D-luciferin Bioluminescent System: a Computational Characterization of its Emitting Electronic State

Departmento de Química
Facultad de Ciencias
Universidad Autónoma de Madrid

A thesis submitted to the program of Theoretical Chemistry and Computational Modeling for the degree of Doctor of Philosophy (Ph.D.)

by

Henar Mateo de la Fuente

Supervisor: Dr. Juan J. Nogueira
Submission: Madrid, January 2026

A winged spark doth soar about —
I never met it near
For Lightning it is oft mistook
When nights are hot and sere —

It's twinkling Travels it pursues
Above the Haunts of men —
A speck of Rapture — first perceived
By feeling it is gone —

Emily Dickinson

Roto el después, el antes, el tal vez mañana, el mundo
se encabrita, y las cigarras, los cascarudos, los bichitos
de luz, debajo de las ramas, se dan cita en el estío galopante,
con pasmosa impertinencia.
No se puede pedir más.
La finitud cabalga como puede.

María Negroni

Acknowledgements

Posiblemente, sea ésta la sección que más gente se lea y precisamente por eso es la que más vértigo me da escribir. No porque no tenga a nadie a quien agradecer, sino más bien por todo lo contrario. Hay tanta gente que ha compartido este camino conmigo... A las que estáis, solo puedo agradeceros enormemente todo el cariño y el apoyo recibido... y las que desgraciadamente no habéis podido acompañarme hasta el final, gracias por todos esos momentos inolvidables, esos abrazos especiales y esos ruidos de tetera.

A mi familia, gracias por estar, en los altos y en los bajos. Mamá, papá, gracias por apoyarme en todas mis decisiones tanto académicas como personales y por brindarme todas las oportunidades necesarias para llegar a ser la mujer que soy hoy. Coco, gracias por lanzarte conmigo a nuestra pequeña aventura artística. Sé que para ti supone un gran esfuerzo y por eso valoro tantísimo esos ratitos que pasamos juntas en el taller, aunque a veces me desesperes por pasar más tiempo con el móvil que con el pincel. Gracias Marta, mi maravillosa novia, por elegirme cada día, gracias por hacerme reír y por aguantar mis tonterías. Pero, sobre todo, gracias por ayudarme a desconectar, porque las dos sabemos que sin ti no habría disfrutado ni la mitad de estos años de tesis (sí, algo he disfrutado). Junto con Marta vino su familia, gracias por tener siempre un plato para mí en vuestra mesa.

Y porque familia también es la que se elige, voy a aprovechar también para agradecer infinitamente a mis amigos. A mis compis del cole y del instituto -Lidia, María y Raquel- por ayudarme a derribar mis propias barreras. A mis amigos de la carrera - Laura, Marta, Darío y Sergio- por aportarme un arcoíris de posibilidades (y a conseguir el grado sin hacer una sola valoración un besito wapas). Pensar en vosotras es recordar momentos icónicos en el aula y viajes divertidísimos por muchos rincones de Europa (y los que nos quedan, porque siempre estarás con nosotras). A mi grupo de sufridoras confinadas -Gema, Silvia, Marta, Lucía, Emilio y Manu- y al resto de compañeras de tesis -Nuria, Javi, Jesús, Darío, Eva, Gustavo, Sergio, Antonio, Paola, Lorena y Vito-. Aunque nos conocimos en momentos diferentes, con todas he podido disfru-

tar congresos e interesantes discusiones científicas, pero también cuencazos y excursiones. Julia, Lucía, gracias por las tardes de *fika* y de esquizofrenia colectiva, pero sobre todo, por el Churro y la Porra. Serafina, gracias por ahorrarme tanto sufrimiento.

Tampoco me puedo olvidar de todos los profesores que, gracias a su gran labor docente, me han impulsado a elegir mi camino, primero el de la ciencia, luego el de la química y, más adelante, el de la Academia. Entre todos me habéis enseñado desde las cosas más básicas como la historia de la tabla periódica o a hacer *ls* hasta los conceptos más complicados de la química. Por supuesto, agradezco a mi director, Juanjo, por su tiempo y su supervisión pero, sobre todo, por sus palabras de reconocimiento al terminar un proyecto. Nunca caen en saco roto. Pero también agradezco el entusiasmo de mis profesores del instituto y de la carrera (hasta he tenido la suerte de que algunos os habéis convertido en compañeros). Hacer una lista sería imposible, pero espero os reconozcáis cuando digo que se agradece cuando os pasáis a saludar por el despacho (pero no cuando os coméis nuestro chocolate) y dedicáis vuestro tiempo en compartir vuestra experiencia docente.

