Supervisors

Prof. Dr. Carles Curutchet

1. *Department de Farmacia i Tecnologia Farmaceutica, i Fisicoquímica,Facultat de farmácia i Ciències de Alimentaci, Universitat de Barcelona (UB)*
2. *Institut de Química teórica i computational (IQTUB), Universitat de Barcelona(UB)*

Supervisors

Dr. Maximilian. Menger

Prof. Dr. Shirin Faraji

Zernike Institute for Advanced Materials, University of Groningen

Master Studies in
Theoretical Chemistry and
Computational Modelling

Master Thesis

Benchmark: quantum-augmented excited state force-fields for protein fluorescence

Benchmark: Campos de fuerza de estado excitado complementados cuánticamente aplicados a la fluorescencia de proteínas

Benchmark: Camps de força d'estat excitat complementats quànticament aplicats a la fluorescència de proteïnes

Mateus Zanotto

June 2022

UNIVERSITAT DE BARCELONA

SUMMARY

Fluorophores embedded in complex systems are a challenge to study computationally as they demand an accurate description of the environment, increasing drastically its simulation cost. Using two different frameworks to calculate transition energies of protein fluorescence, we studied the viability of simulating such systems with a classical MM method with force fields augmented for the excited state by Q-Force. We compared it with the more accurate QM/MM method. The ground state FFs have shown excitation energies in agreement with the QM/MM procedure and the experimental data, with the highest difference of 0.15eV. The excite state FFs reproduce the geometries accurately, but it systematically overestimates the calculated de-excitation. This can be due to the excited state charges, being necessary to improve the classical MM framework regarding the excited state charges assignment.

Los fluoróforos incrustados en sistemas complejos son un desafío para estudiar computacionalmente, ya que exigen una descripción precisa del entorno, lo que aumenta drásticamente su costo de simulación. Usando dos marcos diferentes para calcular las energías de transición de la fluorescencia de proteínas, estudiamos la viabilidad de simular dichos sistemas con un método MM clásico con campos de fuerza aumentados para el estado excitado por Q-Force. Lo comparamos con el método QM/MM más preciso. Los FF de estado fundamental han mostrado energías de excitación de acuerdo con el procedimiento QM/MM y los datos experimentales, con la mayor diferencia de 0,15 eV. Los FF del estado de excitación reproducen las geometrías con precisión, pero sobreestiman sistemáticamente la desexcitación calculada. Esto puede deberse a las cargas del estado excitado, siendo necesario mejorar el marco clásico de MM con respecto a la asignación de cargas del estado excitado.

Els fluoròfors incrustats en sistemes complexos són un desafiament per estudiar computacionalment, ja que exigeixen una descripció precisa de l'entorn, cosa que augmenta dràsticament el cost de simulació. Usant dos marcs diferents per calcular les energies de transició de la fluorescència de proteïnes, estudiem la viabilitat de simular aquests sistemes amb un mètode MM clàssic amb camps de força augmentats per a l'estat excitat per Q-Force. Ho comparem amb el mètode QM/MM més precís. Els FF d'estat fonamental han mostrat energies d'excitació d'acord amb el procediment QM/MM i les dades experimentals, amb la diferència més gran de 0,15 eV. Els FF de l'estat d'excitació reprodueixen amb precisió les geometries, però sobreestimen sistemàticament la desexcitació calculada. Això pot ser degut a les càrregues de l'estat excitat, i cal millorar el marc clàssic de MM respecte a l'assignació de càrregues de l'estat excitat.

IDENTIFICATION AND REFLECTION ON THE SUSTAINABLE DEVELOPMENT GOALS

Intrinsic protein fluorescence has many applications on the monitoring of protein dynamics and bioimaging, providing powerful information about biological systems and proteins conformational transition, binding sites, denaturation, and general dynamics. This is due to the high sensitivity of fluorescent amino acids to its local environment, where small alterations can cause changes in the emission spectra. Nevertheless, this complexity is also a challenge for its study and interpretation of data.

When we apply this field to the Sustainable Development Goals, the area of this research is People, specifically the 3rd goal: Good Health and Well-Being. Investigating a reliable methodology to study those systems computationally can guide experimentalists to interpret the emission spectra in biological system.

The target 3.4 is to reduce mortality from non-communicable diseases and promote mental health. Improving the monitoring of fluorescent proteins can lead to the better understanding of its dynamics such as binding sites for pharmaceuticals, improving drug delivery and consequently non-communicable diseases treatments.

The scope of this project is to perform two different computational approaches for studying embedded fluorophores, the more accurate, but computationally expensive hybrid quantum-mechanics/molecular-mechanics method and the less demanding classical molecular mechanics with a quantum augmented force field. Finding a robust methodology, it's an important step on the pursue of a framework that is both computationally efficient and chemically accurate.

REPORT

Benchmark: Quantum-augmented excited state force-field for protein fluorescence

Author: Mateus Zanotto
Supervisors: Dr. Maximilian F. S. J. Menger[a], Prof. Dr. Shirin Faraji[a] and Prof. Dr. Carles Curutchet[b].
[a]*Zernike Institure for Advanced Materials, University of Groningen, Nijenborgh 4, 9747 AG Groningen, The Netherlands.*
[b]*Departament de Farmàcia i Tecnologia Farmacèutica, i Fisicoquímica Facultat de Farmàcia i Ciències de l'Alimentació, Universitat de Barcelona, Av. Joan XXIII s/n, 08028 Barcelona, Spain.*

ABSTRACT: Fluorophores embedded in complex systems are a challenge to study computationally as they demand an accurate description of the environment, increasing drastically its simulation cost. Using two different frameworks to calculate transition energies of protein fluorescence, we studied the viability of simulating such systems with a classical MM method with force fields augmented for the excited state by Q-Force. We compared it with the more accurate QM/MM method. The ground state FFs have shown excitation energies in agreement with the QM/MM procedure and the experimental data, with the highest difference of 0.15eV. The excite state FFs reproduce the geometries accurately, but it systematically overestimates the calculated de-excitation. This can be due to the excited state charges, being necessary to improve the classical MM framework regarding the excited state charges assignment.

