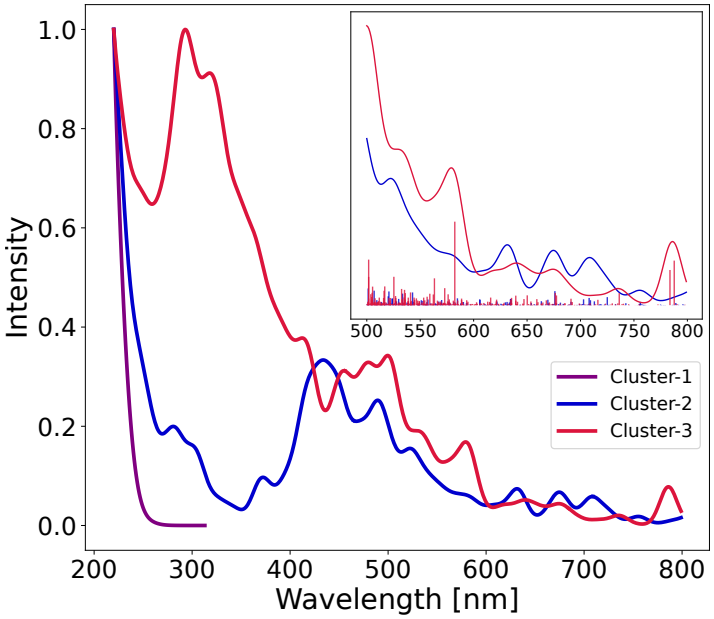


Figure 3.4: Absorption spectra of conformations sampled from the three clusters calculated using QM clusters devoid quantum water molecules in vacuum. The inset shows the spectra in range 500 - 800 nm, and corresponding excitations.

As shown in Figure 3.4, C1 does not exhibit any transitions beyond 300 nm, consistent with what one would expect from transitions involving only the protein backbone. In contrast, C2, which is primarily associated with the charged amino acids Lys and Glu, displays transitions within the 300–800 nm region. This behavior aligns with findings from the previous work of Prasad [1]. However, an important distinction is that our systems were selected using an agnostic approach. Notably, our approach also identifies a distinct cluster, which primarily involves hydrogen-bonding (HB) interactions with the charged amino acid Arg. Our calculations further revealed that this cluster exhibits absorption within the 300–800 nm range. It is important to emphasize that our unsupervised approach is designed to identify structural fingerprints linked to statistically significant hydrogen-bonding patterns, rather than those motifs inherently optically active. Nevertheless, our excited-state calculations demonstrated that certain HB patterns exhibit absorption in the UV-Vis region.

Having demonstrated that our protocol using the clusters emerging from our unsupervised protocol with TD-DFTB produces similar results to those reported by Prasad et. al. [1], we are now in a position to move to applying it to the full protein. With the three clusters in hand we examined how these different motifs modulate the absorption spectra in a more realistic environment. We thus extracted the spectra conducting QM/MM simulations with DFTB on the electronic ground state and then determining the excited states of the system where short-range interactions with nearby protein and water moieties were included in the QM region, while the long-range interactions were treated by explicitly including the protein and water groups with classical electrostatics.

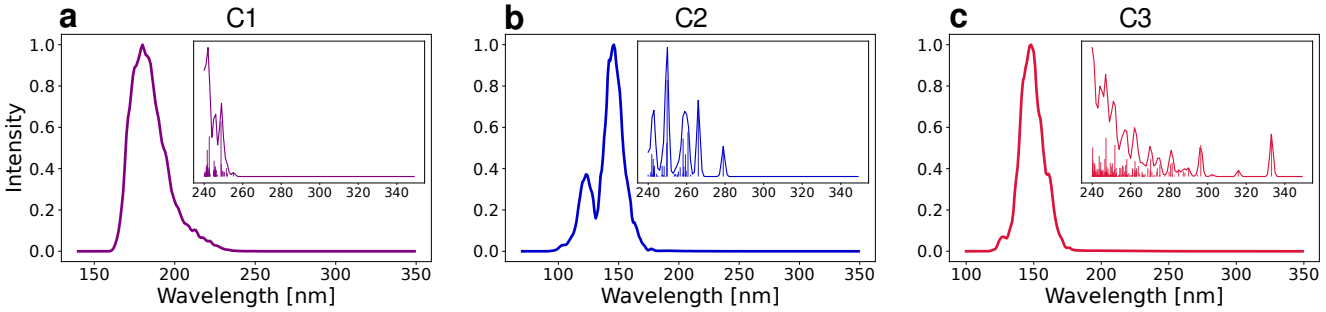


Figure 3.5: Absorption spectra obtained from the three clusters with inset showing a zoom-in of the transitions above 200 nm. Out of the three clusters, C3 is the only one that displays transitions between 300-350nm.

Figure 3.5 compares the absorption spectra obtained for the three clusters averaging over a total of 1000 independent snapshots. At first glance, Figure 3.5 appears to show that the spectra from all three clusters are dominated by a high intensity transition below 200 nm. In particular, C1 is peaked at 180 nm, C2 is characterized by a bi-modal structure involving a peak at 123 and 146 nm and finally, C3 has a single peak at 148 nm. However, upon closer examination C1 and C2 feature a tail in the absorption extending up to 280 nm as seen in the inset plots. It thus appears that the peptide backbone interactions and those involving lysine forming hydrogen bonds with the carboxylate when embedded in a realistic environment, no longer feature low-energy excitations below 300 nm.

In contrast, C3 appears to be the only case where one observes a tail beyond 300 nm with a few some transitions near 340 nm. These transitions albeit weak, indicate that the hydrogen bonds involving arginine appear to be the only ones that introduce electronic states that appear in the mid UVA (315-400nm) region. Thus the inclusion of both the water and protein environment through a QM/MM setup does not lead to a long-tail of UV-absorption as seen in the vacuum simulations. To ensure that this behavior is not an artifact of our TD-DFTB approach, we repeated the calculations for vacuum and QM/MM systems for cluster-2 using the TD-DFT/CAM-B3LYP method, as reported by Prasad [1]. As shown in Figure S4 in the supplementary information, the observed blue shift when transitioning from vacuum to the inclusion of the environment is also present in the TD-DFT calculations.

In order to understand the electronic origins of the low energy tails of the spectrum for each cluster, we examined the molecular orbitals (MO) associated with the lowest energy excitations as shown in the two panels of Figure 3.6. The electronic transitions in C1 correspond to a charge transfer n to π^* excitation from an amide-backbone group of one amino acid to another. In C2 instead, the low energy transitions correspond to a localized n to π^* excitation on the carbonyl group of the glutamic acid side-chain. Finally, the transitions higher than 300 nm found in the tail of C3 involves a HOMO-LUMO n to π^* transition from the carboxylate group of glutamic acid to the guanidinium side-chain of arginine.

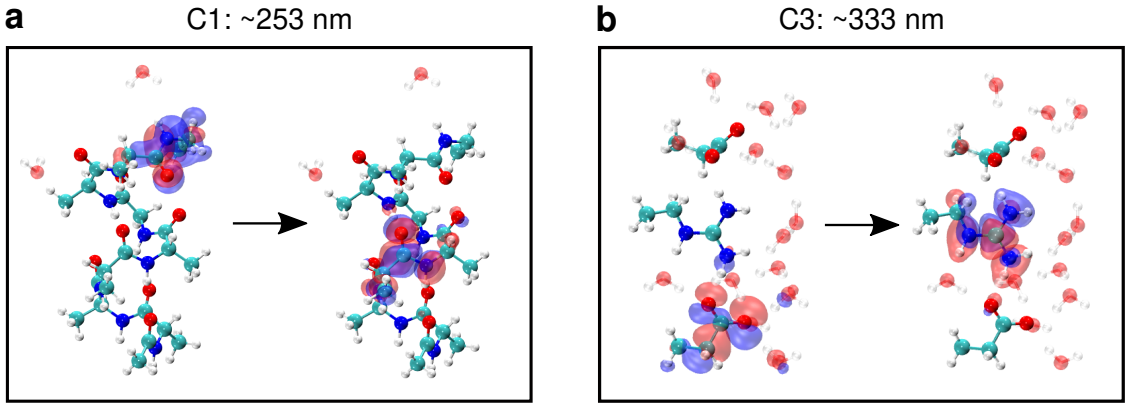


Figure 3.6: Low energy transitions in the first and the third cluster. The left panel a) shows a charge transfer from one amide group to another in C1, and the right panel b) shows a charge transfer from a glutamic acid carboxylate group to a guanidinium group of arginine in C3.

The preceding results show that absorption spectra for all three clusters do not present any excitations above 350 nm. This observation differs from the experiments which display a long tail of absorption extending from 400-800 nm (see Figure 3.7(a)). Over the years, there have been several studies showing the importance of solvation, specifically, the inclusion of explicit water molecules[184, 186, 187, 188, 189] in tuning the optical absorption of organic molecules. Since our studies here involve a full QM/MM framework, we are in a unique position to examine the relative role of QM vs MM waters/protein in tuning the optical absorption.

Panel (b) of Figure 3.7 shows the absorption spectra obtained from our normal QM/MM setup described earlier, to two other situations: i) the first involves carving out the QM region and converting it into a cluster in vacuum with the appropriate capping using hydrogen atoms which can in principle also contain QM water molecules (solid green curve) and ii) same as i) but now without including any waters into the QM cluster (solid red curve). Moving from the full QM/MM to scenario ii) results in a significant red-shift of 200 nm and furthermore introduces a long-tail in the absorption between 500-800 nm. This feature puts the theoretical predictions into closer agreement with the experiments with the caveat of course, that it does not include a realistic description of both the protein and water environment. Upon inclusion of water molecules into the QM region of the cluster, there is a blue-shift of approximately 100 nm moving the first excitations higher up in energy. A similar blue-shift was also observed previously by Prasad and co-workers [1]. On the other hand, if we take our original QM/MM systems and change all the QM waters into MM (therefore keeping only the protein atoms in the QM region) there is a small blue shift in the spectrum of about 20 nm as shown in Figure 3.7(c).

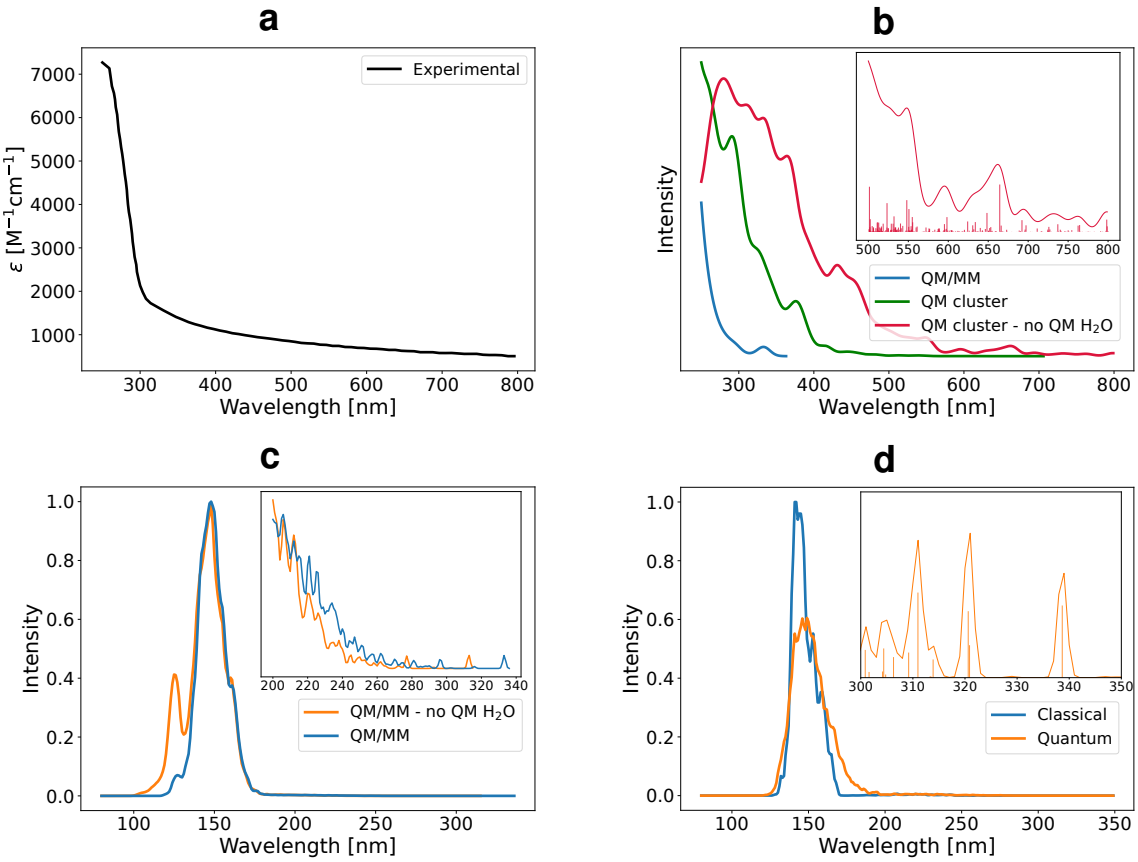


Figure 3.7: Panel (a) show the experimental absorption spectrum of α_3C ranging from 250 nm, with the tail extending to 800 nm, adapted from [1]. Panel (b) shows cluster-3 simulated absorption spectra, smoothened with a Gaussian of 10 nm width, calculated in three ways: 1) using our normal QM/MM setup described in the methods (solid blue curve), 2) using a QM cluster in vacuum (solid green curve) and 3) on QM cluster devoid quantum water molecules (solid red curve). The inset shows absorption spectrum for case 3) and corresponding excitations. Panel c) compares the absorption spectra from the full QM/MM protocol with and without the inclusion of water molecules in the QM part. Finally, panel d) shows the absorption spectra comparing classical and quantum simulations that show the role of nuclear quantum effects.

All in all, our findings show an important and critical challenge of comparing absorption spectra from experiments with theory. On the one hand, it is possible to construct a model that almost quantitatively reproduces the experimental trends, in this case a QM cluster in vacuum. Building systems that in principle account for environmental effects such as the protein and water, introduces bigger discrepancies between the theoretical predictions and experiments. It should however be stressed that our QM/MM approach appears to correctly capture an important region of the spectrum between 300-350nm that does not arise from aromatic groups.

In the last decade, there has been a growing body of theoretical work showing the importance of including nuclear quantum effects (NQEs) into molecular simulations [190]. Specifically, zero-point

energy (ZPE) fluctuations have been shown to affect structural [191, 192], dynamical [192, 193] and electronic [184, 189, 194, 195, 196, 197, 198] properties of hydrogen-bonded systems. Due to the larger structural distortions induced by enhanced fluctuations of vibrational coordinates, absorption spectra can display substantial red-shifts in energy introducing lower-energy excitations that are completely absent in classical simulations.

In order to assess the sensitivity of our results to the role of NQEs we conducted path-integral simulations coupled with a colored noise thermostat (PIGLET) focusing on trying to understand how NQEs would shift energies in cluster-3 which displayed the lowest energy transition. Due to the computational cost, we could only conduct this analysis on the QM cluster in vacuum with a few solvating water molecules. Figure 3.7(d) compares the absorption spectra obtained from the classical runs to PIGLET. In the latter, the lowest energy excitation only reaches 260 nm while in the case of the PIGLET runs this extends to 340 nm. Thus NQEs lead to a red-shift in the spectra of approximately 80 nm (around 1.1 eV). This however still remains blue shifted from the experimental values by at least 200 nm (around 1.4 eV).

3.4 Conclusions and Perspectives

There is currently a very active area of research from both experimental and theoretical fronts looking into the spectroscopic origins of biological systems that display a tendency to absorb UV radiation above 300 nm. A better fundamental understanding of this phenomena holds the promise to allow for designing non-invasive probes for biophysical processes. In this work, we have focused our efforts on studying the optical properties of α_3C a synthetic protein which was recently shown to display near UV-visible absorption creating a long-tail of excitations between 300-800 nm. These previous studies conducted computational work on gas-phase clusters which were shown to reproduce the experimental observations.

Here instead, we employ a data-driven approach where state-of-the-art unsupervised learning approaches are used to automatically discover statistically important structural motifs in the protein. These are then used to conduct QM/MM simulations and optical absorption calculations, where the environment involving both protein and water molecules are included. This is achieved through the use of a tight-binding approach which we have recently demonstrated to give an excellent compromise between computational efficiency and accuracy for studying electronic properties of hydrogen-bond networks.

Our results leads to a somewhat surprising conclusion namely that a realistic inclusion of both the protein and water environment eliminates the long-tail of optical absorption between 400-800 nm (3.1-1.55 eV). Instead, our approach correctly captures features between 300-350 nm which arise from charge-transfer excitations between arginine and carboxylate groups. We propose, by conducting proof-of-concept path-integral simulations that one can expect an additional red-shift of up to 1 eV when quantum effects are included. This could lead to a red-tail absorption extending to approximately 450 nm in the case of cluster-3. Clearly more work is needed to better understand the origins of the discrepancies. On the experimental side, it would be interesting to examine how

the absorption spectra change in α_3C using site-mutagenesis that target different amino acids such as some specific lysine or arginine groups.