I would like to thank Vera for hosting me in Darmstadt, Adam for his research guidance during my stay there and Shweta for the late evening chitchat. But, most importantly, I would like to thank you all for your warm welcome (and for letting me in your Indian restaurant trips). I cannot wait to see you again.

Y para terminar, me gustaría agradecer formalmente a las distintas instituciones que han financiado la realización de esta tesis. Namely:

- La Universidad Autónoma de Madrid a través de la Ayuda para el Fomento de la Investigación en Estudios de Máster y del contrato predoctoral de Formación de Personal Investigador (FPI-UAM).

- The Project HPC-EUROPA3 (INFRAIA-2016-1-730897), with the support of the EC Research Innovation Action under the H2020 Programme and the CINECA computing center.

- The computer resources at Cibeles and the technical support provided by UAM (RES-QH-2025-1-0011).

- El programa Programa Becas Santander a través de la Ayuda Económica para Personal Investigador Predoctoral 2025.

Contents

List of Figures 1
List of Tables 15
List of Schemes 17
List of Listings 18
List of Abbreviations 19

Abstract **23**

Resumen **24**

I Introduction **27**
1.1 Bioluminescence 29
1.1.1 Emitting Species 31
1.1.2 Light Generation in Bioluminescent Systems 35
1.2 Fireflies Bioluminescence 37
1.2.1 Firefly's Life Cycle and Bioluminescent Phenotype 37
1.2.2 D-Luciferin Biochemical Cycle 39
1.2.3 Colour Modulation 46
1.2.4 Applications in Medicine 51

II Computational Methods **57**
2.1 Electronic Structure Theory 60
2.1.1 The Time Independent Schrödinger Equation 60
2.1.2 The Born-Oppenheimer Approximation 62
2.1.3 The Hartree-Fock Approximation 65
2.1.4 The Variational Principle 73
2.1.5 The Molecular Orbitals Theory and Basis Sets 74
2.1.6 The Density Functional Theory 75
2.1.7 The Time Dependent Density Functional Theory 83
2.1.8 Typical Quantum Mechanical Calculations 91
2.2 Nuclear Motion Simulation 100

2.2.1 The Hamilton's Equations of Motion . . . 100
2.2.2 The Integrator . . . 101
2.2.3 Interaction Model . . . 103
2.2.4 Thermodynamic Ensembles . . . 108
2.2.5 Periodic Boundary Conditions . . . 111
2.2.6 Constrained Dynamics . . . 115
2.2.7 The Sampling . . . 116
2.2.8 Typical Analysis for Molecular Dynamics Simulations . . 117
2.3 Hybrid QM/MM Approaches . . . 120
2.4 Interaction Energy . . . 123
2.4.1 MM-Generalized Born Surface Area . . . 123
2.4.2 Energy Decomposition Analysis . . . 127

III Electronic Properties of the Phenolate-Keto Oxyluciferin inside the Wild Type Luciferase 135
3.1 Introduction . . . 139
3.2 Computational Details . . . 139
3.2.1 Static Potential Energy Rigid Scan . . . 139
3.2.2 Molecular Dynamics Simulations . . . 140
3.2.3 Analysis of the Electronic Properties . . . 144
3.3 Results and Discussion . . . 145
3.3.1 Static Scan around the C–C' Bond . . . 145
3.3.2 Absorption and Emission Spectra from the Classical Trajectories . . . 149
3.3.3 Absorption and Emission Spectra from the QM/MM Trajectories . . . 153
3.3.4 Effect of the Intramolecular and Intermolecular Degrees of Freedom on the CT Character . . . 155
3.4 Conclusions . . . 162

IV OLUFF: a Novel Set of Ground and Excited State FF Parameters of the Emitting Oxyluciferin Species 165
4.1 Introduction . . . 169
4.2 Computational Details . . . 170
4.2.1 Parametrization Process . . . 170
4.2.2 Validation of the OLUFF . . . 172
4.3 Results and Discussion . . . 174
4.3.1 Global Fitting Process . . . 175
4.3.2 Potential Energy Profiles and Probability Distributions . 178
4.3.3 QM and OLUFF Minimized Structures . . . 182