1. INTRODUCTION

Shimomura's contributions to the purification and characterization of the Green Fluorescent Protein (GFP) through spectroscopy were essential to the development of the today's widely used GFP-like proteins as markers for protein localization and gene expression monitoring in living cells.[1]

For the discovery and development of the GFP, which can attach to other proteins and mark it with fluorescence, the Nobel Prize in Chemistry of 2008 was awarded jointly to Osamu Shimomura[1,2], Martin Chalfie[3], and Roger Y. Tsien[4,5].

The method that allows monitoring individual fluorescent molecules inside biological systems is the super-resolved fluorescence microscopy. The 2014's Nobel Prize in Chemistry awarded Eric Betzig[6,7], Stefan Hell[8,9], and William E. Moerner[10] for the development of this technique.

This highlights how biophysical fluorescence methods are highly used to monitor and elucidate intracellular structures, dynamics, and interactions from fluorescent signals.[11,12]

When dealing with naturally occurring protein fluorescence, the main three amino acids responsible to its fluorescence are the tryptophan (Trp), tyrosine and phenylalanine.[12] The occurrence of those fluorescent amino acids is quite rare, and between those, the tryptophan dominates most of the fluorescence spectrum of proteins, even when it has more than one fluorescent amino acid.[13]

The tryptophan is much more sensitive to changes in its local environment than the tyrosine and phenylalanine.[14] Because of this complexity it's a challenge to study and interpret its emission, and most scientific investigations focus on the tryptophan residue.

Depending on its accessibility to the solvent, and therefore the polarity of the local environment, the tryptophan can emit from 308nm for completely buried apolar core to 350nm to fully exposed to the solvent. Studying the fluorescence of these bodies can reveal a great variety of information on its structure, its exposure to the solvent, even intramolecular distances, and orientation inside the protein between donor and acceptor moieties based on fluorescence resonance energy transfer (FRET). [5]

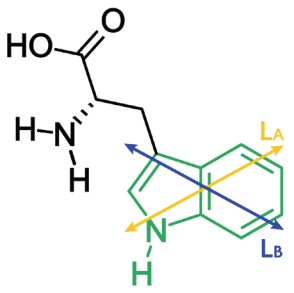

Figure 1. *Tryptophan residue with the fluorescent indole moiety highlighted in green. The transition dipole moment of the L_A and L_B state are represented by the yellow and blue arrows, respectively.*

The tryptophan has two main excited states (E.S.), the L_A and the L_B states, which are categorized based on its transition dipole moment. The L_A state has a strong transition dipole moment and is stabilized by polar environments, being the most responsible for the protein's fluorescence. The L_B state has a lower oscillator strength and less polar, being most stable in apolar environments.[15]

Computationally, the prediction of the fluorescence is highly dependent on how the ground state (G.S.) and the L_A excited state are described. These fluorophores in embedded systems have its charges and geometry strongly affected by the electrostatic field around it.[16]

Because of this deep dependence on the local environment, first-principle simulations must consider not only the electronic structure, but also the surroundings with fidelity.

Therefore, the most common approach to simulate the excited state of the tryptophan inside a protein is the hybrid Quantum-Mechanics / Molecular-Mechanics (QM/ MM) method which combines classical MM with any QM method, manly DFT.[17]

On this investigation we will perform the standard QM/MM and compare its results with a framework that uses QM only to optimize the charges and the molecule-specific force-fields (FFs) parameters with Q-Force [18], a toolkit with protocols to generate quantum mechanically augmented molecular FFs.

This study seeks to bring light on how viable it is to study these complex systems such as protein fluorescence with this computationally less expensive methodology, and what are its limitations.

2. THEORY AND COMPUTATIONAL DETAILS

2.1. Overview of the simulation procedures

The protein of choice for this study is the T4-Lysozyme shown in Figure 2. It has three tryptophans with different exposures to the solvent. This system allows to study the difference in emission due to tryptophans with different access to the solvent in the same protein.

The standard procedure to study an embedded fluorophores is to use a QM/MM framework. This approach is reliable and offers a rich treatment for the environment, but the large computational cost is its main drawback.[19]

A different and cheaper approach by Vivian and Callis[20] to study the tryptophan fluorescence is to use a hybrid procedure that involve QM only for assigning the charges to the indole ring highlighted on Figure 1, the moiety responsible for tryptophan fluorescence. A MD step using a general force field with the assigned charges is used to compute the electrostatic potentials and fields.

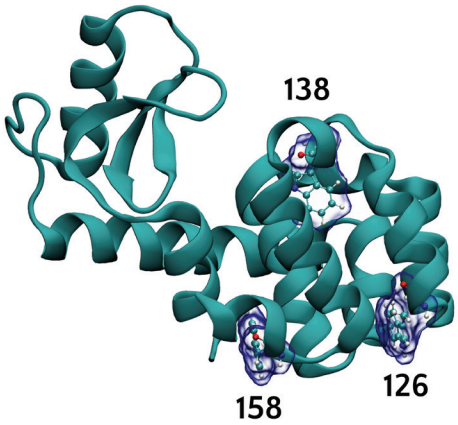

Figure 2. *The T4-Lysozyme (1LYD) has two tryptophans (ResIDs: 126 and 158) more exposed to the solvent and one (ResID: 138) buried inside the protein.*

The transition energies are then simulated by switching the tryptophan geometry by a reference geometry. The reference geometry is obtained from a crystal structure for the ground state and an ab-initio calculation is used for the L_A state geometry.

The strategy for this investigation is to perform the standard framework with QM/MM and a similar hybrid framework, as is follows and it schematized on Figure 3:

(1) The protein is simulated through a classical molecular dynamic (MD) for 1 ms to sample the protein's structure. Using the second half of the MD, we take 10 snapshots of the trajectory for the next step.

(2) To ensure the equilibration of the system on the ground state for each snapshot, we performed a 5ps QM/MM BOMD with 0.5fs timestep, one for each tryptophan residue in the QM region.