Chapter 4

On the Mechanism of Fluorescence in the α_3C Protein

4.1 Introduction

In the previous chapter, we explored the absorption properties of the α_3C protein, which have been reported to span the range of 250-800 nm. Using a combination of unsupervised learning techniques and absorption calculations within a QM/MM framework, we identified three potential chromophoric interactions responsible for absorption in the 250-350 nm region: backbone interactions, amino-carboxylate (lysine-glutamic acid) interactions, and guanidinium-carboxylate (arginine-glutamic acid) interactions.

With these results and tools in hand, we revisit the work of Kumar and collaborators [2] where they demonstrate that excitation in proteins within the 295-305 nm range is not solely due to tryptophan, but can also arise from other, unlabeled chromophores that emit weakly in this region. Through a comparative analysis of the emission spectra of N-Acetyl-L-tryptophanamide (NATA) and the α_3W mutant (a tryptophan-containing variant of α_3C) under 295 nm excitation, they observed that α_3C exhibited weak emission between 310 and 550 nm.

The absorption spectrum of α_3C reported in the 250-800 nm range [1] is particularly intriguing. The observation of weak emission from this protein further suggests that its chromophores are not merely absorbing species but are also capable of radiative decay, implying the existence of relatively stable excited states. Although our calculations did not predict any electronic transitions beyond 350 nm for α_3C, we did observe unconventional absorption features in the 250-350 nm region. Notably, the experimental excitation at 295 nm falls within this predicted range, lending confidence to the relevance of our simulations and enabling meaningful comparison with experimental findings.

In this chapter, we focus on exploring the excited-state dynamics of α_3C, including the key vibrational relaxation pathways responsible for nonradiative decay, as well as the vibrational modes that may stabilize the excited state long enough to produce the observed weak emission. In addition, we aim to identify which of the proposed chromophores contribute to this emission and to estimate their relative contributions.

These investigations require excited-state computations, which are significantly more computationally demanding than the ground-state simulations discussed in the previous chapter. This work is ongoing, and the results presented here are preliminary, pending further simulations and increased statistical sampling.

4.2 Computational Methods

Ten independent configurations were selected from cluster-1, twenty-one from cluster-2, and sixteen from cluster-3. For each configuration, the QM region was defined and QM/MM molecular dynamics simulations were performed for 5 ps in the electronic ground state, following the protocol described in Chapter 3. The final snapshot of each equilibrated ground-state trajectory was then vertically excited, and subsequent QM/MM dynamics were carried out on the lowest singlet excited state (S_1) for 1 ps, again following the procedure outlined in Chapter 3. For each snapshot, a total of ten electronic states were computed. Nonadiabatic couplings associated with the breakdown of the Born–Oppenheimer approximation were not treated explicitly.

It is worth noting that although a simulation time of 1 ps is much shorter than the fluorescence timescale, which typically lies in the nanosecond regime, it is sufficient to capture ultrafast nonradiative decay processes. This timescale therefore enables the identification of primary nonradiative decay pathways and the vibrational modes that contribute to excited-state stabilization, which is the main objective of this study.

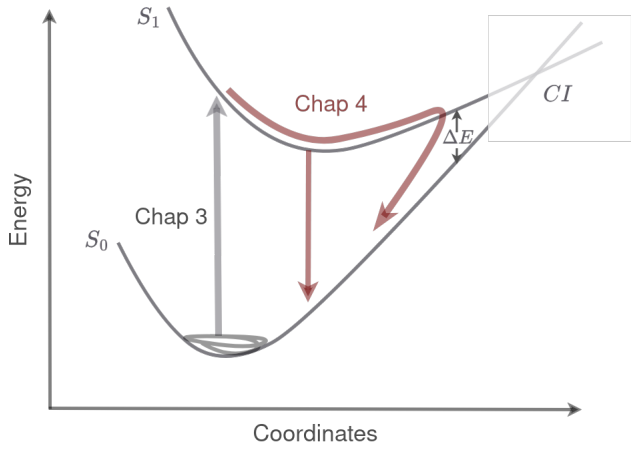


Figure 4.1: Illustration of electronic structure calculations done in Chapter 3 and those covered in this Chapter 4.

An excited-state trajectory was considered to approach a conical intersection sufficient to enable nonradiative decay, when the energy gap between the electronic ground state and the first excited state (denoted as ΔE in Figure 4.1) decreased to 0.2 eV; varying this threshold to values

below 0.5 eV yielded consistent results. Structural changes along the excited-state trajectories were analyzed to extract vibrational modes associated with nonradiative decay within each cluster. For trajectories exhibiting decay, only the part of the trajectory preceding the decay threshold was retained for subsequent analysis. In contrast, from trajectories that did not undergo decay, emission spectra were predicted from the lowest-energy non-decayed electronic transitions sampled after the first 200 fs and subsequently broadened using a Gaussian function with a width of 3 nm.

Subsequent analyses of the secondary structures involved Ramachandran plots [199, 200], which provides a convenient way to visualize energetically allowed geometries of a given amino acid on a polypeptide chain using Φ (C-N-Cα-C) and Ψ (N-Cα-C-N) dihedral angles. Plotting the two kinds of dihedrals from a polypeptide/protein against each other, reveal distinct regions corresponding to two common secondary structures: α-helices and β-sheets, with the former in upper region of the third quadrant (of the [-180°,180°] vs [-180°,180°] graph) where both dihedrals are negative. It also highlights forbidden regions which arise due to steric clashes between atoms.

4.3 Results and Discussion

Here we explore the types of excitations that occur in the three clusters, relative stability of excited state structures, the vibrational mode via which they decay, and finally, the calculated emission spectra.

4.3.1 Initial Excitations

As discussed in the previous chapter, the lowest-energy electronic transitions $S_0 \rightarrow S_1$ observed across all three clusters correspond to n $\rightarrow \pi^*$ excitations, each originating from distinct regions of the protein.

In cluster-1, which involves structural regions centered on interchain hydrogen bonds along the protein backbone, the transitions are characterized by charge transfer between consecutive turns of the helical backbone. In this case, the n molecular orbital is localized on one chain, while the corresponding π^* orbital resides on the adjacent chain.

Cluster-2 consists of side-chain structures centered on hydrogen bonding between the carboxylate group of glutamic acid and the amino group of lysine. Here, the excitations are localized within the carboxylate group, with both the n and π^* molecular orbitals residing on the same group.

Finally in cluster-3, which involves side-chain interactions between the carboxylate group of glutamic acid and the guanidinium group of arginine, the transitions involve charge transfer from the carboxylate group to the guanidinium group.

These represent the lowest-energy transitions activated upon excitation of the final equilibrated

frame in the electronic ground state. The corresponding lowest excitation energies for clusters 1, 2, and 3 were approximately 250 nm, 224 nm, and 266 nm, respectively.

4.3.2 Energy Gap Time Evolution

After the 1 ps dynamics of the configurations in the lowest excited state, our analysis began by examining the temporal evolution of the energy gap across all three clusters. We observed that, within each cluster, some trajectories exhibited clear decay behaviors while others did not. In cluster-1, among the ten simulated trajectories, three showed a clear decay, while seven did not. Of these seven, three trajectories approached the decay threshold but did not cross it, whereas four showed no decay at all. Accordingly, we categorized the trajectories into three groups: strictly decayed, almost decayed, and strictly non-decayed. In cluster-2, out of twenty-one trajectories, twelve were strictly decayed and nine were strictly non-decayed. In cluster-3, only one out of sixteen trajectories did not exhibit decay.

Figure 4.2 illustrates the time series of the energy gap for these trajectories. Strictly decayed trajectories are shown in orange, the mean and standard deviation of the almost decayed trajectories are represented in grey, and the mean and standard deviation of the strictly non-decayed trajectories are shown in blue for each cluster.

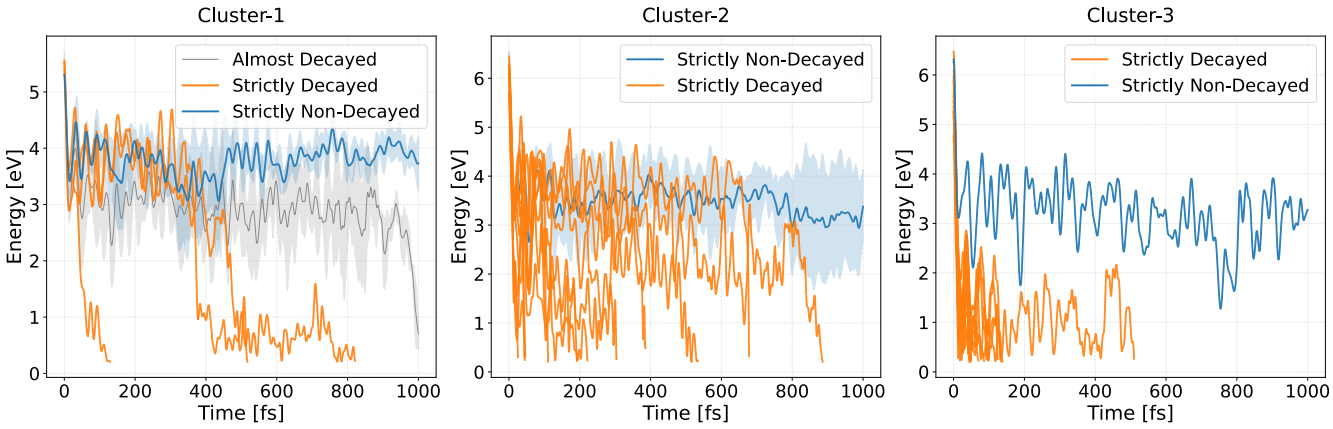


Figure 4.2: Energy gap time evolution in cluster-1 (on left panel), cluster-2 (middle) and cluster-3 (right panel). Orange curve are for individual strictly decayed trajectories, while blue curves and shaded region are for average and standard deviation of strictly not decayed trajectories. The grey curve and shaded region in cluster-1 is for the mean and standard deviation of energy gap in almost decayed trajectories.

Recognizing that 40% of the structures in cluster-1 and 43% in cluster-2 remain stable in excited state throughout the simulation, whereas only 6% in cluster-3 do not decay, we infer that cluster-3 appears to contribute the least to emission of the α_3C protein. Moreover, since over 90% of the trajectories in this cluster decay within the first 200 fs of the dynamics, it likely accesses the conical intersection more rapidly than the other clusters. Together, these observations indicate

that cluster-3 plays a minimal role in the emission of α_3C. It is noteworthy that across all the three cases, more trajectories decay than those which do not. This is in agreement with the experimental emission which was found to be weak.

Following this observation, we proceeded to investigate the underlying structural factors facilitating the decay in these trajectories, and the estimated emission from the non-decayed ones in each cluster to see in which region each contributes.

4.3.3 Cluster-1 Vibrational Relaxation Modes and Estimated Emission

We systematically explore the vibrational modes of structures native to each cluster starting with cluster-1 which comprises of structures located along the backbone. We identified the Cα-N (alpha carbon - nitrogen) bond stretching as the most prominent vibrational mode that correlated with the energy gap. In all the cluster-1 decayed trajectories, the energy gap consistently decreases with the Cα-N bond stretching. This correlation is illustrated in Figure 4.3 which shows the energy gap versus Cα-N bond length.

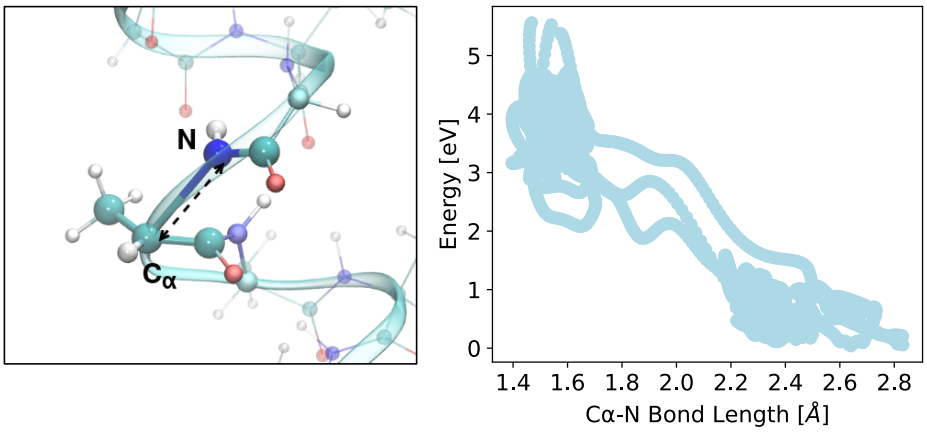


Figure 4.3: Illustration of the stretched Cα-N bond on the protein backbone (left panel), and the correlation between energy gap and the Cα-N bond length in cluster-1 strictly decayed trajectories.

Looking beyond the Cα-N bond stretching, which represents the most prominent mode correlated with energy gap, several additional structural parameters also exhibit notable correlations, including the C-O bond stretching, and the C-N-Cα angle. Therefore we examined the the secondary structure itself as a whole, with the help of Ramachandran plots [199, 200].

Figure 4.4 presents the two-dimensional probability density of the Ψ/Φ backbone dihedral angles in the Ramachandran plots, for the three categories of trajectories: strictly decayed (left panel, shown in orange), almost decayed (middle, shown in grey), and strictly non-decayed (right, shown in blue). The plots show that in all the three cases, the Ψ/Φ values are more dense in the α-helices

region as expected. However, the helices on left and middle explore farther Ψ/Φ values from those of the ideal α-helix, which show that the α-helical structure is distorted in those cases. The helical structure of the strictly not decayed remain relatively stable.

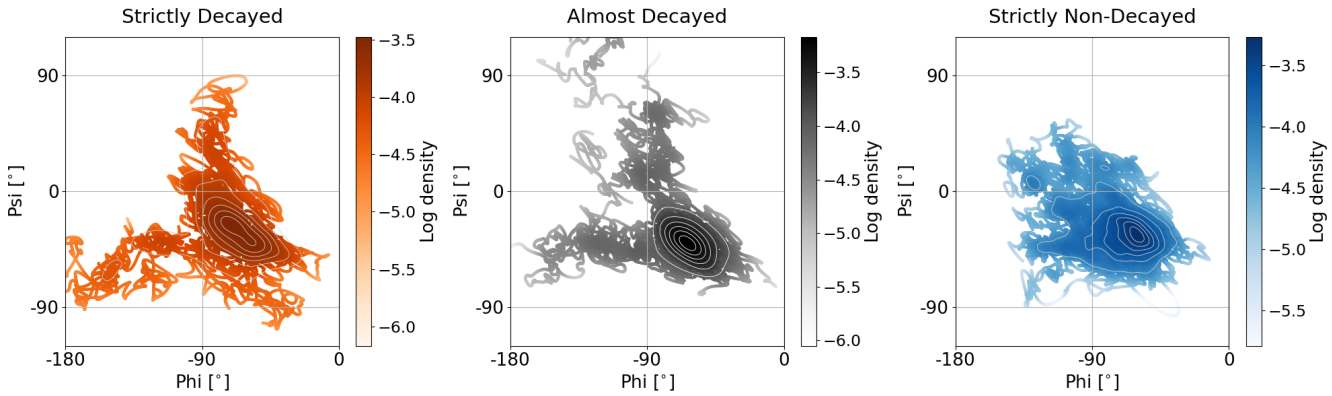


Figure 4.4: Two-dimensional probability density of Ψ/Φ dihedral angles in the Ramachandran plots for the three trajectory categories within cluster-1. The strictly decayed in orange (left panel), the almost decayed in grey (middle panel), and the strictly non-decayed group in blue (right panel).

Having observed the distortion of the helical structure, we next investigated whether it arose as a result of excitation or was already present in the ground state. We assess this by comparing ground states and excited state Ramachandran plots for the decayed trajectories. These plots are shown in Figure 4.7, where the excited state are in orange (left panel) and the corresponding ground state in blue (right panel). The results indicate that, in the ground state, structures occupy predominantly the α-helical region, confirming that the excited state conformations were initially close to pure helical prior to excitation.

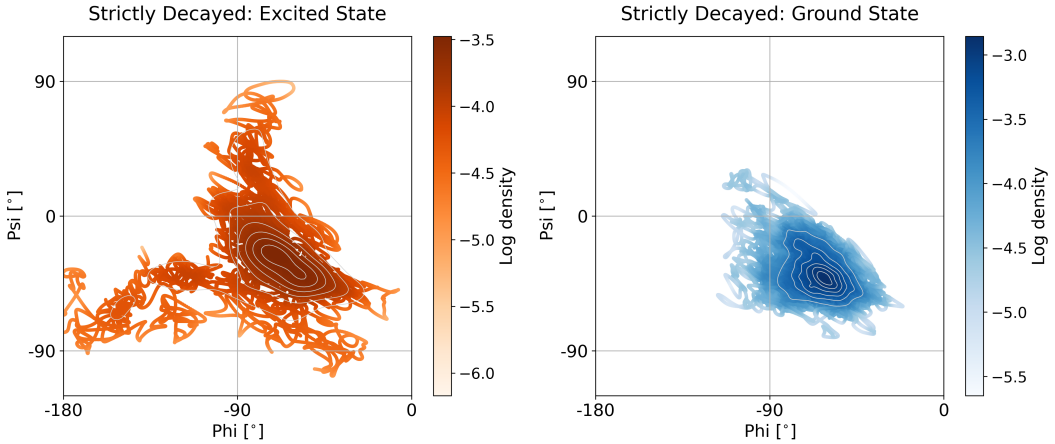


Figure 4.5: Two-dimensional probability density of Ψ/Φ dihedral angles in the Ramachandran plots for the strictly decayed cluster-1 excited state trajectories (left panel) and corresponding ground state categories (right panel).