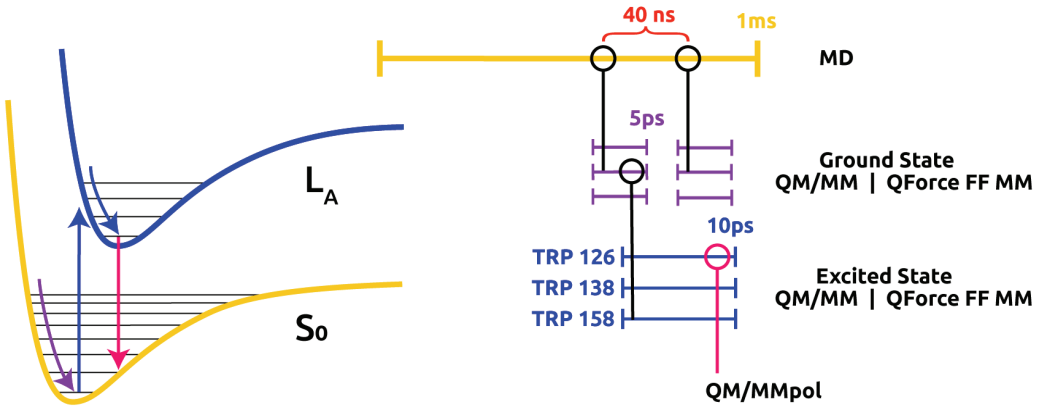

Figure 3. *MD Simulations methods scheme. (1) The protein is first sampled with a 1ms classical MD, then (2) a step is performed to relax the G.S. further. (3) The system is excited to the first excited state, relaxed and sampled. (4) Lastly the absorption and emission are calculated with QM/MMPol.*

(3) The trajectories are continued with a 10ps QM/MM BOMD with the same timestep in order to electronic excite the tryptophans, relax the surroundings and sampled 100 structures of the indole in the L_A state for the second half of the QM/MM. This step will be referred as QM sampling.

For the framework using Q-Force FFs, the GS QM / MM (step 2) and ES QM/MM (step 3) are substituted by a classical molecular mechanic with the indole moiety being described by one augmented Q-Force FF for the GS and one for the ES. This step will be referred as MM sampling.

(4) Finally, we calculate the absorption and the emission of each tryptophans in the protein, for both the QM/MM sampling and the MM sampling with Q-Force FF, for the GS and ES geometries using the polarizable QM/MM method (QM / MMPol) [21].

For each of the 10 snapshots from the step (1), there was sampled 100 structures on the GS (step 2) and on the ES (step 3), totalizing 1000 conformations to be post-processed by QM/MMPol for each transition reported.

2.2. Classical Simulations

The seed structure for the studied protein T4-Lysozyme (1LYD, resolution 2.00Å) used in this work was obtained from the RCSB Protein Data Bank (PDB).[22,23] The optimal pH for 1LYD was calculated using the H++ web server, which added the hydrogens accordingly to the standard protonation for most amino acids.[24] Only the protonation state of the histidine residue (HIS) was replaced by the HIP residue index, which corresponds to the histidine protonated on both N-epsilon and N-delta. The PDB file was then edited to remove all crystallographic waters.

The modified PDB file was solvated by adding the protein to a box containing ~28000 OPC water molecules.[25] Nine chloride ions were included in order to balance the charges and neutralize the system. The 1LYD protein was described with the ff19SB Force Field.[26] The cut-off radius for non-bonding interactions was set to 8Å.

The energy minimization was done in two steps, the first minimized only the solvent and the ions, maintaining the protein frozen, and then the second minimized the whole system. This step is important to eliminate any high energy structure from the system, which can lead to problems in the heating and production steps.

The heating was performed in two steps, the first increased the temperature from 0K to 300K in the NVT ensemble through 250ps followed by the second heating step with 250ps in 300K using the NPT ensemble, all with 2fs timestep.

Following, a production run for the protein for a period of 1 ms with classical molecular dynamic (MD) using a topology with Hydrogen Mass Repartitioning (HMR).

This approximation redistributes some of the heavy atom's mass to the hydrogen connected to them, allowing it to use larger time steps as the hydrogen will be less flexible. It also diminishes the instability related to high-frequency hydrogen motion. With this, the timestep was increased to 4fs. All steps were done with SHAKE. All MD runs were performed with the Amber 20 suite of program[27].

From the production step we took 10 snapshots, from 640ns to 1000ns, each 40ns.

2.3. QM/MM and QM/MMPol Simulations

The multiscale QM/MM method have different approaches based on how it can be implemented. The Subtractive scheme is done by first calculating the MM energy for the entire system (S) and then the QM energy for only the inner subsystem with the link atom (I+L). The total energy is then calculated by subtracting the MM energy from the inner subsystem.

$$E_{QM/MM} = E_{MM}(S) + E_{QM}(I+L) - E_{MM}(I+L)$$

In this scheme, the coupling of the subsystem is done entirely at the MM level of theory. This approach doesn't describe the Coulombic interactions between the atomic charges in the QM and the MM, making the subtractive scheme troublesome for the electrostatic interactions.

The improved subtractive ONIOM method allows the system to be subdivided into a *n*-layered subsystem, which can be treated with QM or MM deliberately. In this scheme the MM charges can be included into the QM Hamiltonian, allowing electrostatic embedding.

For the additive QM/MM scheme, the energy is obtained with the following equation:

$$E_{QM/MM} = E_{MM}(O) + E_{QM}(I+L) + E_{QM\text{--}MM}(I\text{--}O)$$

Where the MM energy is calculated for the outer (O) region and an additional term describes explicitly the coupling between the inner QM and the outer MM systems. How the coupling term is written is what defines a QM/MM method.

The steps (2) and (3) with QM/MM were done with the electrostatic embedding.[28] This scheme incorporates the MM point charges into the QM Hamiltonian, which enables the inner system to adapt to the rigid charges of the MM region and be polarized by it.

The QM region was treated at the DFT level for the ground state and at TD-DFT level for the excited state, both with PBE0 and 6-31G(d) basis set, using the Gaussian 16 program.[29] To inspect the separation between the indole's L_A and L_B excited states, a TD-DFT calculation in the indole ring with a an C_α in the position 3 (3-methyl indole, 3MI) has been made. We found a separation of 0.8eV between the first L_A state and the L_B state for this level of theory.

The next complexity step regarding QM/MM is to have the QM electric field acting back into the MM charges. This approach is called polarizable embedding, which have two main approaches. The first where the QM acts into the MM charges, but the MM region doesn't act back into the QM region. And the second where the QM Hamiltonian incorporates a MM polarizable self-energy term, and the effective Hamiltonian can be calculated self-consistently.

The QM/MMPol[30] method used for calculating the transition energies is the later model, where full mutual polarization effects are accounted. The effective Hamiltonian for this model is the following:

$$\hat{\boldsymbol{H}}_{\boldsymbol{eff}} = \hat{\boldsymbol{H}}_{\boldsymbol{0}}(\boldsymbol{I}) + \hat{\boldsymbol{H}}^{\boldsymbol{el}}_{\boldsymbol{QM/MM}}(\boldsymbol{I}\text{--}\boldsymbol{O}) + \hat{\boldsymbol{H}}^{\boldsymbol{pol}}_{\boldsymbol{QM/MM}}(\boldsymbol{I}\text{--}\boldsymbol{O}) + \hat{\boldsymbol{H}}^{\boldsymbol{el}}_{\boldsymbol{MM}}(\boldsymbol{O}) + \hat{\boldsymbol{H}}^{\boldsymbol{pol}}_{\boldsymbol{MM}}(\boldsymbol{O})$$

In which the $\hat{H}_0$ is the Hamiltonian for the isolated QM system, $\hat{H}^{el}_{QM/MM}$ and $\hat{H}^{pol}_{QM/MM}$ are the electrostatic and polarizable Hamiltonians for the QM/MM coupling energy terms. Finally, the $\hat{H}^{el}_{MM}$ and $\hat{H}^{pol}_{MM}$ are the electronic self-energy for the charges and the polarization energy terms for the MM region. In this work the cut-off radius of the polarized MM is 15Å around the QM region.

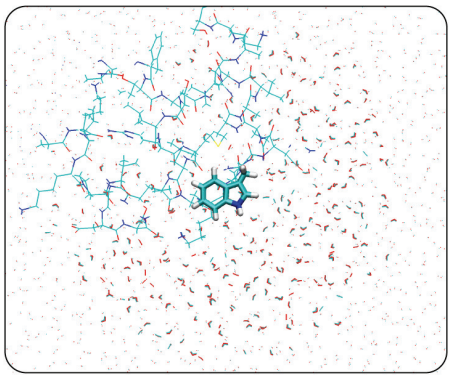

Figure 4. *Trp126 inputs for QM/MMPol. The regions with the indole in the QM region represented as cylinders (licorice), the environment with MM polarizability effects around (15Å around the QM region) as lines and the MM without polarizability (30Å around the QM region) as dots.*

When dealing with embedded chromophores, the environment strongly affects its charges and relaxed structure. The electrostatic effects are the main contribution from the environment, but the polarizability can't be neglected for systems with strong transition dipole moment between the G.S. and the E.S such as the tryptophans.

With this, we use the linear response of the time-dependent density functional theory (TD-DFT) to perform the single point transition energy calculations. All the calculations were performed with the same level of theory as the others. For the QM/MMPol we used a locally modified development version of the Gaussian package[29].

The vertical excitation and de-excitation energies are obtained with the Linear Response (LR) formalist and a State Specific (SS) correction. First the excitation energy is obtained by electronic exciting the G.S. to an E.S. with the solvent polarization frozen (ω^0). Then, with LR formulation, the electronic response (R) of the environment is calculated in respect to the transition density (ρ^T).

The SS correction is obtained by making the electrostatic potential that arises from the excited state density self-consistent with the environment.

All reported excitation energies are solely the vertical transition energy (ω^0+R(ρ^T)), while the de-excitation energies also include the SS corrections.

2.4. Force Field Parameters

Q-Force[18] is a toolkit to augment transferable Force Fields with specific parameters derived from QM calculations. The nonbonded parameters are retained from the transferable FFs (e.g., OPLS, AMBER, GROMOS, CHARMM), which are rigorously tested in respect to thermodynamics properties.

This approach allows the augmented FF for nonstandard molecules to be combined with force fields that already have been carefully parametrized for systems like proteins, that demand a rigorous description for the backbone for example.

In order to optimize the FF force constants, equilibrium distances, angles and dihedrals, Q-Force uses the optimized geometry, the second derivative matrix (Hessian matrix), and relaxed dihedrals scans from a QM calculation.

First Q-Force determines the rigid terms and obtain the force constants (k) using a linear least-square fitting in the Hessian Matrix. For the bond and angle terms, it fits to the harmonic potential.

The dihedrals are fitted using three different functions. For the rigid, there are the proper dihedral that describes the interaction between atoms *ijk* and *jkl* with atoms *i* and *l* in different planes; and the improper dihedrals, that describes the planar groups. Finally, the flexible dihedrals that are the ones with multiple minima.

The flexible dihedrals are scanned with a QM calculation by fixing the flexible dihedral at each interval and optimizing the rest of the geometry. Q-Force creates these scan inputs automatically.

The differences in each function for the potentials are further discussed on section 3.1.

The Q-Force FF was generated for the indole rings, as this is the moiety responsible for the fluorescence. The backbone was maintained being described by the ff19SB. The nonbonded parameters were taken from the general Amber force-field (GAFF2).

The geometry optimization and frequency calculations for the indole were obtained at the DFT level of theory with PBE0 functional and 6-31G(d) basis set. The indole ring was solvated with water using a Polarizable Continuum Model (PCM) solvation model with the Integral Equation Formalism variation (IEFPCM) to better reproduce the geometry in the excited state, which is stabilized by the electrostatic interactions with the solvent. For the excited state FF, the geometry was optimized to the first E.S. (root=1). The RESP charges for both G.S. and E.S. FFs for the indole ring were calculated at the same level of theory. All calculations were performed using Gaussian 16 program.[29]

The timesteps and total simulation time were also kept the same. The information about how the Q-Force FF parameters were obtained are discussed below.