To investigate how localized or delocalized the helical distortions are, and to determine when the distortions begin, relative to the decay time, we further analyzed the 2D probability densities of the Ψ/Φ dihedrals in the Ramachandran plots for the decayed trajectories. Specifically, we compared two different trajectories slices: one corresponding to configurations with an energy gap greater than 2 eV and the other with an energy gap below 2 eV. For each structure or trajectory, the analysis focused on the most affected helical turn, identified as the one with the largest Cα-N bond stretching. The results are shown in Figure 4.6 below.

The left panel presents the data over the entire simulation period up to decay, while the middle and right panels correspond to the first and second trajectory slices, respectively. Comparison between the left panels of Figures 4.4 and 4.6 indicates that the distortions are not confined to a single helical turn. The distinct Ψ/Φ distributions suggest that some combinations present in Figure 4.4 arise from helical turns not analyzed in Figure 4.6. Furthermore, the results shown in the middle and right panels of Figure 4.6 reveal that distortions in the helical conformations appear well before the onset of decay, indicating a gradual loss of structural integrity rather than an abrupt transition.

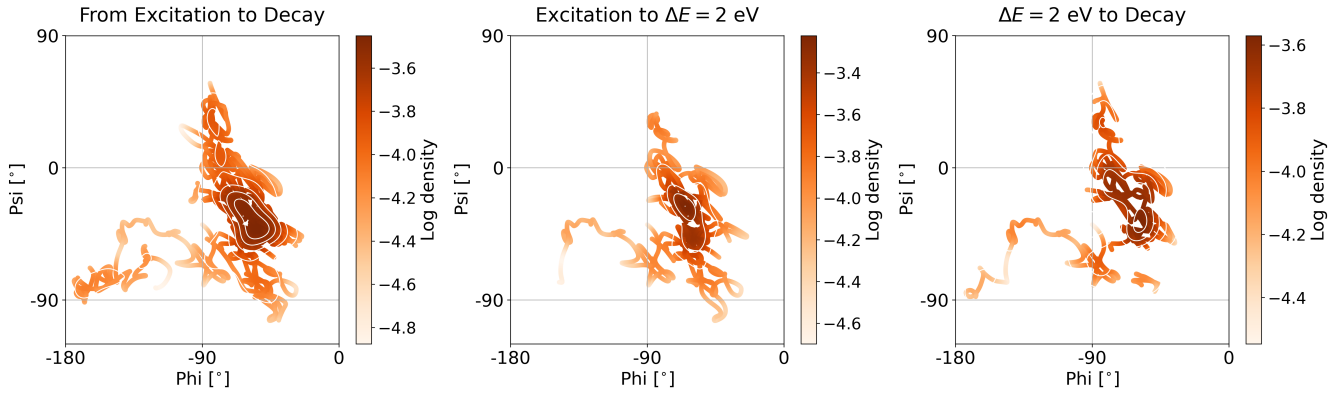


Figure 4.6: Two-dimensional probability density of Ψ/Φ dihedral angles in the Ramachandran plots for decayed trajectories in cluster-1 plotted using dihedrals on the most affected helical turn. From the excitation to decay (left panel), from excitation to energy gap of 2 eV (middle panel), and from energy gap of 2 eV to decay (right panel).

Overall, these findings suggest the following scenario: the ground-state structures, initially in their native helical conformation, may undergo vibrational relaxation upon excitation, leading to gradual helical distortion. Vibrational modes such as Cα–N and C–O bond stretching appear to drive the system toward a conical intersection. In cases where vibrational relaxation does not significantly distort the secondary structure, the configuration remains trapped in the lowest excited state, which may result in emission. This behavior is illustrated in Figure 4.7 below.

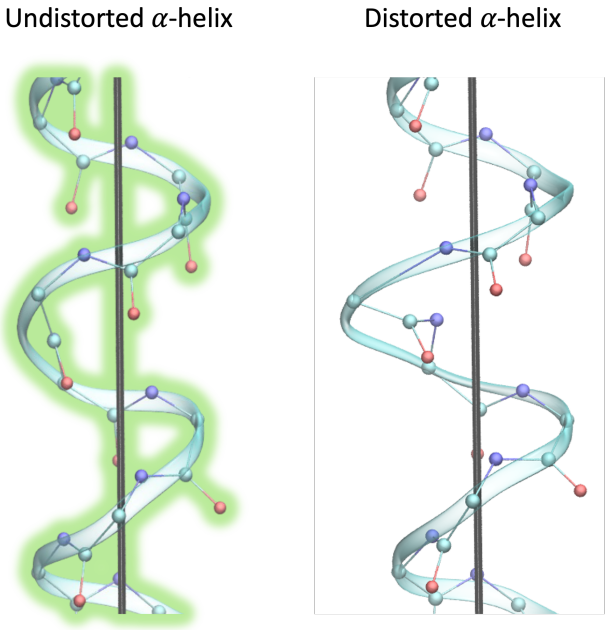


Figure 4.7: Illustration of cluster-1 structures that exhibit emission: the glowing, undistorted α-helix (left panel), and the non-emissive, distorted helical structure (right panel).

Such emission were computed using 40% of the trajectories in cluster-1 which did not decay within 1 ps. This is under the assumption that they would remain sufficiently long in the S_1 state to emit radiation. The left panel Figure 4.8 show the normalized estimated emission which lies below 400 nm, the vertical line is the lowest $S_0 \rightarrow S_1$ transition. This emission is weak as can be seen from the right panel of Figure 4.8. It show distribution of the oscillator strength of the emission which is below 0.01.

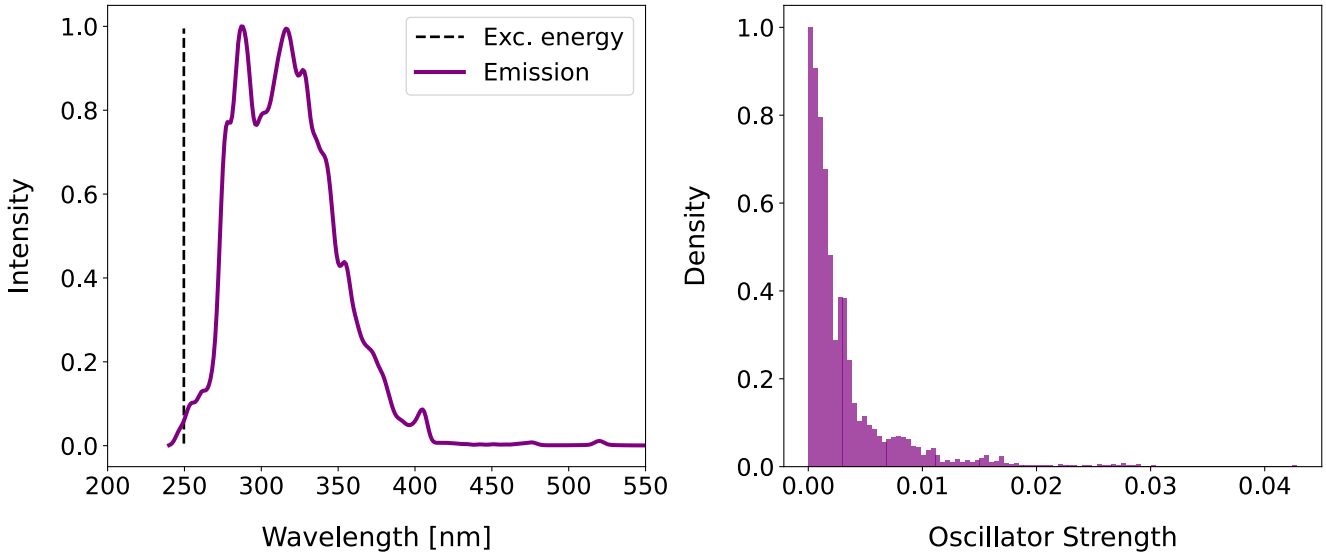


Figure 4.8: The calculated lowest excitation energy and emission spectra in cluster-1 (on left) and the corresponding oscillator strength (on right).

4.3.4 Cluster-2 Vibrational Relaxation Modes and Estimated Emission

The excitations in cluster-2 are n $\rightarrow \pi^*$ localized on carboxylate group of glutamic acid side chain. In this cluster, proton transfer was identified as the dominant decay pathway. The left panel of Figure 4.9 show representative molecular structures within this cluster, which are characterized by a hydrogen bond between a carboxylate group and amino group. In this context, proton transfer refers to the movement of a hydrogen atom from the nitrogen atom where it is initially bonded, to the oxygen atom of the carboxylate group. Accordingly, the proton transfer coordinate is defined as the difference between the N-H and O-H bond lengths.

The right panel of Figure 4.9 show proton transfer distribution for cluster-2 trajectories, in ground state (blue) and in the excited state (orange). In absence of proton transfer, the O-H bond is longer than the N-H bond, indicating that H and O are not in close proximity. This behavior is observed in ground state, where proton transfer coordinate assumes positive values, typically between 0 and 1, or in some cases, significantly higher. In contrast, upon proton transfer, the

proton transfer coordinate crosses zero to negative values, as observed in excited state.

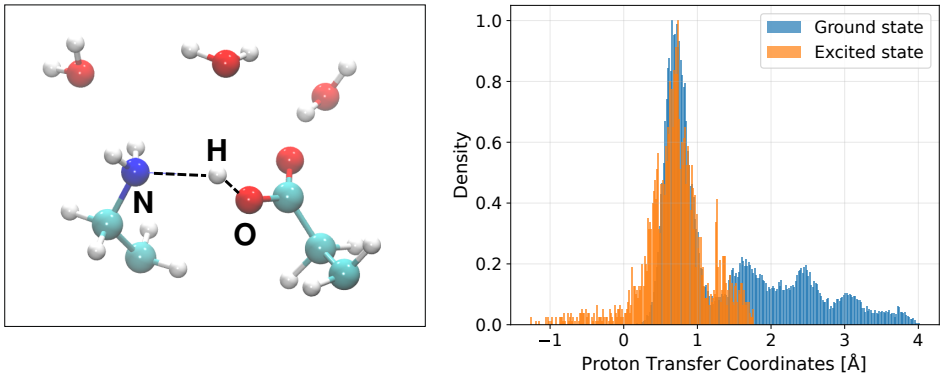


Figure 4.9: Left panel: illustration of the proton (H) transfer from a nitrogen (N) atom on the side chain of lysine amino acid, to an oxygen (O) atom on glutamic acid side chain. Right panel: density plot of proton transfer coordinates in cluster-2, for ground state (in blue) and excited state (in orange).

This mode was found to correlate well with the energy gap. Specifically, as the proton move from the nitrogen atom towards the oxygen atom, the energy gap decreases. This is shown in a box plot for energy gap versus proton transfer coordinate. In the plot, the lower and upper boundaries of each box represent the 25^{th} and 75^{th} percentile of the energy gap distribution respectively, while the horizontal line within the box indicate the median value. The lines extending from the box indicate the minimum and maximum energy gap values. Overall the figure show a clear trend of that decreasing energy gap as the proton transfer occurs.

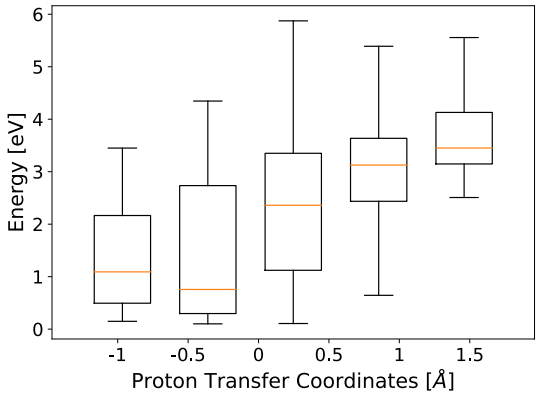


Figure 4.10: A box plot for the correlation of energy gap with proton transfer coordinates in cluster-2.

We also examined several other vibrational modes that have been reported to influence the stabilization or destabilization of the the excited state, including the stretching of the C-O bond

[63] in on the carboxylate group in our system. However these modes were found not to be directly correlated with the energy gap. Compared to the ground state, these bonds get longer in the excited state, irrespective of whether the dynamics decay or not. This may be attributed to the fact that it is at this carboxylate group that the initial local excitation happens. Figure 4.11 depicts the distribution of bond lengths between the carbon and oxygen atoms of the carboxylate groups in cluster-2 for electronic ground and excited state, decayed and nondecayed structures.

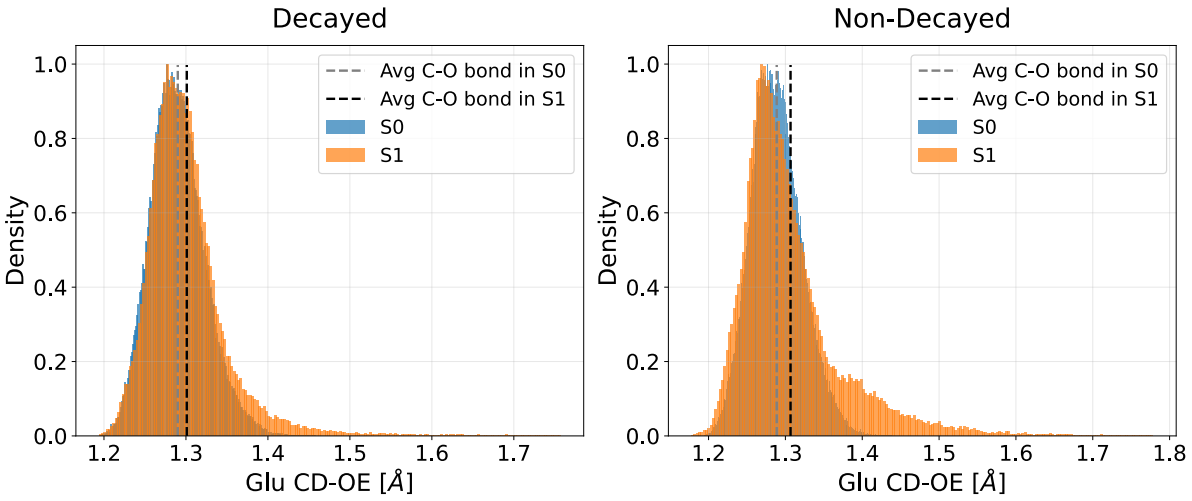


Figure 4.11: Distribution of C-O bond lengths, in ground and excited states for decayed (on left) and non-decayed dynamics (on right) in cluster-2. Vertical dashed lines indicate the average bond length in each state.

Assuming that 46% of the trajectories that did not decay within 1 ps would remain sufficiently long in S_1 state to emit radiation, their overall computed emission and corresponding oscillator strength is presented in Figure 4.12. Upon excitation around 224 nm, indicated as a vertical dashed line, cluster-2 is expected to emit in range 250-550 nm, in agreement with the reported range [2]. However, the oscillator strength distribution, on the right panel of the same figure, indicates that this emission very weak, below 0.002, ten times less than that of cluster-1.

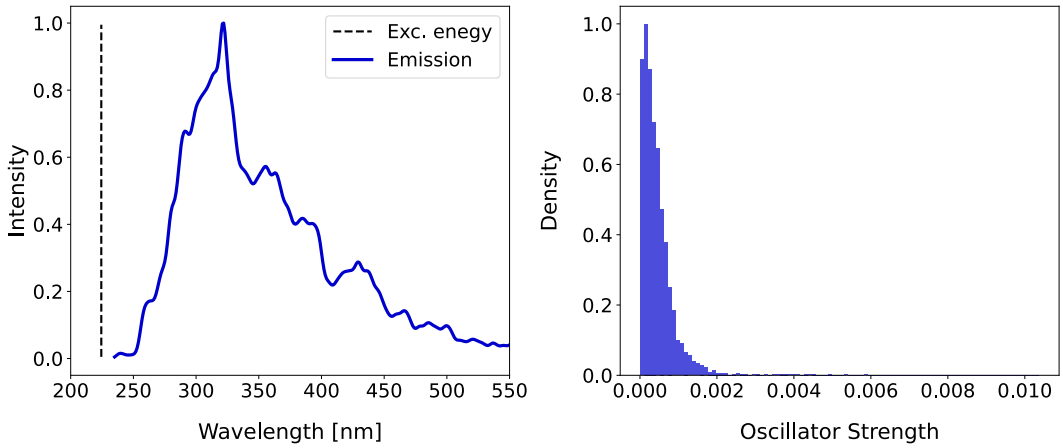


Figure 4.12: The calculated lowest excitation energy and emission spectra in cluster-2 (on left) and the corresponding oscillator strength distribution (on right).