3. RESULTS AND DISCUSSION

3.1. Q-Force developments

The available Q-Force version generated force fields in GROMACS format. Part of this work was to add support to AMBER format for Q-Force. Converting AMBER format to GROMACS format can be done easily and there is several software, such as ParmEd, that can achieve this. The conversion from GROMACS to AMBER is more complex thought, as the improper dihedrals on AMBER are described with the same cosine function as the other dihedrals, while on GROMACS format the improper dihedrals are described by the harmonic potential:

The parameters used in Amber Force Fields are given by the following Hamiltonian:

$$\begin{aligned} E_{total} &= \sum_{bonds} k_r(r-r_0)^2 \\ &+ \sum_{angles} k_\theta(\theta-\theta_0)^2 \\ &+ \sum_{dihedrals} V_n[1+cos(n\phi-\gamma)] \\ &+ \sum_{i=1}^{N-1}\sum_{j>1}^{N}\left[\frac{A_{ij}}{R_{ij}^{12}}-\frac{B_{ij}}{R_{ij}^{6}}+\frac{q_iq_j}{\epsilon R_{ij}}\right] \end{aligned}$$

The bond and angles parameters are described as a harmonic potential for both software and converting them can be done by only changing the units from kJ mol-1 to kcal mol-1.

For the nonbonded interactions, the parameters are retained from transferable FFs (GAFF2 in this case) as these parameters are necessary to better describe the interaction between the Q-Force FF and the transferable FF for the rest of the system.

For the rigid dihedrals, which are the proper and improper ones, GROMACS uses the harmonic potential following for these terms:

$$V_n = k_\phi(\phi-\phi_0)^2$$

The improper dihedral is the term that describes the 4-body interactions when there is a central atom or a cross interaction, as shown below.

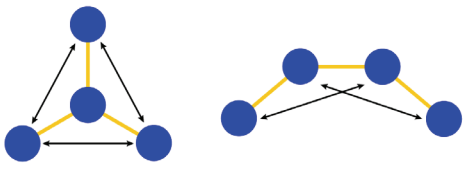

Figure 5. *Improper dihedral interactions constrain the conjugated rings and double bonds. This are the interactions responsible for keeping molecules planar in a force field.*

In order to make these dihedrals compatible with Amber format, which is a Fourier expansion, the parameters were transferred through the following relationship:

$$V_n = \frac{2k_\phi}{n^2}$$

This approximation reproduces the minima of the harmonic potential from GROMACS as a cosine function in AMBER and maintains the improper dihedrals planar during the classical simulation.

The flexible dihedrals are the ones with multiple minima, and in GROMACS they are described as the Ryckaert-Ballemans[31] dihedral potential:

$$V_{rb} = \sum_{n=0}^{5} C_n\big(cos(\phi-\pi)\big)^n$$

Which can be analytically described by both software, AMBER and GROMACS, as the four terms Fourier series:

$$\begin{aligned} V_F = \frac{1}{2}\,[&C_1(1+cos(\phi)\ +\ C_2(1-cos(2\phi)\ + \\ &C_3(1+cos(3\phi)\ +\ C_4(1-cos(4\phi)] \end{aligned}$$

At the time of publication of this study, Q-Force toolkit with support to AMBER is still being built to be more intuitive for the end user. Q-Force is freely available on GitHub (https://github.com/selimsami/qforce) with various examples and tutorials.

3.2. Protein structure analysis

The first step (1) of this work was to sample the 1LYD protein for 1ms with a classical MD.

The Root-Mean-Square Deviation (RMSD) calculated measures the average distance between the atoms in the backbone of a protein through the simulation, with its first geometry as reference.

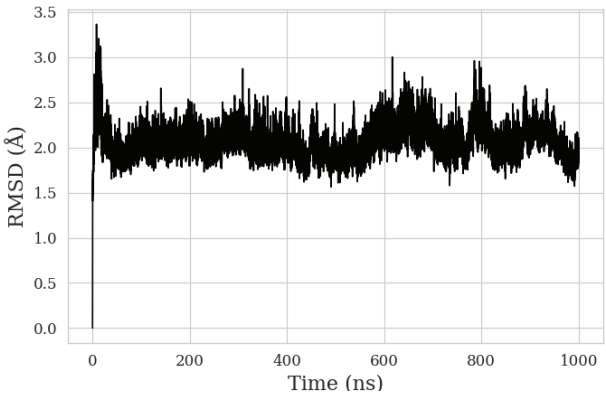

Figure 6. *1LYD RMSD during the production run of 1ms. The deviations after 50ns are always below 3Å.*

It's possible to see on Figure 6 that the 1LYD protein stabilizes the fluctuations of the RMSD profile between 1.5 and 3.0Å after 50ns.

In order to analyse the tryptophans mobility inside the protein, we calculated in Figure 7 the Root-Mean-Square Fluctuation (RMSF) by residue, measuring the mass-weighted fluctuation of each residue throughout the classical MD simulation.

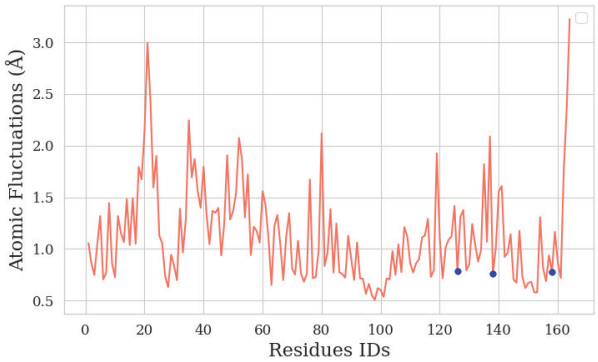

Figure 7. *RMSF by residue, with the 3 Trps present on 1LYD (Trp126, Trp138 and Trp158) highlighted in blue. As all tryptophans are in a low flexibility region, it's expected that the environment remains similar throughout the simulation for each tryptophan.*

Quantifying the mobility of each residue, we observe that the 3 tryptophans have a low mobility, fluctuating 0.7Å through the simulation. This measurement is important to indicate if the environment for each of the residues remains similar.

The radial distribution function (RDF, g(r)) measures the radial distance between two particles. The function was calculated between the oxygen present in the solvent water molecules and the nitrogen in the indole averaged over the MD run.

The tryptophans 126 and 158 have the first solvation shell located at 3.0Å from the nitrogen in the indole ring and we don't observe a second solvation shell. The tryptophan 138 has a lower density of waters around the nitrogen, having it only solvation shell around 3.8Å.

The Trp138 is less exposed to the water as it was shown in

Figure 7 As the L_A state is stabilized by polar environments, the Trp138 have a lower stabilization and a potential energy surface (PES) higher than the other two Residues. Because of that, its emission should present a blueshift.

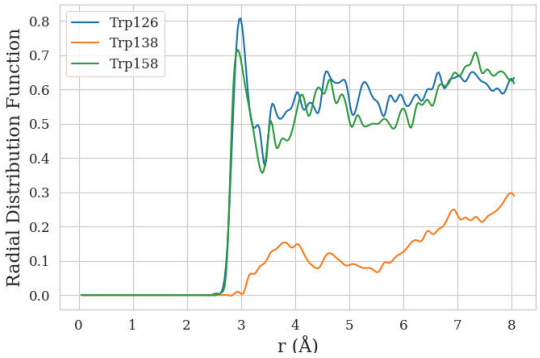

Figure 8. *RDFs measuring the distance between the oxygen in the solvent water molecules and the nitrogen in the indole. The Trp126 and Trp158 have a similar hydration and the Trp138 is the least hydrated from the tryptophans as it is in an inner region of the protein.*

3.3. Protein's Fluorescence

The tryptophan has a reported absorption of 280nm (4.43eV), with a small peak at 290nm (4.28eV)[32]. Because the optical absorption of proteins is mostly due to the tryptophan, mostly studies in the literature measures the emission of protein's fluorescence by exciting it with wavelengths between 280 and 300nm.[33,34] For the wild-type T4-Lysozyme and for the three mutant proteins containing only one tryptophan each (T4-lysozyme W126, T4-lysozyme W138 and T4-lysozyme W158) the reported emission is 330nm (3.76eV)[34].

The averaged absorptions and emissions for the experimental data and both sampling methods, QM/MM and classical MM are reported on Table 1.

The QM/MM sampled structures showed an average absorption close to the experimental, and virtually the same for all the tryptophans.

Regarding the emission, the Trp138 shows a difference of 0.11 and 0.15eV from the Trp126 and Trp158 emissions, respectively. The PBE0 method, together with M06, shows the smallest mean absolute error in terms of transition energy (0.22 - 0.23eV)[35] when comparing with the functionals B3LYP, M06-2X, CAM-B3LYP, and LC-PBE for TD-DFT.

Table 1. *Experimental and average absorptions and emissions calculated with QM/MMPol. The averaged energies with its std. deviation are obtained from the sampled structures for each Trp in both methods.*

	Experimental results			QM/MM Structures			QForce FF Structures		
Residue	**Abs. (eV)**	**Em. (eV)**	**Stoke Shift (eV)**	**Avg. Abs. (eV)**	**Avg. Em. (eV)**	**Stoke Shift (eV)**	**Avg. Abs. (eV)**	**Avg. Em. (eV)**	**Stoke Shift (eV)**
T4 - 126	4.43*	3.76†	0.67	4.64 ±0.14	3.73 ±0.25	0.91 ±0.29	4.58 ±0.15	4.18 ±0.18	0.40 ±0.23
T4 - 138	4.43*	3.76†	0.67	4.63 ±0.14	3.84 ±0.31	0.79 ±0.34	4.55 ±0.16	4.17 ±0.17	0.38 ±0.23
T4 - 158	4.43*	3.76†	0.67	4.64 ±0.14	3.69 ±0.28	0.95 ±0.31	4.58 ±0.16	4.19 ±0.17	0.39 ±0.23

** Reported Abs=280nm at Mansoor, B.Eggum(1968)[33] for the tryptophan. † Reported Em=330nm at Harris and Hudson (1991)[34] T4-Lysozyme protein*

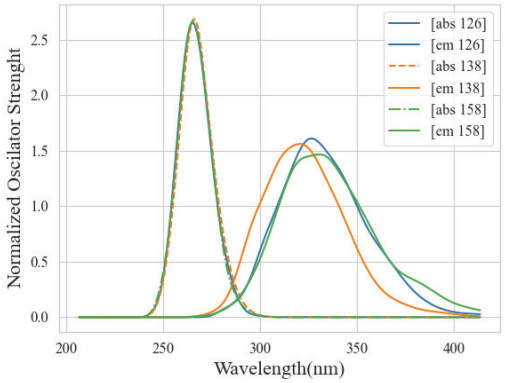

Figure 9. *Absorption and emission spectra for the QM sampled structures. The vertical transitions were broadened by a gaussian broadening function (σ=0.1).*

The difference of emission between the Trp138 and the Trp126 and Trp158 falls under the method's error, which still didn't achieve the chemical accuracy of 0.10eV. Due to this, it's not possible to say that the difference in emissions is only because of the molecular environment.

The vertical transition energies were broadened by a gaussian broadening function (σ=0.1) and plotted to better visualize the data. On the broadened spectra it's possible to observe the differences between the emissions for QM (Figure 9) and the classical geometries (Figure 10).

Regarding the spectra obtained from the classically sampled structures, all the absorption and emissions are virtually the same for the three tryptophans. The standard deviations for the emissions are lower than the ones sampled with QM/MM. This indicates that the geometries for the excited state FF could be less flexible than the higher-level theory framework. This will be investigated further on this study.

Another reason that can explain this is that the excited state charges are optimized to the L_A state only on the beginning of the simulation. As discussed before, the electrostatic effects are the major contribution for the excited state geometry. The indole's sensitivity to the variations on its charge and structure seems to be highly affected by the lack of update on the charges and a possibly less flexible geometry for the FFs, in comparison with the QM structures, on the final calculated emission.