4.3.5 Cluster-3 Vibrational Relaxation Modes and Estimated Emission

Lastly, in cluster-3, n$\rightarrow \pi^*$ transitions are HOMO-LUMO charge transfers, with HOMO located on a carboxylate group of glutamic acid and LUMO on guanidinium group of arginine residue. We found that the primary mode driving decay of all of this cluster's trajectories is arginine deplanarization.

In its ideal geometry, the guanidinium group's central carbon atom and the three nitrogen atoms to which it is bonded, all lie on the same plane. Arginine deplanarization refers to the distance from the central carbon to the plane formed by the three nitrogens, as illustrated in panel (a) and (b) of Figure 4.13: for the carbon and the three nitrogen atoms in plane and the carbon out of the plane made by the three nitrogens respectively. This distance exhibits greater fluctuations in the excited state compared to the ground state. In ground state it remain bounded by 0.15 Å, whereas in the excited state it can reach values up 0.4 Å, as shown in the histogram in panel (c) of Figure 4.13.

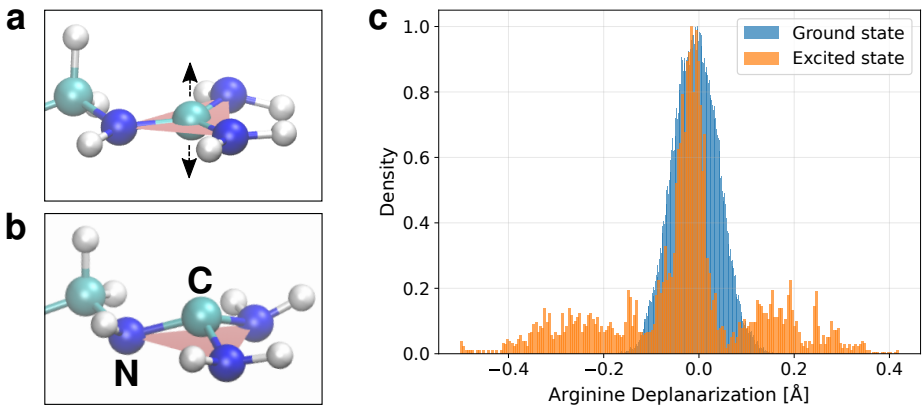


Figure 4.13: Panel (a) depicts carbon lying in the three nitrogen plane, and (b) carbon out of the plane their plane (arginine deplanarization). Panel (c) show density plot of arginine deplanalization in cluster-3, for ground state (in blue), and excited state (orange) structures.

The correlation between arginine deplanarization and the energy gap is illustrated in the box plot shown in Figure 4.14. As evident from the figure, both maximum and median values of energy gap drop sharply once the deplanarization fluctuations exceeds 0.15 Å. This increase in plane distortions accelerates the system's progression towards the conical intersection.

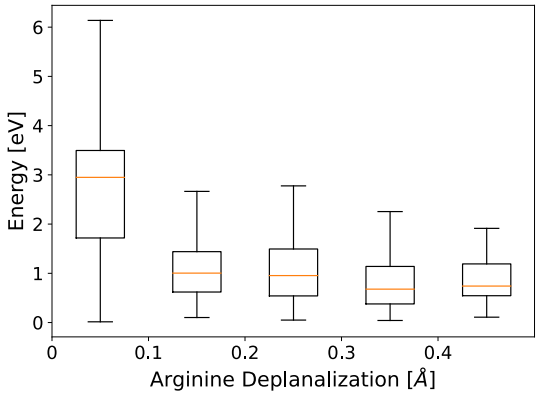


Figure 4.14: A box plot for correlation between energy gap and arginine deplanarization.

Even if both clusters 2 and 3 are centered on hydrogen bonding that involves carboxylate group, unlike cluster-2, the excited carboxylate C-O bond in cluster-3 do not deviate from their ground state values in a similar way. For the decayed group, C-O bond lengths are slightly shorter than their corresponding ground state values, whereas, for the non-decayed structures, they become significantly longer, as illustrated in Figure 4.15.

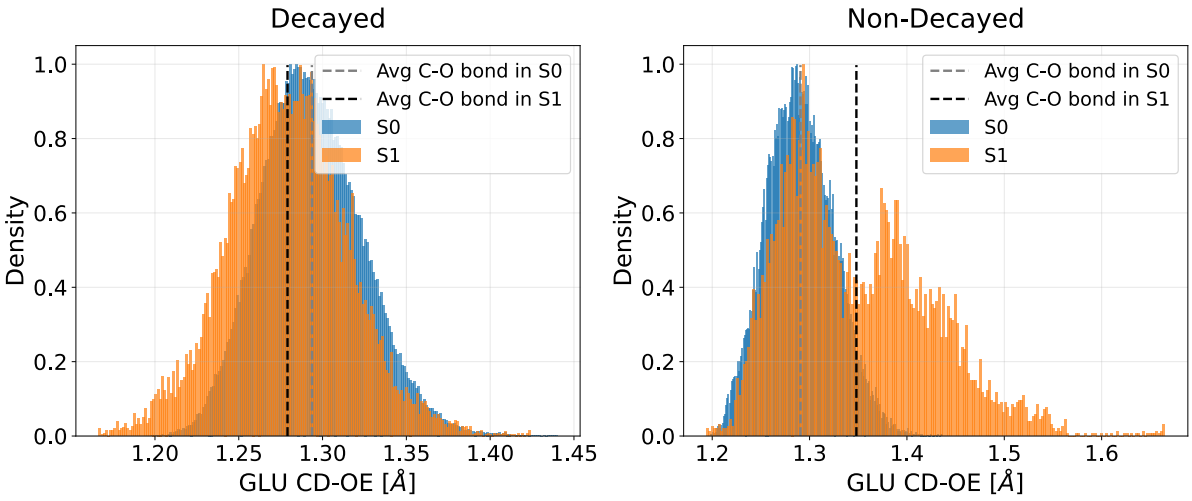


Figure 4.15: Distribution of the C-O bond lengths in carboxylate groups, in ground and excited states for decayed (on left) and non-decayed dynamics (on right) in cluster-3. Vertical dashed lines indicate the average bond length in each state.

Therefore, we observe that the primary vibrational mode responsible for driving the decay of Cluster-3 structures - arginine deplanarization - is coupled to slightly shorter C-O bond lengths. In contrast, excited-state structures that do not decay exhibit larger fluctuations in C-O bond lengths but maintain much lower arginine deplanarization, closely resembling the ground-state configuration.

Cluster-3 has been identified to be the one that is more prone to low energy electronic excitations, but also the one which contributes the least to the emission of α_3C. However, we also sought to identify the spectral region to which the surviving long-lived structures would contribute to emission. Therefore, we estimated its emission spectrum using its nondecayed structures, and the corresponding oscillator strength. From Figure 4.16 below, we can see that this cluster exhibits the emission in the whole reported range, 300-350 nm, with peak in range 350-400 nm. Its oscillator strength however, is very small, below 0.001.

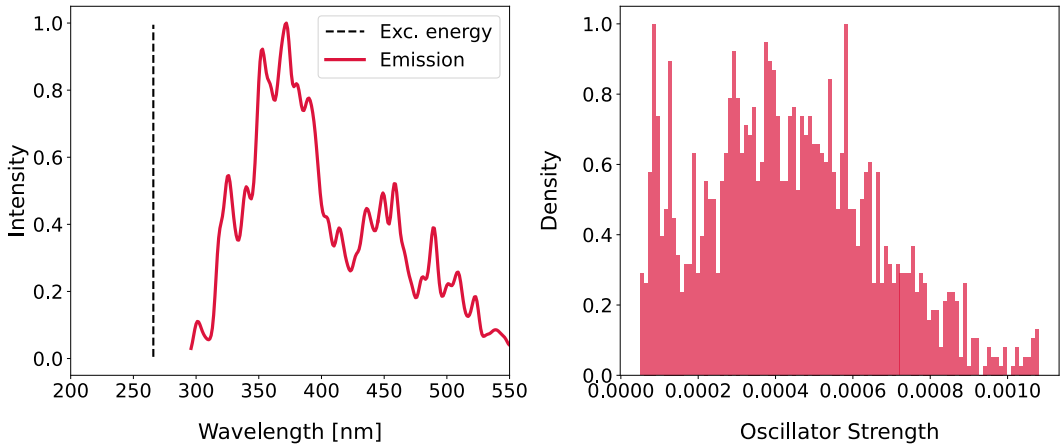


Figure 4.16: The calculated lowest excitation energy and emission spectra in cluster-2 (on left) and the corresponding oscillator strength distribution (on right).

4.4 Relative Emission of the Three Clusters

Comparison of the results obtained for each of these clusters suggests that, when their secondary structure is not distorted, chromophores in cluster-1 may emit in range 250-400 nm, exhibiting the maximum intensity between 280-320 nm. In cluster-2, as long as the vibrational modes hinder the proton transfer, emission may occur primarily between 250-400 nm, exhibiting the maximum intensity around 320 nm, and low intensity extending up to 550 nm. Structures in cluster-3 where vibrational relaxations hinder the arginine deplanalization, exhibit a more red-shifted emission spectrum, spanning 300-550 nm, with maximum intensity around 400 nm. These trends are illustrated in Figure 4.17 (left panel).

The right panel shows distributions of their corresponding oscillator strengths on a logarithmic scale. Although all oscillator strengths are weak (below 0.04), their magnitudes differ substantially: for cluster-3, values lie between 10^{-4}-10^{-3}; for cluster-2 they can reach up to 10^{-2}; and for cluster-1, they extend an order of magnitude higher, up to 10^{-1}.

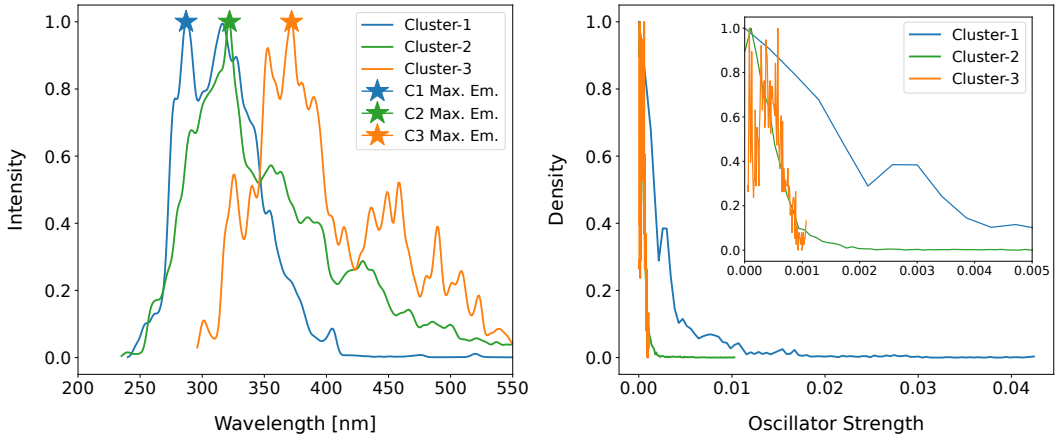


Figure 4.17: Comparison between computed emission (left panel) and density of the oscillator strength (right panel), for cluster-1 (blue curve), cluster-2 (green curve) and cluster-3 (orange curve).

Altogether, these results indicate that cluster-3 contributes the least to the emission of α_3C, with its weak oscillator strength and primary emission occurring around 400 nm and beyond. Cluster-2 shows an intermediate contribution, emitting at slightly higher energies with moderately increased oscillator strength. In contrast, cluster-1 contributes predominantly around 300 nm, exhibiting the highest oscillator strength among the three. Combined, these findings reproduce a pattern consistent with the reported experimental emission spectrum, which displays a maximum around 350 nm and strong intensity between 310-400 nm.

4.5 Conclusions and Perspectives

In this chapter we focused on anomalous emission by an nonaromatic α_3C protein that have been reported by Kumar and the collaborators [2]. Their study showed that the protein exhibits weak emission in 310-550 nm region of the spectrum, when excited at 295 nm.

Using the main hydrogen bonding motifs for local interactions in α_3C that have been identified in Chapter 3, we conducted their excited dynamics in QM/MM framework, where the electronic structure of the clusters configurations were described by DFTB/TD-DFTB, taking into account the effects of their environment, that includes both protein and solvent.

Our preliminary results provide insight into the origin of the weak emission observed in α_3C, as well as the main vibrational modes through which the system relaxes nonradiativelly back to the ground state. Assuming that the structures stable during our simulation time remain sufficiently long in the excited state, and that the relative proportions of decayed and non-decayed trajectories within the three clusters would not change significantly with a larger sampling size, we draw the following conclusion: chromophores in cluster-1 are likely the primary contributors to the observed

weak emission of the α_3C, whereas cluster-3 contributes the least. The combined contributions of all clusters reproduce a spectrum similar to that reported in the experiment [2].

The dominant vibratonal relaxation pathways responsible for nonradiative decay differ among the clusters: helical distrortions, mainly driven by the Cα-N bond stretching, in cluster-1, proton transfer in cluster-2 and arginine deplanalization in cluster-3. Fluorophores in this context, correspond to structures within each cluster in which the vibrational modes are not significantly active.

These findings, together with those discussed in the previous chapter, suggest that many cluster-3 structures may enhance absorption but not emission. The secondary α-helical structure of the α_3C protein appears to play a crucial role in stabilizing n$\rightarrow \pi^*$ transitions in the excited states. A similar conclusion has been drawn in studies investigating the optical properties of amyloids, where emission has been associated with their cross-β sheet secondary structure [38, 45]. Moreover, a recent study by Gonzalez and the collaborators reported nonaromatic fluorescence arising from a single α-helix in solution, which they attributed to α-helical folding [57].

Based on our current identification of the possible decay-associated vibrational modes across the three clusters, the next step will be to verify whether these modes indeed drive the decay process through constrained dynamics simulations. It would also be of interest to perform non-adiabatic ab initio molecular dynamics to examine the relative population of the excited states and to determine whether the protein returns to its native structure after relaxing back to the ground state.

Chapter 5

Conclusions and Perspectives

In this work, we explore the optical properties of nonaromatic systems using α_3C as a case study. This topic has attracted significant research interest because these systems display unconventional optical behaviors that challenge the long-standing belief that only aromatic and conjugated compounds absorb light beyond 250 nm. As discussed in the Introduction [1], several studies have reported systems exhibiting such unconventional absorption and emission properties [45, 46, 47, 48, 49, 50, 54, 55, 56, 57], and various suggestions have been proposed to explain the origins of these effects [38, 46, 52, 54, 55, 63]. However the general mechanism remains unclear.

The focus of our study is on a synthetic, and highly charged nonaromatic protein, α_3C, which have been reported by Prasad and collaborators to absorb light in the 250-800 nm range, in its monomeric form [1]. In their work they investigated the origins of this absorption, assuming that the key interactions involved lysine and glutamic acid residues. They performed the the electronic structure calculations on dimers of these amino acids in vacuum, and successfully reproduced the experimental absorption spectra. From this, they concluded that the origin of long tail arises form charge transfer interactions between spatial proximal (5-6 Å) glutamic acid and lysine residues.

Additionally, the same protein has been reported by Kumar and the collaborators [2] to exhibit fluorescence. In a comparative analysis of the emission spectra of N-Acetyl-L-tryptophanamide (NATA) and the α_3W mutant (a tryptophan-containing variant of α_3C) under 295 nm excitation, they observed that α_3C displayed weak emission between 310 and 550 nm.

While lysine-glutamic acid interactions may indeed play an important role, we argue that they are not sufficient to account for all relevant interactions in α_3C. Given that the protein contains highly charged side chains and exists in aqueous solution, it may be inaccurate to neglect solvent effects in electronic structure calculations. This limitation becomes especially significant when charged residues are spatially close, as their interactions may be mediated by solvent molecules and other nearby side chains.

In this work, we investigated the origins of the absorption and emission properties of α_3C in an agnostic manner. Rather than selecting specific interactions a priori, we use unsupervised machine learning approaches to identify relevant structural motifs. These motifs then serve as initial

conditions for optical property calculations, which are performed within a QM/MM framework to account for environmental effects. The theoretical foundations of these methods are presented in Chapter 2. In Chapter 3, we establish absorption properties by performing electronic ground-state optimizations followed by single-point excited-state calculations. In Chapter 4, we examine emission properties through excited-state dynamics simulations in the lowest excited state.

We identify three main types of local structures in the system: backbone interactions , amino-carboxylate interactions (primarily between lysine and glutamic acid), and guanidinium-carboxylate interactions (between arginine and glutamic acid). Our absorption spectra calculations reveal unconventional absorption features spanning 250-350 nm, with the arginine–glutamic acid interactions contributing to all transitions beyond 300 nm. However, the simulated absorption tail is much shorter than what is observed experimentally. In particular, our calculations do not predict any absorption between 400 and 800 nm. Transitions in this range appear only when environmental interactions are neglected-a simplification we consider physically inaccurate. To investigate the source of this discrepancy between simulated and experimental results, we examined the influence of nuclear quantum effects on the absorption spectra. We found that these effects significantly broaden the spectrum and redshift it by approximately 100 nm, extending the calculated absorption to around 450 nm.