The classical structures show an overestimation on the de-excitation energies. This could be caused by an overpolarization arisen from the MM charges being placed too close of the QM region. This effect could arise from the conjunction of the water solvation in the QM calculation used for the FF parameter fitting in with the QM/MMPol. Consequently, the system accounts for the electrostatic interactions for parametrizing the FF and again for the calculated transitions.

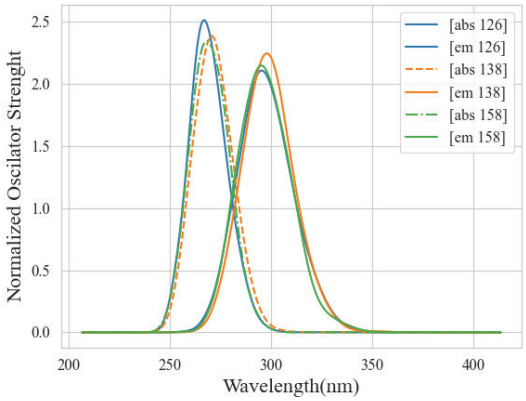

Figure 10. *Absorption and emission spectra for the MM sampled structures. The broadening is made the same way as the QM spectra.*

Still, the usage of solvation model can't be excluded totally for the geometry optimization as it is an important to reproduce the excited state geometry that is highly stabilized by the solvent. The possible overpolarization is still to be investigated.

Regarding the Stoke Shift, that is the difference in the peaks for the absorption and the emission, it's possible to observe on Table 1 a shift of 0.67eV for the experimental data. The QM sampling seems to overestimate this shift (0.80ev - 0.95eV), while the MD sampling underestimates it (0.38eV - 0.30eV). Both shifts fall again in the error of the method.

3.4. Q-Force structure analysis

As discussed before, the overestimation of the emission energies for the MM sampling could be cause by the final structures using the augmented FFs. To analyse the influence solely of the structure in the emission of the tryptophans sampled with Q-Force's FF we calculated the emission for the structures without solvation for both frameworks.

We compared the previously reported emission with both the electrostatic and polarizability effects of the environment with this second one without any embedding.

Table 2. *Emission with and without (w/o) full mutual polarizable effects.*

	QM Stucutre			MM structure		
Residue	**Elec.+Pol. effects (eV)**	**w/o Elec+Pol effects (eV)**	**Diff.**	**Elec.+Pol. effects (eV)**	**w/o Elec+Pol effects (eV)**	**Diff.**
Trp126	3.73 ±0.25	4.07 ±0.23	-0.34	4.18 ±0.18	4.47 ±0.15	-0.29
Trp138	3.84 ±0.31	3.99 ±0.21	-0.15	4.16 ±0.17	4.28 ±0.17	-0.12
Trp158	3.69 ±0.28	4.00 ±0.24	-0.31	4.19 ±0.17	4.38 ±0.16	-0.19

With this procedure it's possible to decouple the electrostatic and polarizability effects and understand which influence the geometry of the tryptophans have into the emission.

The mutual polarization contribution is the same for both frameworks, meaning that the structures should not be the problem in the systemically larger de-excitation energies.

4. CONCLUSION

We have investigated the excited state transitions of the tryptophan inside a complex system with two different frameworks, the physically more accurate QM/MM, and the classical treatment with quantum augmented force fields by Q-Force.

The ground state FF has shown absorption energies in well agreement with the more sophisticated QM/MM procedure and the experimental data, with its variation inside the error's method of 0.22eV.

For the excited state FF, the emissions for the Trp126, Trp138 and Trp158 were 0.45, 0.33 and 0.50eV higher than the energies from the QM/MM method.

The environmental electrostatic and polarization effects on the indole highly influence its charges and structure, and therefore the transition energies. The emission's divergence for the classical framework can be explained by the charges that are kept fixed through the 10ps simulation, while on Vivian and Callis[20] approach they update the charges each 10fs. Updating the charges in this framework is still a technical challenge.

Regarding Q-Force FFs, it was shown that the final structures for G.S. FFs reproduce the excitation energies from the QM structures, and that the E.S. FFs geometries reproduce the polarization contribution. It's systematic difference in the de-excitation still has to be investigated

Furthermore, to understand where this systemically larger de-excitation energies comes from and better describe the excite state with a force field, the next strategies can be considered:

1. The indole sensitivity to the charges and its variation is sufficiently high to make it necessary to update it during the MM.
2. The nonbonded parameters from generalized force fields are extensively parametrized for the ground state. Using a nonbonded parameters for the excited state for the chromophore can describe better the interaction with its surroundings.
3. Fitting the force field parameters with a continuum solvation model in the QM calculation may cause an overpolarization. This effect is still to be investigated.

5. ACKNOWLEDGEMENT

Thanks to Erasmus Mundus program that enabled me to have this opportunity of studying abroad. To Carles Curutchet and Shirin Faraji that were always open to me gave the means to perform this investigation as well as many scientific insights and teaches. To Maximilian that taught me so much and showed himself not only as a supervisor, but also as a friend.

And a special thanks to my wife Natália for her love and support, which made all this journey happier, and for keeping me sane.

6. REFERENCE

1. Shimomura, O., Johnson, F. H. & Saiga, Y. Extraction, Purification and Properties of Aequorin, a Bioluminescent Protein from the Luminous Hydromedusan,Aequorea. *Journal of Cellular and Comparative Physiology* **59**, 223–239 (1962).
2. SHIMOMURA, O. & JOHNSON, F. H. Regeneration of the photoprotein aequorin. *Nature* **256**, 236–238 (1975).
3. Chalfie, M., Tu, Y., Euskirchen, G., Ward, W. W. & Prasher, D. C. Green Fluorescent Protein as a Marker for Gene Expression. *Science (1979)* **263**, 802–805 (1994).
4. Cubitt, A. B. *et al.* Understanding, improving and using green fluorescent proteins. *Trends in Biochemical Sciences* **20**, 448–455 (1995).
5. Giepmans, B. N. G., Adams, S. R., Ellisman, M. H. & Tsien, R. Y. The Fluorescent Toolbox for Assessing Protein Location and Function. *Science (1979)* **312**, 217–224 (2006).
6. Betzig, E. Proposed method for molecular optical imaging. *Optics Letters* **20**, 237 (1995).
7. Betzig, E. *et al.* Imaging Intracellular Fluorescent Proteins at Nanometer Resolution. *Science (1979)* **313**, 1642–1645 (2006).
8. Hell, S. W. & Wichmann, J. Breaking the diffraction resolution limit by stimulated emission: stimulated-emission-depletion fluorescence microscopy. *Optics Letters* **19**, 780 (1994).
9. Hell, S. W. & Kroug, M. Ground-state-depletion fluorscence microscopy: A concept for breaking the diffraction resolution limit. *Applied Physics B Lasers and Optics* **60**, 495–497 (1995).
10. Moerner, W. E. & Kador, L. Optical detection and spectroscopy of single molecules in a solid. *Physical Review Letters* **62**, 2535–2538 (1989).
11. Lopez, A. J. & Martínez, L. Parametric models to compute tryptophan fluorescence wavelengths from classical protein simulations. *Journal of Computational Chemistry* **39**, 1249–1258 (2018).
12. Rao, G. *et al. Application of Fluorescence Sensing to Bioreactors*. *Topics in Fluorescence Spectroscopy* (2006). doi:10.1007/0-306-47060-8_13.
13. Lakowicz, J. R. *Topics in Fluorescence Spectroscopy*. *Topics in Fluorescence Spectroscopy* (Springer US, 1994). doi:10.1007/b112911.
14. Callis, P. R. *Predicting fluorescence lifetimes and spectra of biopolymers*. *Methods in Enzymology* vol. 487 (Elsevier Inc., 2011).
15. Brisker-Klaiman, D. & Dreuw, A. Explaining level inversion of the La and Lb states of indole

and indole derivatives in polar solvents. *ChemPhysChem* **16**, 1695–1702 (2015).

16. Menger, M. F. S. J., Caprasecca, S. & Mennucci, B. Excited-State Gradients in Polarizable QM/MM Models: An Induced Dipole Formulation. *Journal of Chemical Theory and Computation* **13**, 3778–3786 (2017).
17. Senn, H. M. & Thiel, W. QM/MM methods for biomolecular systems. *Angewandte Chemie - International Edition* **48**, 1198–1229 (2009).
18. Sami, S., Menger, M. F. S. J., Faraji, S., Broer, R. & Havenith, R. W. A. Q-Force: Quantum Mechanically Augmented Molecular Force Fields. *Journal of Chemical Theory and Computation* **17**, 4946–4960 (2021).
19. Corbella, M., Cupellini, L., Lipparini, F., Scholes, G. D. & Curutchet, C. Spectral Variability in Phycocyanin Cryptophyte Antenna Complexes is Controlled by Changes in the α-Polypeptide Chains. *ChemPhotoChem* **3**, 945–956 (2019).
20. Vivian, J. T. & Callis, P. R. Mechanisms of tryptophan fluorescence shifts in proteins. *Biophysical Journal* **80**, 2093–2109 (2001).
21. Curutchet, C. *et al.* Photosynthetic Light-Harvesting Is Tuned by the Heterogeneous Polarizable Environment of the Protein. *J Am Chem Soc* **133**, 3078–3084 (2011).
22. Berman, H. M. The Protein Data Bank. *Nucleic Acids Research* **28**, 235–242 (2000).
23. Rose, D. R. *et al.* Crystal structure of T4-lysozyme generated from synthetic coding DNA expressed in Escherichia coli. *"Protein Engineering, Design and Selection"* **2**, 277–282 (1988).
24. Anandakrishnan, R., Aguilar, B. & Onufriev, A. v. H++ 3.0: automating pK prediction and the preparation of biomolecular structures for atomistic molecular modeling and simulations. *Nucleic Acids Research* **40**, W537–W541 (2012).
25. Izadi, S., Anandakrishnan, R. & Onufriev, A. v. Building Water Models: A Different Approach. *The Journal of Physical Chemistry Letters* **5**, 3863–3871 (2014).
26. Tian, C. *et al.* Ff19SB: Amino-Acid-Specific Protein Backbone Parameters Trained against Quantum Mechanics Energy Surfaces in Solution. *Journal of Chemical Theory and Computation* **16**, 528–552 (2020).
27. D.A. Case, K. Belfon, I.Y. Ben-Shalom, S.R. Brozell, D.S. Cerutti, T.E. Cheatham, III, V. W. D. C. *et al. AMBER 2020*. (University of California, 2020).
28. Vreven, T. *et al.* Combining Quantum Mechanics Methods with Molecular Mechanics Methods in ONIOM. *Journal of Chemical Theory and Computation* **2**, 815–826 (2006).
29. M. J. Frisch, G. W. Trucks, H. B. Schlegel, G. E. Scuseria, M. A. Robb, J. R. Cheeseman, G. Scalmani, V. Barone, G. A. Petersson, H. Nakatsuji, X. Li, M. Caricato, A. V. Marenich, J. Bloino, B. G. Janesko, R. Gomperts, B. Mennucci, H. P. Hratchian, J. V., J. B. F. and D. J. F. Gaussian 16, Revision C.01. (2016).
30. Curutchet, C. *et al.* Electronic Energy Transfer in Condensed Phase Studied by a Polarizable QM/MM Model. *Journal of Chemical Theory and Computation* **5**, 1838–1848 (2009).
31. Ryckaert, J.-P. & Bellemans, A. Molecular dynamics of liquid alkanes. *Faraday Discussions of the Chemical Society* **66**, 95 (1978).
32. Jacquemin, D., Planchat, A., Adamo, C. & Mennucci, B. TD-DFT assessment of functionals for optical 0-0 transitions in solvated dyes. *Journal of Chemical Theory and Computation* **8**, 2359–2372 (2012).
33. Eggum, B. O. Determination of Tryptophan. *Acta Agriculturae Scandinavica* **18**, 127–131 (1968).
34. Harris, D. L. & Hudson, B. S. Fluorescence and molecular dynamics study of the internal motion of the buried tryptophan in bacteriophage T4 lysozyme: Effects of temperature and alteration of nonbonded networks. *Chemical Physics* **158**, 353–382 (1991).