Finally we focused on the emission properties of α_3C. It was found that the arginine-glutamic acid structures, although associated with low electronic energy transitions, contribute the least to the experimentally observed weak emission. These structures primarily relax from the excited state back to the ground state through arginine deplanarization. In contrast, backbone interactions appear to play a more significant role: as long as the α-helical structure remains undistorted, these regions exhibit the strongest emission within the 250-400 nm range. When considering contributions from all three types of interactions together, the resulting emission spectrum is in good agreement with the experimentally reported one.

These results highlight challenges and opportunities from both computational and experimental perspectives. On the computational side, it would be valuable to investigate the absorption properties while explicitly including NQEs to better assess their influence and enable comparison with existing results. Experimentally, since our findings suggest that arginine-glutamic acid interactions are more prone to absorb photons at longer wavelengths, it would be interesting to study α_3C variants with certain residues mutated to arginine. We speculate that such mutations might enhance absorption but not emission. As our results indicate that emission depends primarily on maintaining the α-helical structure, it would also be worthwhile to identify the smallest peptide capable of exhibiting near-UV absorption and emission.

Appendix A

Supporting Information for Chapter 3

Structural Stability of α_3C over 1 μs MD Trajectory

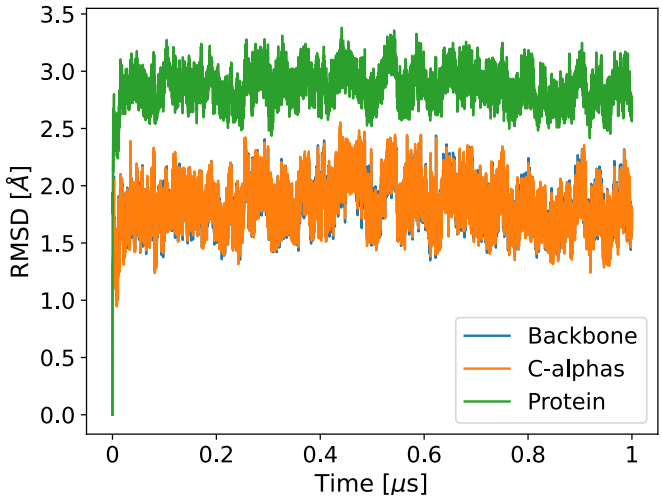


Figure A.1: Root mean square displacement of α_3C calculated over a one microsecond trajectory.

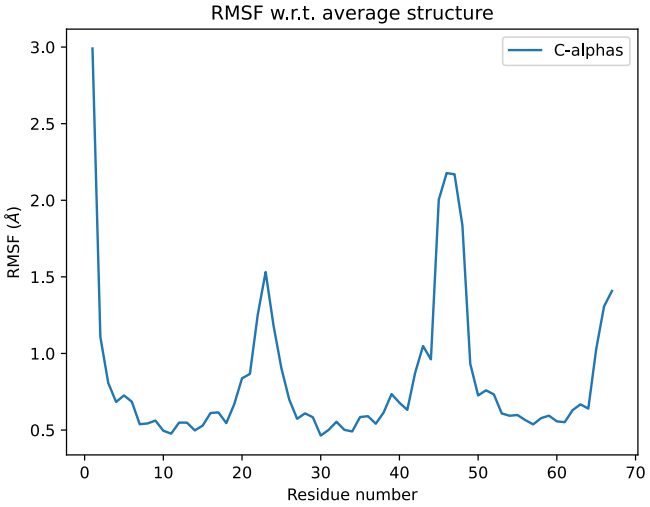


Figure A.2: Root mean square fluctuation of residues of α_3C calculated over a one microsecond trajectory.

SOAP Cutoff Radius

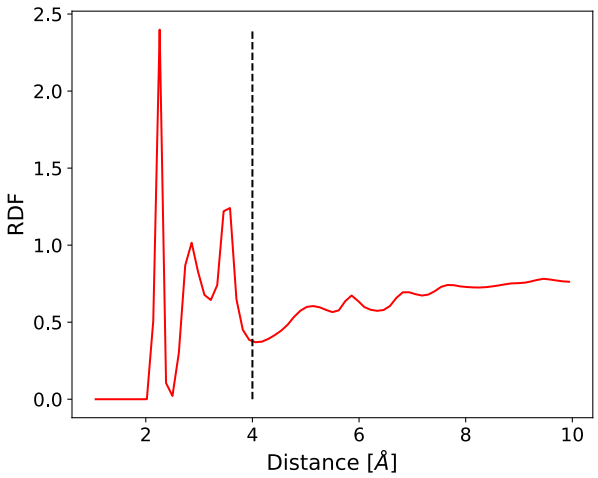


Figure A.3: Radial distribution function between all nitrogen atoms in α_3C protein and all oxygen atoms in the system. We chose R_{cut}=4 Å the value below which all the sharp peaks were found.

DPA Clusters for Various Z and R_{cut} Parameters

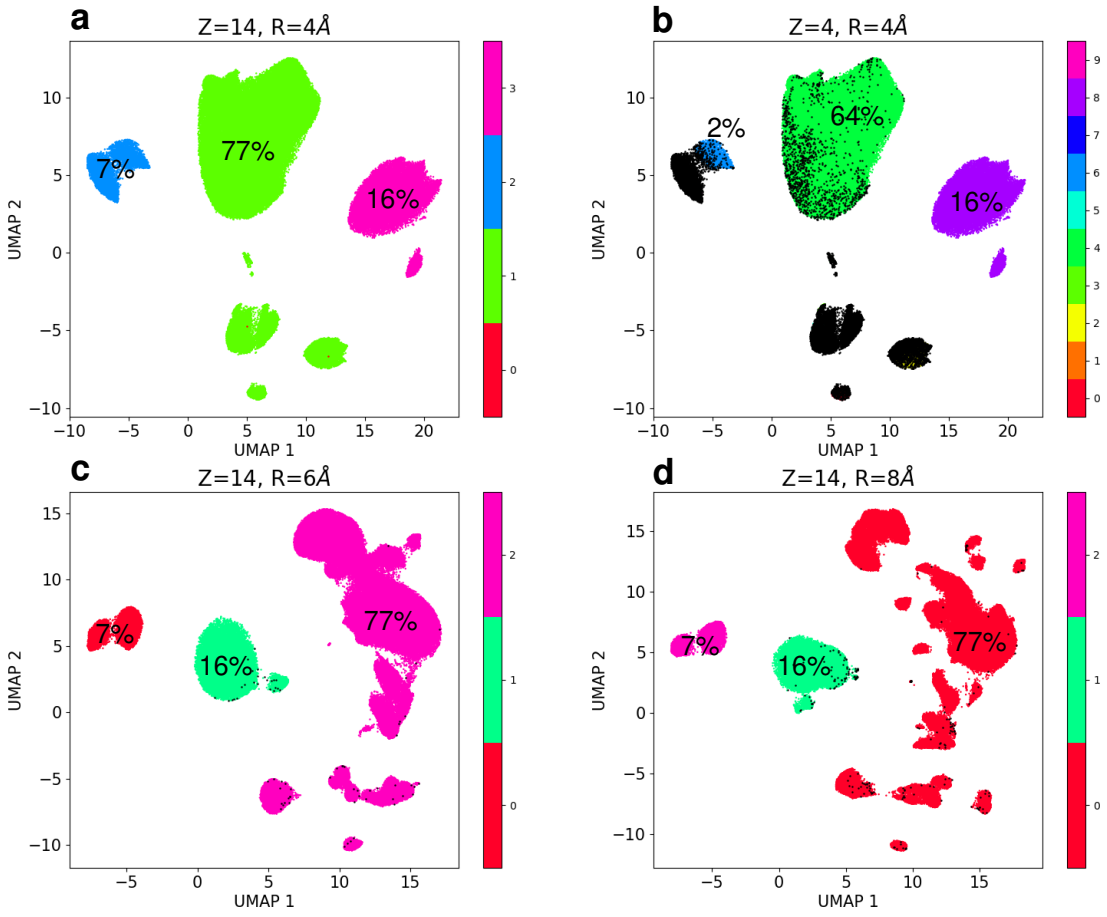


Figure A.4: DPA clusters for various Z and R_{cut} values. Panel a: $Z = 14$ and $R_{cut} = 4$Å, panel b: $Z = 4$ and $R_{cut} = 4$Å, panel c: $Z = 14$ and $R_{cut} = 6$Å, and panel d: $Z = 14$ and $R_{cut} = 8$Å.

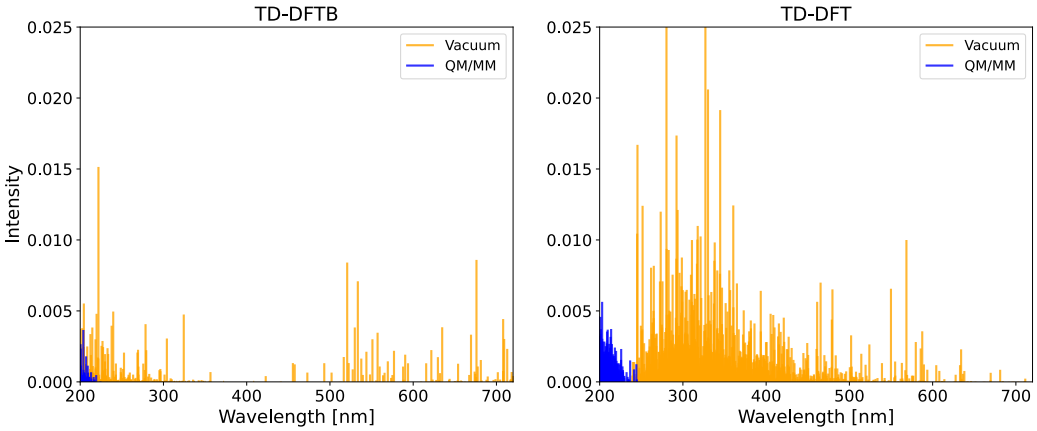
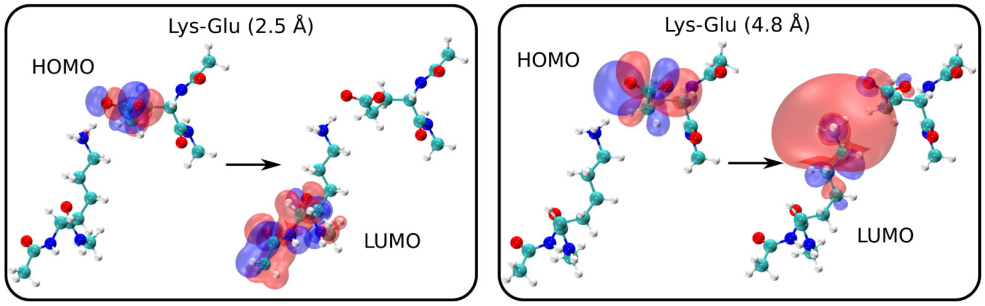


Figure A.5: Excition energy obtained from 100 conformations of Cluster-2 calculated using TD-DFTB and TD-DFT/CAM-B3LYP/6-311G(d,p), in QM/MM (blue lines) and in vacuum (orange lines).

Cluster-2 Absorption: TD-DFTB vs TD-DFT



First excitation energy [nm]

Lys-Glu (2.5 Å)		Lys-Glu (4.8 Å)	
TD-DFTB	TD-DFT	TD-DFTB	TD-DFT
352	304	678	671

Figure A.6: Comparison between TD-DFTB and TD-DFT predictions of the lowest-energy electronic transition wavelengths for two characteristic Lys-Glu dimers at distances of 2.5 Å (a) and 4.5 Å(b). The upper panel illustrates the molecular orbitals involved in the charge transfer transition. The lower panel provides a numerical comparison of the lowest-energy electronic transition wavelengths (in nm) calculated using TD-DFTB and CAM-B3LYP for the two dimers. The corresponding first excitation energies are 3.52, 4.08, 1.83 and 1.85 eV for 352, 304, 678, 671 nm respectivelly.

The Stability of the Hydrogen Bonds throughout the 5 ps QM/MM Trajectory

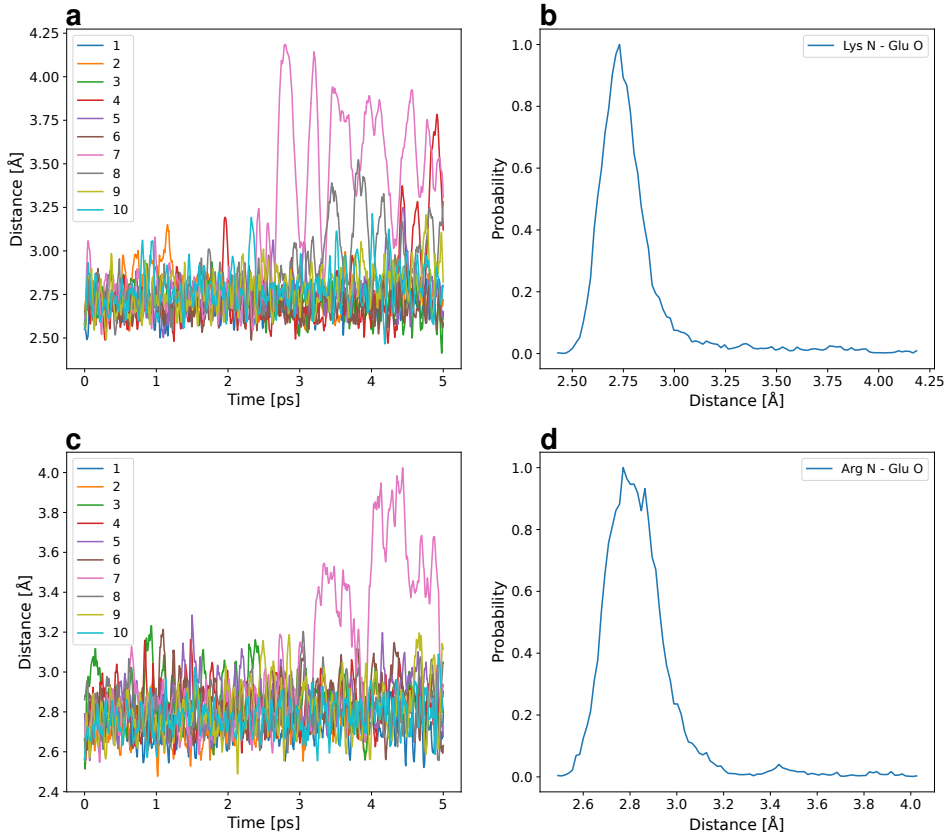


Figure A.7: Nitrogen - Oxygen bond lengths. Panel (a) shows the time series of the QM/MM dynamics for lysine - glutamic acid salt bridge in cluster-2, with its corresponding distribution in panel (b). Panel (c) shows the time series in the QM/MM dynamics for arginine - glutamic acid salt bridge in cluster-3, with the corresponding distribution in panel (d).

The Convergence of the Absorption Spectra QM/MM Trajectory-wise

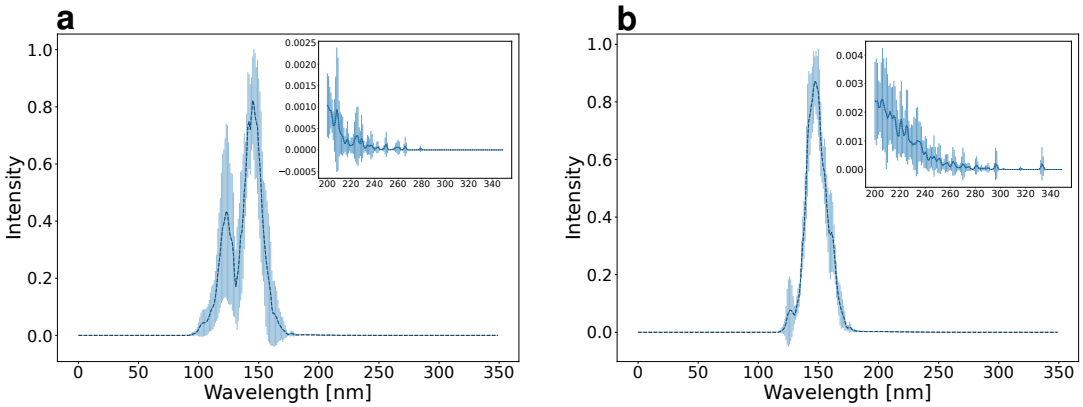


Figure A.8: Average and standard deviation for the absorption spectra calculated on configuration sampled from cluster-2 (panel a) and cluster-3 (panel b).

Bibliography

[1] Saumya Prasad, Imon Mandal, Shubham Singh, Ashim Paul, Bhubaneswar Mandal, Ravindra Venkatramani, and Rajaram Swaminathan. Near uv-visible electronic absorption originating from charged amino acids in a monomeric protein. *Chemical science*, 8(8):5416–5433, 2017.

[2] Amrendra Kumar, Shah E Alom, Dileep Ahari, Anurag Priyadarshi, Mohd Ziauddin Ansari, and Rajaram Swaminathan. Role of charged amino acids in sullying the fluorescence of tryptophan or conjugated dansyl probe in monomeric proteins. *Biophysical Journal*, 121(3):416a, 2022.

[3] Maria d'Errico, Elena Facco, Alessandro Laio, and Alex Rodriguez. Automatic topography of high-dimensional data sets by non-parametric density peak clustering. *Information Sciences*, 560:476–492, 2021.

[4] Rico Gutzler, Manish Garg, Christian R Ast, Klaus Kuhnke, and Klaus Kern. Light–matter interaction at atomic scales. *Nature Reviews Physics*, 3(6):441–453, 2021.

[5] Dario Leister. Photobiology: Introduction, overview and challenges. *Frontiers in Photobiology*, 1:1253330, 2023.

[6] Andrew Lininger, Giovanna Palermo, Alexa Guglielmelli, Giuseppe Nicoletta, Madhav Goel, Michael Hinczewski, and Giuseppe Strangi. Chirality in light–matter interaction. *Advanced Materials*, 35(34):2107325, 2023.

[7] Xinjie Yang, Jianjian Huang, Juan Guo, Shuran Fang, Zhiming Wang, Guojiao Wu, Yuzhou Wu, and Fangrui Zhong. Bridging chemistry and biology for light-driven new-to-nature enantioselective photoenzymatic catalysis. *Chemical Society Reviews*, 54(11):5157–5188, 2025.

[8] Siegfried Wahl, Moritz Engelhardt, Patrick Schaupp, Christian Lappe, and Iliya V Ivanov. The inner clock—blue light sets the human rhythm. *Journal of biophotonics*, 12(12):e201900102, 2019.

[9] Brian Wardle. *Principles and applications of photochemistry*. John Wiley & Sons, 2009.

[10] Robert W Schoenlein, Linda A Peteanu, Richard A Mathies, and Charles V Shank. The first step in vision: femtosecond isomerization of rhodopsin. *Science*, 254(5030):412–415, 1991.

[11] Sofía Miñano, Stuart Golodetz, Tommaso Cavallari, and Graham K Taylor. Through hawks' eyes: synthetically reconstructing the visual field of a bird in flight. *International Journal of Computer Vision*, 131(6):1497–1531, 2023.

[12] Mathilde Delacoux and Fumihiro Kano. Fine-scale tracking reveals visual field use for predator detection and escape in collective foraging of pigeon flocks. *Elife*, 13:RP95549, 2024.

[13] Mani Shrestha, Adrian G Dyer, Skye Boyd-Gerny, Bob BM Wong, and Martin Burd. Shades of red: bird-pollinated flowers target the specific colour discrimination abilities of avian vision. *New Phytologist*, 198(1):301–310, 2013.

[14] Jingyu Liu, Tao-Yuan Du, Xuan Deng, Bo Li, Qianqian Cheng, Jiangong Hu, and Kaiyao Huang. Theory for ultrafast energy transfer in photosynthesis. *Physical Chemistry Chemical Physics*, 27(6):2908–2919, 2025.

[15] Steven HD Haddock, Mark A Moline, and James F Case. Bioluminescence in the sea. *Annual review of marine science*, 2(1):443–493, 2010.

[16] Ana Belén Muñoz-García, Iacopo Benesperi, Gerrit Boschloo, Javier J Concepcion, Jared H Delcamp, Elizabeth A Gibson, Gerald J Meyer, Michele Pavone, Henrik Pettersson, Anders Hagfeldt, et al. Dye-sensitized solar cells strike back. *Chemical Society Reviews*, 50(22):12450–12550, 2021.

[17] Ruby Baby, Peter Daniel Nixon, Nallapaneni Manoj Kumar, MSP Subathra, and Nallamuthu Ananthi. A comprehensive review of dye-sensitized solar cell optimal fabrication conditions, natural dye selection, and application-based future perspectives. *Environmental Science and Pollution Research*, 29(1):371–404, 2022.

[18] Jessica Barichello, Paolo Mariani, Luigi Vesce, Donatella Spadaro, Ilaria Citro, Fabio Matteocci, Antonino Bartolotta, Aldo Di Carlo, and Giuseppe Calogero. Bifacial dye-sensitized solar cells for indoor and outdoor renewable energy-based application. *Journal of Materials Chemistry C*, 12(7):2317–2349, 2024.

[19] Johannes Karges. Clinical development of metal complexes as photosensitizers for photodynamic therapy of cancer. *Angewandte chemie international edition*, 61(5):e202112236, 2022.

[20] Shambo Mohanty, Vaibhavi Meghraj Desai, Rupesh Jain, Mukta Agrawal, Sunil Kumar Dubey, and Gautam Singhvi. Unveiling the potential of photodynamic therapy with nanocarriers as a compelling therapeutic approach for skin cancer treatment: current explorations and insights. *RSC advances*, 14(30):21915–21937, 2024.

[21] João CS Simões, Sophia Sarpaki, Panagiotis Papadimitroulas, Bruno Therrien, and George Loudos. Conjugated photosensitizers for imaging and pdt in cancer research. *Journal of medicinal chemistry*, 63(23):14119–14150, 2020.

[22] Trailokya Bhattarai, Abasifreke Ebong, and Mohammad Yasin Akhtar Raja. A review of light-emitting diodes and ultraviolet light-emitting diodes and their applications. In *photonics*, volume 11, page 491. MDPI, 2024.

[23] Hakan Inan, Muhammet Poyraz, Fatih Inci, Mark A Lifson, Murat Baday, Brian T Cunningham, and Utkan Demirci. Photonic crystals: emerging biosensors and their promise for point-of-care applications. *Chemical society reviews*, 46(2):366–388, 2017.

[24] David Thomson, Aaron Zilkie, John E Bowers, Tin Komljenovic, Graham T Reed, Laurent Vivien, Delphine Marris-Morini, Eric Cassan, Léopold Virot, Jean-Marc Fédéli, et al. Roadmap on silicon photonics. *Journal of Optics*, 18(7):073003, 2016.

[25] Van-Nghia Nguyen, Minh Viet Nguyen, Huong Pham Thi, Anh-Tuan Vu, and Truong Xuan Nguyen. Recent advances in near-infrared organic photosensitizers for photodynamic cancer therapy. *Biomaterials Science*, 2025.

[26] Yunhua Zhang, Chengyuan Qian, Yuncong Chen, Weijiang He, and Zijian Guo. Phototherapy via modulation of β-amyloid in combating alzheimer's disease. *Aggregate*, 6(5):e70020, 2025.

[27] Susan Monro, Katsuya L Colon, Huimin Yin, John Roque III, Prathyusha Konda, Shashi Gujar, Randolph P Thummel, Lothar Lilge, Colin G Cameron, and Sherri A McFarland. Transition metal complexes and photodynamic therapy from a tumor-centered approach: challenges, opportunities, and highlights from the development of tld1433. *Chemical reviews*, 119(2):797–828, 2018.

[28] Francesco D'Amico, Bas de Jong, Matteo Bartolini, Daniele Franchi, Alessio Dessì, Lorenzo Zani, Xheila Yzeiri, Emanuela Gatto, Annalisa Santucci, Aldo Di Carlo, et al. Recent advances in organic dyes for application in dye-sensitized solar cells under indoor lighting conditions. *Materials*, 16(23):7338, 2023.

[29] Sumit Sahil Malhotra, Mukhtar Ahmed, Manoj Kumar Gupta, and Azaj Ansari. Metal-free and natural dye-sensitized solar cells: recent advancements and future perspectives. *Sustainable Energy & Fuels*, 8(18):4127–4163, 2024.

[30] Martin A Green, Ewan D Dunlop, Jochen Hohl-Ebinger, Masahiro Yoshita, Nikos Kopidakis, Karsten Bothe, David Hinken, Michael Rauer, and Xiaojing Hao. Solar cell efficiency tables (version 60). *Progress in Photovoltaics*, 30(7), 2022.

[31] Angellina Ebenezer Anitha and Marius Dotter. A review on liquid electrolyte stability issues for commercialization of dye-sensitized solar cells (dssc). *Energies*, 16(13):5129, 2023.

[32] Nicholas J Turro, Vaidhyanathan Ramamurthy, and Juan C Scaiano. *Principles of molecular photochemistry: an introduction*. University science books, 2009.

[33] Joseph R Lakowicz. *Principles of fluorescence spectroscopy*. Springer, 2006.

[34] Helen H Fielding and Graham A Worth. Using time-resolved photoelectron spectroscopy to unravel the electronic relaxation dynamics of photoexcited molecules. *Chemical Society Reviews*, 47(2):309–321, 2018.

[35] Michael Kasha. Characterization of electronic transitions in complex molecules. *Discussions of the Faraday society*, 9:14–19, 1950.

[36] Annalisa Bortolotti, Yin How Wong, Stine S Korsholm, Noor Hafizan B Bahring, Sara Bobone, Saad Tayyab, Marco van de Weert, and Lorenzo Stella. On the purported "backbone fluorescence" in protein three-dimensional fluorescence spectra. *Rsc Advances*, 6(114):112870–112876, 2016.

[37] Tomáš Polívka and Villy Sundström. Ultrafast dynamics of carotenoid excited states- from solution to natural and artificial systems. *Chemical reviews*, 104(4):2021–2072, 2004.

[38] Luca Grisanti, Marin Sapunar, Ali Hassanali, and Naa Došlić. Toward understanding optical properties of amyloids: a reaction path and nonadiabatic dynamics study. *Journal of the American Chemical Society*, 142(42):18042–18049, 2020.

[39] Salmahaminati and Daniel Roca-Sanjuan. The photophysics and photochemistry of phenylalanine, tyrosine, and tryptophan: A casscf/caspt2 study. *ACS omega*, 9(33):35356–35363, 2024.

[40] Katarzyna Guzow, Mariusz Szabelski, Alicja Rzeska, Jerzy Karolczak, Hanna Sulowska, and Wiesław Wiczk. Photophysical properties of tyrosine at low ph range. *Chemical physics letters*, 362(5-6):519–526, 2002.

[41] Gernot Frenking. Perspective on "quantentheoretische beiträge zum benzolproblem. i. die elektronenkonfiguration des benzols und verwandter beziehungen" hückel e (1931) z phys 70: 204–286. *Theoretical Chemistry Accounts*, 103(3):187–189, 2000.

[42] Debasish Barman, Kavita Narang, Retwik Parui, Nehal Zehra, Mst Nasima Khatun, Laxmi Raman Adil, and Parameswar Krishnan Iyer. Review on recent trends and prospects in π-conjugated luminescent aggregates for biomedical applications. *Aggregate*, 3(5):e172, 2022.

[43] Paolo Coghi and Carmine Coluccini. Literature review on conjugated polymers as light-sensitive materials for photovoltaic and light-emitting devices in photonic biomaterial applications. *Polymers*, 16(10):1407, 2024.

[44] Amar BT Ghisaidoobe and Sang J Chung. Intrinsic tryptophan fluorescence in the detection and analysis of proteins: a focus on förster resonance energy transfer techniques. *International journal of molecular sciences*, 15(12):22518–22538, 2014.

[45] Fiona TS Chan, Gabriele S Kaminski Schierle, Janet R Kumita, Carlos W Bertoncini, Christopher M Dobson, and Clemens F Kaminski. Protein amyloids develop an intrinsic fluorescence signature during aggregation. *Analyst*, 138(7):2156–2162, 2013.

[46] Dorothea Pinotsi, Alexander K Buell, Christopher M Dobson, Gabriele S Kaminski Schierle, and Clemens F Kaminski. A label-free, quantitative assay of amyloid fibril growth based on intrinsic fluorescence. *ChemBioChem*, 14(7):846, 2013.

[47] Dorothea Pinotsi, Luca Grisanti, Pierre Mahou, Ralph Gebauer, Clemens F Kaminski, Ali Hassanali, and Gabriele S Kaminski Schierle. Proton transfer and structure-specific fluorescence in hydrogen bond-rich protein structures. *Journal of the American Chemical Society*, 138(9):3046–3057, 2016.

[48] Nicole Balasco, Carlo Diaferia, Elisabetta Rosa, Alessandra Monti, Menotti Ruvo, Nunzianna Doti, and Luigi Vitagliano. A comprehensive analysis of the intrinsic visible fluorescence emitted by peptide/protein amyloid-like assemblies. *International journal of molecular sciences*, 24(9):8372, 2023.

[49] Chyi Wei Chung, Amberley D Stephens, Edward Ward, Yuqing Feng, Molly Jo Davis, Clemens F Kaminski, and Gabriele S Kaminski Schierle. Label-free characterization of amyloids and alpha-synuclein polymorphs by exploiting their intrinsic fluorescence property. *Analytical Chemistry*, 94(13):5367–5374, 2022.

[50] Jonathan Pansieri, Véronique Josserand, Sun-Jae Lee, Anaëlle Rongier, Daniel Imbert, Marcelle Moulin Sallanon, Enikö Kövari, Thomas G Dane, Charlotte Vendrely, Odette Chaix-Pluchery, et al. Ultraviolet–visible–near-infrared optical properties of amyloid fibrils shed light on amyloidogenesis. *Nature photonics*, 13(7):473–479, 2019.

[51] Zohar A Arnon, Topaz Kreiser, Boris Yakimov, Noam Brown, Ruth Aizen, Shira Shaham-Niv, Pandeeswar Makam, Muhammad Nawaz Qaisrani, Emiliano Poli, Antonella Ruggiero, et al. On-off transition and ultrafast decay of amino acid luminescence driven by modulation of supramolecular packing. *Iscience*, 24(7), 2021.

[52] Xiaohong Chen, Weijian Luo, Huili Ma, Qian Peng, Wang Zhang Yuan, and Yongming Zhang. Prevalent intrinsic emission from nonaromatic amino acids and poly (amino acids). *Science China Chemistry*, 61(3):351–359, 2018.

[53] Marta Cadeddu, Davide Carboni, Luigi Stagi, Luca Malfatti, Maria F Casula, Francesca Caboi, and Plinio Innocenzi. Design of dual-emitting nonaromatic fluorescent polymers through thermal processing of l-glutamic acid and l-lysine. *Macromolecules*, 57(2):514–527, 2024.

[54] Debarshi Banerjee, Sonika Chibh, Om Shanker Tiwari, Gonzalo Díaz Mirón, Marta Monti, Hadar R Yakir, Shweta Pawar, Dror Fixler, Linda JW Shimon, Ehud Gazit, et al. Crystallization of l-cysteine in heavy water induces intrinsic fluorescence. *Angewandte Chemie*, 137(29):e202505331, 2025.

[55] Amberley D Stephens, Muhammad Nawaz Qaisrani, Michael T Ruggiero, Gonzalo Díaz Mirón, Uriel N Morzan, Mariano C González Lebrero, Saul TE Jones, Emiliano Poli, Andrew D Bond, Philippa J Woodhams, et al. Short hydrogen bonds enhance nonaromatic protein-related fluorescence. *Proceedings of the National Academy of Sciences*, 118(21):e2020389118, 2021.

[56] Anshuman Shukla, Sourav Mukherjee, Swati Sharma, Vishal Agrawal, KV Radha Kishan, and Purnananda Guptasarma. A novel uv laser-induced visible blue radiation from protein

crystals and aggregates: scattering artifacts or fluorescence transitions of peptide electrons delocalized through hydrogen bonding? *Archives of biochemistry and biophysics*, 428(2):144–153, 2004.

[57] Carmen González-González, Roi Lopez-Blanco, Juan A González-Vera, Sara D'Ingiullo, David Bouzada, Manuel Melle-Franco, Angel Orte, and M Eugenio Vázquez. Non-aromatic fluorescence from single α-helical peptides. *Cell Reports Physical Science*, 2025.

[58] Yang Yu, Soeun Gim, Dongyoon Kim, Zohar A Arnon, Ehud Gazit, Peter H Seeberger, and Martina Delbianco. Oligosaccharides self-assemble and show intrinsic optical properties. *Journal of the American Chemical Society*, 141(12):4833–4838, 2019.

[59] Jiayu Wu, Yuhuan Wang, Pan Jiang, Xiaolong Wang, Xin Jia, and Feng Zhou. Multiple hydrogen-bonding induced nonconventional red fluorescence emission in hydrogels. *Nature Communications*, 15(1):3482, 2024.

[60] Yangbin Xie, Zhen Wang, Mingcai Zhang, Xuan Wu, Ying Sun, Jincai Wu, Chun-Lin Sun, Baoxin Zhang, and Xiaobo Pan. Cluster-triggered excitation-dependent phosphorescence emission in polymorphic arylborinic acid. *Inorganic Chemistry Communications*, page 112895, 2024.

[61] Pai Liu, Weiqiang Fu, Peter Verwilst, Miae Won, Jinwoo Shin, Zhengxu Cai, Bin Tong, Jianbing Shi, Yuping Dong, and Jong Seung Kim. Mdm2-associated clusterization-triggered emission and apoptosis induction effectuated by a theranostic spiropolymer. *Angewandte Chemie*, 132(22):8513–8517, 2020.

[62] Roger Bresolí-Obach, José A Castro-Osma, Santi Nonell, Agustín Lara-Sánchez, and Cristina Martín. Polymers showing cluster triggered emission as potential materials in biophotonic applications. *Journal of Photochemistry and Photobiology C: Photochemistry Reviews*, page 100653, 2024.

[63] Gonzalo Díaz Mirón, Jonathan A Semelak, Luca Grisanti, Alex Rodriguez, Irene Conti, Martina Stella, Jayaramakrishnan Velusamy, Nicola Seriani, Nadja Došlić, Ivan Rivalta, et al. The carbonyl-lock mechanism underlying non-aromatic fluorescence in biological matter. *Nature Communications*, 14(1):7325, 2023.

[64] Scott A Hollingsworth and Ron O Dror. Molecular dynamics simulation for all. *Neuron*, 99(6):1129–1143, 2018.

[65] Martin Karplus and J Andrew McCammon. Molecular dynamics simulations of biomolecules. *Nature structural biology*, 9(9):646–652, 2002.

[66] Paul K Weiner and Peter A Kollman. Amber: Assisted model building with energy refinement. a general program for modeling molecules and their interactions. *Journal of computational chemistry*, 2(3):287–303, 1981.

[67] Chris Oostenbrink, Alessandra Villa, Alan E Mark, and Wilfred F Van Gunsteren. A biomolecular force field based on the free enthalpy of hydration and solvation: the gromos force-field parameter sets 53a5 and 53a6. *Journal of computational chemistry*, 25(13):1656–1676, 2004.

[68] Bernard R Brooks, Robert E Bruccoleri, Barry D Olafson, David J States, S a Swaminathan, and Martin Karplus. Charmm: a program for macromolecular energy, minimization, and dynamics calculations. *Journal of computational chemistry*, 4(2):187–217, 1983.

[69] Herman JC Berendsen, James PM Postma, Wilfred F van Gunsteren, and Jan Hermans. Interaction models for water in relation to protein hydration. In *Intermolecular forces: proceedings of the fourteenth Jerusalem symposium on quantum chemistry and biochemistry held in jerusalem, israel, april 13–16, 1981*, pages 331–342. Springer, 1981.

[70] Daniel J Price and Charles L Brooks III. A modified tip3p water potential for simulation with ewald summation. *The Journal of chemical physics*, 121(20):10096–10103, 2004.

[71] William L Jorgensen, Jayaraman Chandrasekhar, Jeffry D Madura, Roger W Impey, and Michael L Klein. Comparison of simple potential functions for simulating liquid water. *The Journal of chemical physics*, 79(2):926–935, 1983.

[72] Alex D MacKerell Jr, Donald Bashford, MLDR Bellott, Roland Leslie Dunbrack Jr, Jeffrey D Evanseck, Martin J Field, Stefan Fischer, Jiali Gao, H Guo, Sookhee Ha, et al. All-atom empirical potential for molecular modeling and dynamics studies of proteins. *The journal of physical chemistry B*, 102(18):3586–3616, 1998.

[73] Yue Shi, Zhen Xia, Jiajing Zhang, Robert Best, Chuanjie Wu, Jay W Ponder, and Pengyu Ren. Polarizable atomic multipole-based amoeba force field for proteins. *Journal of chemical theory and computation*, 9(9):4046–4063, 2013.

[74] Loup Verlet. Computer" experiments" on classical fluids. i. thermodynamical properties of lennard-jones molecules. *Physical review*, 159(1):98, 1967.

[75] William C Swope, Hans C Andersen, Peter H Berens, and Kent R Wilson. A computer simulation method for the calculation of equilibrium constants for the formation of physical clusters of molecules: Application to small water clusters. *The Journal of chemical physics*, 76(1):637–649, 1982.

[76] Nosé Shuichi. Constant temperature molecular dynamics methods. *Progress of Theoretical Physics Supplement*, 103:1–46, 1991.

[77] Giovanni Bussi, Davide Donadio, and Michele Parrinello. Canonical sampling through velocity rescaling. *The Journal of chemical physics*, 126(1), 2007.

[78] Michele Parrinello and Aneesur Rahman. Polymorphic transitions in single crystals: A new molecular dynamics method. *Journal of Applied physics*, 52(12):7182–7190, 1981.

[79] Adu Offei-Danso, Ali Hassanali, and Alex Rodriguez. High-dimensional fluctuations in liquid water: Combining chemical intuition with unsupervised learning. *Journal of Chemical Theory and Computation*, 18(5):3136–3150, 2022.

[80] Gareth A Tribello and Piero Gasparotto. Using dimensionality reduction to analyze protein trajectories. *Frontiers in molecular biosciences*, 6:46, 2019.

[81] Angel E García. Large-amplitude nonlinear motions in proteins. *Physical review letters*, 68(17):2696, 1992.

[82] Manel A Balsera, Willy Wriggers, Yoshitsugu Oono, and Klaus Schulten. Principal component analysis and long time protein dynamics. *The Journal of Physical Chemistry*, 100(7):2567–2572, 1996.

[83] Luigi Bonati, Enrico Trizio, Andrea Rizzi, and Michele Parrinello. A unified framework for machine learning collective variables for enhanced sampling simulations: mlcolvar. *The Journal of Chemical Physics*, 159(1), 2023.

[84] Aldo Glielmo, Brooke E Husic, Alex Rodriguez, Cecilia Clementi, Frank Noé, and Alessandro Laio. Unsupervised learning methods for molecular simulation data. *Chemical Reviews*, 121(16):9722–9758, 2021.

[85] Edward Danquah Donkor, Alessandro Laio, and Ali Hassanali. Do machine-learning atomic descriptors and order parameters tell the same story? the case of liquid water. *Journal of Chemical Theory and Computation*, 19(14):4596–4605, 2023.

[86] Jianqing Fan and Jinchi Lv. A selective overview of variable selection in high dimensional feature space. *Statistica Sinica*, 20(1):101, 2010.

[87] Marie-Claire Bellissent-Funel, Ali Hassanali, Martina Havenith, Richard Henchman, Peter Pohl, Fabio Sterpone, David Van Der Spoel, Yao Xu, and Angel E Garcia. Water determines the structure and dynamics of proteins. *Chemical reviews*, 116(13):7673–7697, 2016.

[88] Giacomo Fiorin, Michael L Klein, and Jérôme Hénin. Using collective variables to drive molecular dynamics simulations. *Molecular Physics*, 111(22-23):3345–3362, 2013.

[89] Robert B Best, Gerhard Hummer, and William A Eaton. Native contacts determine protein folding mechanisms in atomistic simulations. *Proceedings of the National Academy of Sciences*, 110(44):17874–17879, 2013.

[90] Giovanni Bussi and Alessandro Laio. Using metadynamics to explore complex free-energy landscapes. *Nature Reviews Physics*, 2(4):200–212, 2020.

[91] Albert P Bartók, Risi Kondor, and Gábor Csányi. On representing chemical environments. *Physical Review B—Condensed Matter and Materials Physics*, 87(18):184115, 2013.

[92] Yasuhiro Matsunaga, Yasuaki Komuro, Chigusa Kobayashi, Jaewoon Jung, Takaharu Mori, and Yuji Sugita. Dimensionality of collective variables for describing conformational changes of a multi-domain protein. *The Journal of Physical Chemistry Letters*, 7(8):1446–1451, 2016.

[93] Cristina Caruso, Annalisa Cardellini, Martina Crippa, Daniele Rapetti, and Giovanni M Pavan. Timesoap: Tracking high-dimensional fluctuations in complex molecular systems via time variations of soap spectra. *The Journal of Chemical Physics*, 158(21), 2023.

[94] Jörg Behler and Michele Parrinello. Generalized neural-network representation of high-dimensional potential-energy surfaces. *Physical review letters*, 98(14):146401, 2007.

[95] Linfeng Zhang, Jiequn Han, Han Wang, Wissam Saidi, Roberto Car, et al. End-to-end symmetry preserving inter-atomic potential energy model for finite and extended systems. *Advances in neural information processing systems*, 31, 2018.

[96] Ralf Drautz. Atomic cluster expansion for accurate and transferable interatomic potentials. *Physical Review B*, 99(1):014104, 2019.

[97] Debarshi Banerjee, Khatereh Azizi, Colin K Egan, Edward Danquah Donkor, Cesare Malosso, Solana Di Pino, Gonzalo Díaz Mirón, Martina Stella, Giulia Sormani, Germaine Neza Hozana, et al. Aqueous solution chemistry in silico and the role of data-driven approaches. *Chemical Physics Reviews*, 5(2), 2024.

[98] Giulia Sormani, Alex Rodriguez, and Ali Hassanali. Opportunities and challenges in unsupervised learning: The case of aqueous electrolyte solutions. *Journal of Chemical Theory and Computation*, 21(16):8060–8072, 2025.

[99] Sandip De, Albert P Bartók, Gábor Csányi, and Michele Ceriotti. Comparing molecules and solids across structural and alchemical space. *Physical Chemistry Chemical Physics*, 18(20):13754–13769, 2016.

[100] Dmitrii Maksimov, Carsten Baldauf, and Mariana Rossi. The conformational space of a flexible amino acid at metallic surfaces. *International Journal of Quantum Chemistry*, 121(3):e26369, 2021.

[101] Justus Johann Lange, Andrea Anelli, Jochem Alsenz, Martin Kuentz, Patrick J O'Dwyer, Wiebke Saal, Nicole Wyttenbach, and Brendan T Griffin. Comparative analysis of chemical descriptors by machine learning reveals atomistic insights into solute–lipid interactions. *Molecular Pharmaceutics*, 2024.

[102] Albert P Bartók, James Kermode, Noam Bernstein, and Gábor Csányi. Machine learning a general-purpose interatomic potential for silicon. *Physical Review X*, 8(4):041048, 2018.

[103] Aleks Reinhardt, Chris J Pickard, and Bingqing Cheng. Predicting the phase diagram of titanium dioxide with random search and pattern recognition. *Physical Chemistry Chemical Physics*, 22(22):12697–12705, 2020.

[104] Benjamin A Helfrecht, Rocio Semino, Giovanni Pireddu, Scott M Auerbach, and Michele Ceriotti. A new kind of atlas of zeolite building blocks. *The Journal of Chemical Physics*, 151(15), 2019.

[105] Solana Di Pino, Edward Danquah Donkor, Veronica M Sánchez, Alex Rodriguez, Giuseppe Cassone, Damian Scherlis, and Ali Hassanali. Zundeig: The structure of the proton in liquid water from unsupervised learning. *The Journal of Physical Chemistry B*, 127(45):9822–9832, 2023.

[106] Edward Danquah Donkor, Adu Offei-Danso, Alex Rodriguez, Francesco Sciortino, and Ali Hassanali. Beyond local structures in critical supercooled water through unsupervised learning. *The Journal of Physical Chemistry Letters*, 15(15):3996–4005, 2024.

[107] Riccardo Capelli, Francesco Muniz-Miranda, and Giovanni M Pavan. Ephemeral ice-like local environments in classical rigid models of liquid water. *The Journal of Chemical Physics*, 156(21), 2022.

[108] Bartomeu Monserrat, Jan Gerit Brandenburg, Edgar A Engel, and Bingqing Cheng. Liquid water contains the building blocks of diverse ice phases. *Nature communications*, 11(1):5757, 2020.

[109] Elizaveta Levina and Peter Bickel. Maximum likelihood estimation of intrinsic dimension. *Advances in neural information processing systems*, 17, 2004.

[110] Lucas Lacasa and Jesús Gómez-Gardenes. Correlation dimension of complex networks. *Physical review letters*, 110(16):168703, 2013.

[111] Iuri Macocco, Aldo Glielmo, Jacopo Grilli, and Alessandro Laio. Intrinsic dimension estimation for discrete metrics. *Physical Review Letters*, 130(6):067401, 2023.

[112] Ian Jolliffe. Principal component analysis. In *International encyclopedia of statistical science*, pages 1094–1096. Springer, 2011.

[113] Leland McInnes, John Healy, and James Melville. Umap: Uniform manifold approximation and projection for dimension reduction. *arXiv preprint arXiv:1802.03426*, 2018.

[114] Joshua B Tenenbaum, Vin de Silva, and John C Langford. A global geometric framework for nonlinear dimensionality reduction. *science*, 290(5500):2319–2323, 2000.

[115] Laurens van der Maaten and Geoffrey Hinton. Visualizing data using t-sne. *Journal of machine learning research*, 9(Nov):2579–2605, 2008.

[116] Elena Facco, Maria d'Errico, Alex Rodriguez, and Alessandro Laio. Estimating the intrinsic dimension of datasets by a minimal neighborhood information. *Scientific reports*, 7(1):12140, 2017.

[117] Absalom E Ezugwu, Abiodun M Ikotun, Olaide O Oyelade, Laith Abualigah, Jeffery O Agushaka, Christopher I Eke, and Andronicus A Akinyelu. A comprehensive survey of clustering algorithms: State-of-the-art machine learning applications, taxonomy, challenges, and future research prospects. *Engineering Applications of Artificial Intelligence*, 110:104743, 2022.

[118] Bernard W Silverman. *Density estimation for statistics and data analysis*. Routledge, 2018.

[119] YP Mack and Murray Rosenblatt. Multivariate k-nearest neighbor density estimates. *Journal of Multivariate Analysis*, 9(1):1–15, 1979.

[120] Alex Rodriguez, Maria d'Errico, Elena Facco, and Alessandro Laio. Computing the free energy without collective variables. *Journal of chemical theory and computation*, 14(3):1206–1215, 2018.

[121] Kenneth M Merz Jr. Using quantum mechanical approaches to study biological systems. *Accounts of Chemical Research*, 47(9):2804–2811, 2014.

[122] Laura E Ratcliff, Stephan Mohr, Georg Huhs, Thierry Deutsch, Michel Masella, and Luigi Genovese. Challenges in large scale quantum mechanical calculations. *Wiley Interdisciplinary Reviews: Computational Molecular Science*, 7(1):e1290, 2017.

[123] Markus Bursch, Jan-Michael Mewes, Andreas Hansen, and Stefan Grimme. Best-practice dft protocols for basic molecular computational chemistry. *Angewandte Chemie*, 134(42):e202205735, 2022.

[124] Soohaeng Yoo, Federico Zahariev, Sarom Sok, and Mark S Gordon. Solvent effects on optical properties of molecules: A combined time-dependent density functional theory/effective fragment potential approach. *The Journal of chemical physics*, 129(14), 2008.

[125] Cheng Cai, Weiqiang Tang, Chongzhi Qiao, Bo Bao, Peng Xie, Shuangliang Zhao, and Honglai Liu. A reaction density functional theory study of solvent effect in the nucleophilic addition reactions in aqueous solution. *Green Energy & Environment*, 7(4):782–791, 2022.

[126] Guillaume Jeanmairet, Maximilien Levesque, and Daniel Borgis. Tackling solvent effects by coupling electronic and molecular density functional theory. *Journal of Chemical Theory and Computation*, 16(11):7123–7134, 2020.

[127] Feliu Maseras and Keiji Morokuma. Imomm: A new integrated ab initio+ molecular mechanics geometry optimization scheme of equilibrium structures and transition states. *Journal of Computational Chemistry*, 16(9):1170–1179, 1995.

[128] Mats Svensson, Stephane Humbel, Robert DJ Froese, Toshiaki Matsubara, Stefan Sieber, and Keiji Morokuma. Oniom: a multilayered integrated mo+ mm method for geometry optimizations and single point energy predictions. a test for diels- alder reactions and pt (p (t-bu) 3) 2+ h2 oxidative addition. *The Journal of Physical Chemistry*, 100(50):19357–19363, 1996.

[129] Stefan Dapprich, István Komáromi, K Suzie Byun, Keiji Morokuma, and Michael J Frisch. A new oniom implementation in gaussian98. part i. the calculation of energies, gradients, vibrational frequencies and electric field derivatives. *Journal of Molecular Structure: THEOCHEM*, 461:1–21, 1999.

[130] Lung Wa Chung, WMC Sameera, Romain Ramozzi, Alister J Page, Miho Hatanaka, Galina P Petrova, Travis V Harris, Xin Li, Zhuofeng Ke, Fengyi Liu, et al. The oniom method and its applications. *Chemical reviews*, 115(12):5678–5796, 2015.

[131] Uriel N Morzan, Diego J Alonso de Armino, Nicolas O Foglia, Francisco Ramirez, Mariano C Gonzalez Lebrero, Damian A Scherlis, and Dario A Estrin. Spectroscopy in complex environments from qm–mm simulations. *Chemical reviews*, 118(7):4071–4113, 2018.

[132] Mattia Bondanza, Michele Nottoli, Lorenzo Cupellini, Filippo Lipparini, and Benedetta Mennucci. Polarizable embedding qm/mm: the future gold standard for complex (bio) systems? *Physical Chemistry Chemical Physics*, 22(26):14433–14448, 2020.

[133] Hans Martin Senn and Walter Thiel. Qm/mm methods for biomolecular systems. *Angewandte Chemie International Edition*, 48(7):1198–1229, 2009.

[134] Kittusamy Senthilkumar, Jon I Mujika, Kara E Ranaghan, Frederick R Manby, Adrian J Mulholland, and Jeremy N Harvey. Analysis of polarization in qm/mm modelling of biologically relevant hydrogen bonds. *Journal of The Royal Society Interface*, 5(suppl_3):207–216, 2008.

[135] Hai Lin, Yan Zhang, Soroosh Pezeshki, Bo Wang, Xin-Ping Wu, Laura Gagliardi, and Donald G Truhlar. Qmmm 2018. *University of Minnesota: Minneapolis, MN, USA*, 2018.

[136] Hai Lin and Donald G Truhlar. Redistributed charge and dipole schemes for combined quantum mechanical and molecular mechanical calculations. *The Journal of Physical Chemistry A*, 109(17):3991–4004, 2005.

[137] PH König, M Hoffmann, Th Frauenheim, and Q Cui. A critical evaluation of different qm/mm frontier treatments with scc-dftb as the qm method. *The Journal of Physical Chemistry B*, 109(18):9082–9095, 2005.

[138] Axel D Becke. Perspective: Fifty years of density-functional theory in chemical physics. *The Journal of chemical physics*, 140(18), 2014.

[139] Pierre Hohenberg and Walter Kohn. Inhomogeneous electron gas. *Physical review*, 136(3B):B864, 1964.

[140] Jörg Neugebauer and Tilmann Hickel. Density functional theory in materials science. *Wiley Interdisciplinary Reviews: Computational Molecular Science*, 3(5):438–448, 2013.

[141] Gabriel R Schleder, Antonio CM Padilha, Carlos Mera Acosta, Marcio Costa, and Adalberto Fazzio. From dft to machine learning: recent approaches to materials science–a review. *Journal of Physics: Materials*, 2(3):032001, 2019.

[142] Robert O Jones. Density functional theory: Its origins, rise to prominence, and future. *Reviews of modern physics*, 87(3):897–923, 2015.

[143] Jens K Nørskov, Frank Abild-Pedersen, Felix Studt, and Thomas Bligaard. Density functional theory in surface chemistry and catalysis. *Proceedings of the National Academy of Sciences*, 108(3):937–943, 2011.

[144] Valeria Butera. Density functional theory methods applied to homogeneous and heterogeneous catalysis: a short review and a practical user guide. *Physical Chemistry Chemical Physics*, 26(10):7950–7970, 2024.

[145] Konstantinos D Vogiatzis, Mikhail V Polynski, Justin K Kirkland, Jacob Townsend, Ali Hashemi, Chong Liu, and Evgeny A Pidko. Computational approach to molecular catalysis by 3d transition metals: challenges and opportunities. *Chemical reviews*, 119(4):2453–2523, 2018.

[146] Yaxiao Ma, Jing Xiong, Peng Zhang, Yuanfeng Li, Sicheng Zhang, Zhenpeng Wang, Linsheng Xu, Haoqi Guo, Kaixuan Chen, and Yuechang Wei. Recent progress on density functional theory calculations for catalytic control of air pollution. *ACS ES&T Engineering*, 4(1):47–65, 2023.

[147] Mark E Casida and Miquel Huix-Rotllant. Progress in time-dependent density-functional theory. *Annual review of physical chemistry*, 63(1):287–323, 2012.

[148] Claudio Garino and Luca Salassa. The photochemistry of transition metal complexes using density functional theory. *Philosophical Transactions of the Royal Society A: Mathematical, Physical and Engineering Sciences*, 371(1995):20120134, 2013.

[149] Michael J Gillan, Dario Alfe, and Angelos Michaelides. Perspective: How good is dft for water? *The Journal of chemical physics*, 144(13), 2016.

[150] J. C. Slater and G. F. Koster. Simplified lcao method for the periodic potential problem. *Phys. Rev.*, 94:1498–1524, Jun 1954.

[151] Ka Hung Lee, Udo Schnupf, Bobby G Sumpter, and Stephan Irle. Performance of density-functional tight-binding in comparison to ab initio and first-principles methods for isomer geometries and energies of glucose epimers in vacuo and solution. *ACS omega*, 3(12):16899–16915, 2018.

[152] Michael Gaus, Albrecht Goez, and Marcus Elstner. Parametrization and benchmark of dftb3 for organic molecules. *Journal of Chemical Theory and Computation*, 9(1):338–354, 2013.

[153] Gonzalo Díaz Mirón, Carlos R Lien-Medrano, Debarshi Banerjee, Uriel N Morzan, Michael A Sentef, Ralph Gebauer, and Ali Hassanali. Exploring the mechanisms behind non-aromatic fluorescence with the density functional tight binding method. *Journal of Chemical Theory and Computation*, 20(9):3864–3878, 2024.

[154] Fernand Spiegelman, Nathalie Tarrat, Jérôme Cuny, Leo Dontot, Evgeny Posenitskiy, Carles Martí, Aude Simon, and Mathias Rapacioli. Density-functional tight-binding: basic concepts and applications to molecules and clusters. *Advances in physics: X*, 5(1):1710252, 2020.

[155] Anshuman Kumar, Pablo R Arantes, Aakash Saha, Giulia Palermo, and Bryan M Wong. Gpu-enhanced dftb metadynamics for efficiently predicting free energies of biochemical systems. *Molecules*, 28(3):1277, 2023.

[156] Dirk Porezag, Th Frauenheim, Th Köhler, Gotthard Seifert, and R Kaschner. Construction of tight-binding-like potentials on the basis of density-functional theory: Application to carbon. *Physical Review B*, 51(19):12947, 1995.

[157] Marcus Elstner, Dirk Porezag, G Jungnickel, J Elsner, M Haugk, Th Frauenheim, Sandor Suhai, and Gotthard Seifert. Self-consistent-charge density-functional tight-binding method for simulations of complex materials properties. *Physical Review B*, 58(11):7260, 1998.

[158] Michael Gaus, Qiang Cui, and Marcus Elstner. Dftb3: Extension of the self-consistent-charge density-functional tight-binding method (scc-dftb). *Journal of chemical theory and computation*, 7(4):931–948, 2011.

[159] Van Quan Vuong, Jissy Akkarapattiakal Kuriappan, Maximilian Kubillus, Julian J Kranz, Thilo Mast, Thomas A Niehaus, Stephan Irle, and Marcus Elstner. Parametrization and benchmark of long-range corrected dftb2 for organic molecules. *Journal of Chemical Theory and Computation*, 14(1):115–125, 2018.

[160] Robert S Mulliken. Electronic population analysis on lcao–mo molecular wave functions. i. *The Journal of chemical physics*, 23(10):1833–1840, 1955.

[161] Thomas A Niehaus, S Suhai, F Della Sala, P Lugli, Marcus Elstner, Gotthard Seifert, and Th Frauenheim. Tight-binding approach to time-dependent density-functional response theory. *Physical Review B*, 63(8):085108, 2001.

[162] Thomas A Niehaus. Approximate time-dependent density functional theory. *Journal of Molecular Structure: THEOCHEM*, 914(1-3):38–49, 2009.

[163] Mark E Casida. Time-dependent density functional response theory for molecules. In *Recent Advances In Density Functional Methods: (Part I)*, pages 155–192. World Scientific, 1995.

[164] Mark E Casida. Time-dependent density-functional theory for molecules and molecular solids. *Journal of Molecular Structure: THEOCHEM*, 914(1-3):3–18, 2009.

[165] Monja Sokolov, Beatrix M Bold, Julian J Kranz, Sebastian Hofener, Thomas A Niehaus, and Marcus Elstner. Analytical time-dependent long-range corrected density functional tight binding (td-lc-dftb) gradients in dftb+: implementation and benchmark for excited-state geometries and transition energies. *Journal of Chemical Theory and Computation*, 17(4):2266–2282, 2021.

[166] Ove Christiansen, Henrik Koch, and Poul Jørgensen. The second-order approximate coupled cluster singles and doubles model cc2. *Chemical Physics Letters*, 243(5-6):409–418, 1995.

[167] Beatrix M Bold, Monja Sokolov, Sayan Maity, Marius Wanko, Philipp M Dohmen, Julian J Kranz, Ulrich Kleinekathöfer, Sebastian Höfener, and Marcus Elstner. Benchmark and performance of long-range corrected time-dependent density functional tight binding (lc-td-dftb) on rhodopsins and light-harvesting complexes. *Physical Chemistry Chemical Physics*, 22(19):10500–10518, 2020.

[168] Germaine Neza Hozana, Gonzalo Diaz Miron, and Ali Hassanali. Data-driven discovery of the origins of uv absorption in the alpha-3c protein. *The Journal of Physical Chemistry B*, 129(19):4728–4737, 2025.

[169] Cecilia Tommos, Kathleen G Valentine, Melissa C Martínez-Rivera, Li Liang, and Veronica R Moorman. Reversible phenol oxidation and reduction in the structurally well-defined 2-mercaptophenol-α3c protein. *Biochemistry*, 52(8):1409–1418, 2013.

[170] E Lindahl, MJ Abraham, B Hess, and D Van Der Spoel. Gromacs 2019 source code (2018). *URL https://doi. org/10.5281/zenodo*, 2424363, 2022.

[171] Lauri Himanen, Marc OJ Jäger, Eiaki V Morooka, Filippo Federici Canova, Yashasvi S Ranawat, David Z Gao, Patrick Rinke, and Adam S Foster. Dscribe: Library of descriptors for machine learning in materials science. *Computer Physics Communications*, 247:106949, 2020.

[172] Aldo Glielmo, Iuri Macocco, Diego Doimo, Matteo Carli, Claudio Zeni, Romina Wild, Maria d'Errico, Alex Rodriguez, and Alessandro Laio. Dadapy: Distance-based analysis of data-manifolds in python. *Patterns*, page 100589, 2022.

[173] Martin Ester, Hans-Peter Kriegel, Jörg Sander, Xiaowei Xu, et al. A density-based algorithm for discovering clusters in large spatial databases with noise. In *kdd*, volume 96, pages 226–231, 1996.

[174] Martin J Field, Paul A Bash, and Martin Karplus. A combined quantum mechanical and molecular mechanical potential for molecular dynamics simulations. *Journal of computational chemistry*, 11(6):700–733, 1990.

[175] Jaewoon Jung, Takaharu Mori, Chigusa Kobayashi, Yasuhiro Matsunaga, Takao Yoda, Michael Feig, and Yuji Sugita. Genesis: a hybrid-parallel and multi-scale molecular dynamics simulator with enhanced sampling algorithms for biomolecular and cellular simulations. *Wiley Interdisciplinary Reviews: Computational Molecular Science*, 5(4):310–323, 2015.

[176] Chigusa Kobayashi, Jaewoon Jung, Yasuhiro Matsunaga, Takaharu Mori, Tadashi Ando, Koichi Tamura, Motoshi Kamiya, and Yuji Sugita. Genesis 1.1: A hybrid-parallel molecular dynamics simulator with enhanced sampling algorithms on multiple computational platforms, 2017.

[177] Robert B Best, Xiao Zhu, Jihyun Shim, Pedro EM Lopes, Jeetain Mittal, Michael Feig, and Alexander D MacKerell Jr. Optimization of the additive charmm all-atom protein force field targeting improved sampling of the backbone ϕ, ψ and side-chain $\chi1$ and $\chi2$ dihedral angles. *Journal of chemical theory and computation*, 8(9):3257–3273, 2012.

[178] Chien-Cheng Huang, A Chatterji, Godehard Sutmann, Gerhard Gompper, and Roland G Winkler. Cell-level canonical sampling by velocity scaling for multiparticle collision dynamics simulations. *Journal of computational physics*, 229(1):168–177, 2010.

[179] Ben Hourahine, Bálint Aradi, Volker Blum, Frank Bonafe, Alex Buccheri, Cristopher Camacho, Caterina Cevallos, MY Deshaye, T Dumitrică, A Dominguez, et al. Dftb+, a software package for efficient approximate density functional theory based atomistic simulations. *The Journal of chemical physics*, 152(12), 2020.

[180] U Ryde. Qm/mm calculations on proteins. *Methods in enzymology*, 577:119–158, 2016.

[181] Michele Ceriotti, Joshua More, and David E Manolopoulos. i-pi: A python interface for ab initio path integral molecular dynamics simulations. *Computer Physics Communications*, 185(3):1019–1026, 2014.

[182] Venkat Kapil, Mariana Rossi, Ondrej Marsalek, Riccardo Petraglia, Yair Litman, Thomas Spura, Bingqing Cheng, Alice Cuzzocrea, Robert H Meißner, David M Wilkins, et al. i-pi 2.0: A universal force engine for advanced molecular simulations. *Computer Physics Communications*, 236:214–223, 2019.

[183] Michele Ceriotti and David E Manolopoulos. Efficient first-principles calculation of the quantum kinetic energy and momentum distribution of nuclei. *Physical review letters*, 109(10):100604, 2012.

[184] Yu Kay Law and Ali A Hassanali. Role of quantum vibrations on the structural, electronic, and optical properties of 9-methylguanine. *The Journal of Physical Chemistry A*, 119(44):10816–10827, 2015.

[185] Takeshi Yanai, David P Tew, and Nicholas C Handy. A new hybrid exchange–correlation functional using the coulomb-attenuating method (cam-b3lyp). *Chemical physics letters*, 393(1-3):51–57, 2004.

[186] Maja Parac, Markus Doerr, Christel M Marian, and Walter Thiel. Qm/mm calculation of solvent effects on absorption spectra of guanine. *Journal of computational chemistry*, 31(1):90–106, 2010.

[187] Akira Nakayama, Gaku Arai, Shohei Yamazaki, and Tetsuya Taketsugu. Solvent effects on the ultrafast nonradiative deactivation mechanisms of thymine in aqueous solution: excited-state qm/mm molecular dynamics simulations. *The Journal of Chemical Physics*, 139(21), 2013.

[188] Roberto Improta, Fabrizio Santoro, and Lluis Blancafort. Quantum mechanical studies on the photophysics and the photochemistry of nucleic acids and nucleobases. *Chemical reviews*, 116(6):3540–3593, 2016.

[189] Subrahmanyam Sappati, Ali Hassanali, Ralph Gebauer, and Prasenjit Ghosh. Nuclear quantum effects in a hiv/cancer inhibitor: The case of ellipticine. *The Journal of Chemical Physics*, 145(20), 2016.

[190] Thomas E Markland and Michele Ceriotti. Nuclear quantum effects enter the mainstream. *Nature Reviews Chemistry*, 2(3):0109, 2018.

[191] Michele Ceriotti, Jérôme Cuny, Michele Parrinello, and David E Manolopoulos. Nuclear quantum effects and hydrogen bond fluctuations in water. *Proceedings of the National Academy of Sciences*, 110(39):15591–15596, 2013.

[192] Michele Ceriotti, Wei Fang, Peter G Kusalik, Ross H McKenzie, Angelos Michaelides, Miguel A Morales, and Thomas E Markland. Nuclear quantum effects in water and aqueous systems: Experiment, theory, and current challenges. *Chemical reviews*, 116(13):7529–7550, 2016.

[193] Fabrizio Santoro, James A Green, Lara Martinez-Fernandez, Javier Cerezo, and Roberto Improta. Quantum and semiclassical dynamical studies of nonadiabatic processes in solution: achievements and perspectives. *Physical Chemistry Chemical Physics*, 23(14):8181–8199, 2021.

[194] YK Law and AA Hassanali. The importance of nuclear quantum effects in spectral line broadening of optical spectra and electrostatic properties in aromatic chromophores. *The Journal of Chemical Physics*, 148(10), 2018.

[195] Margaret L Berrens, Arpan Kundu, Marcos F Calegari Andrade, Tuan Anh Pham, Giulia Galli, and Davide Donadio. Nuclear quantum effects on the electronic structure of water and ice. *The Journal of Physical Chemistry Letters*, 15(26):6818–6825, 2024.

[196] Francesco Ambrosio, Giacomo Miceli, and Alfredo Pasquarello. Structural, dynamical, and electronic properties of liquid water: A hybrid functional study. *The Journal of Physical Chemistry B*, 120(30):7456–7470, 2016.

[197] Unmesh Mondal, Ivan Girotto, Ali Hassanali, and Prasenjit Ghosh. Effect of quantum delocalization on temperature dependent double proton transfer in molecular crystals of terephthalic acid. *The Journal of Physical Chemistry B*, 127(23):5263–5272, 2023.

[198] László Turi, Bence Baranyi, and Ádám Madarász. 2-in-1 phase space sampling for calculating the absorption spectrum of the hydrated electron. *Journal of Chemical Theory and Computation*, 20(10):4265–4277, 2024.

[199] Gopalasamudram Narayana Ramachandran. Stereochemistry of polypeptide chain configurations. *J. Mol. Biol.*, 7:95–99, 1963.

[200] GN T Ramachandran and V Sasisekharan. Conformation of polypeptides and proteins. *Advances in protein chemistry*, 23:283–437, 1968.