4.3.4 Dynamic Infrared Spectra . . . 183
4.3.5 Principal Components Analysis . . . 187
4.4 Conclusions . . . 189

V Structure of the Oxyluciferin Emitting Species in Aqueous and Luciferase Media 191
5.1 Introduction . . . 195
5.2 Computational Details . . . 195
5.2.1 System Setup . . . 195
5.2.2 Classical Molecular Dynamics . . . 196
5.3 Results and Discussion . . . 197
5.3.1 Aqueous Trajectories . . . 197
5.3.2 Trajectories of the OLU/luciferase Complex . . . 201
5.4 Conclusions . . . 210

VI Characterization of the Nuclear Coordinate for the Phenolate-Keto Oxyluciferin Red-Shift Emission from Quantum Mechanical Calculations 213
6.1 Introduction . . . 217
6.2 Computational Details . . . 218
6.2.1 Classical Molecular Dynamics . . . 218
6.2.2 Principal Components Analysis . . . 218
6.2.3 Fitting of Emission Energies . . . 219
6.2.4 Scan along the Energy Shift Dimension . . . 220
6.3 Results and Discussion . . . 220
6.3.1 Fitting of Emission Energies . . . 220
6.3.2 Scan along the Energy Shift Dimension . . . 222
6.4 Conclusions . . . 225

VII Classical and QM/MM Energy Decomposition Analysis of Oxyluciferin/Luciferase in the Electronic Excited State 227
7.1 Introduction . . . 231
7.2 Computational Details . . . 232
7.2.1 Interaction Energy . . . 233
7.3 Results and Discussion . . . 235
7.3.1 Phenolate-Keto Oxyluciferin . . . 235
7.3.2 Phenol-Keto, Phenol-Enolate and Phenolate-Enolate Oxyluciferins . . . 242
7.4 Conclusions . . . 247

VIII Structure of the Phenolate-Keto Oxyluciferin Emitting Species inside the SER286ASN Mutated Luciferase 249
8.1 Introduction . . . 253
8.2 Computational Details . . . 254
8.2.1 System Setup . . . 255
8.2.2 Classical Molecular Dynamics . . . 255
8.3 Results and Discussion . . . 257
8.3.1 Structure of the Oxyluciferin Chromophores . . . 257
8.3.2 The Role of Water Molecules . . . 258
8.3.3 OLU/luciferase Interactions . . . 260
8.3.4 The Role of AMP . . . 264
8.4 Conclusions . . . 266

IX Conclusions 269

X Conclusiones 275

Bibliography 281

A Structure and Abbreviations of Amino Acids II

B Electronic Properties of the Phenolate-Keto Oxyluciferin inside the Wild Type Luciferase VI
B.1 Oxyluciferin Parameters for the Classical Molecular Dynamics Simulations . . . VI
B.1.1 Atom Types and Charges . . . VI
B.1.2 Stretching . . . VII
B.1.3 Bending . . . VIII
B.1.4 Torsion . . . X
B.1.5 Non-Bonding Parameters . . . XIII
B.2 Variation of the Oscillator Strength and Charge Transfer for different Density Functionals . . . XIV
B.3 Absorption and Emission Spectra from the Classical and QM/MM Trajectories and Angle Contributions. . . . XV
B.4 Probability Distribution of the SCCS Dihedral Angles for the 500 ns Trajectory . . . XVI
B.5 Environment of the Two Clusters with the Smallest and Largest CT Number Values from the QM/MM Trajectory of the S_0 State. XVII
B.6 Total Binding Free Energy per Residue of the Classical S_1 Trajectories Without and With AMP . . . XIX

B.7 Representation of the Three Positions Found for the Analysed Snapshots of the S_0 QM/MM Trajectory XX
B.8 Charge Transfer Number Variation with Respect to the Distance of the Closest Amino-acids of the Protein XXI

C OLUFF: a Novel Set of Ground and Excited State FF Parameters of the Emitting Oxyluciferin Species XXII
C.1 System Preparation and Molecular Dynamics simulations XXII
C.2 CAM-B3LYP Basis Set Comparison XXVII
C.3 Nature of the S_1 Transition XXVII
C.4 Convergence of the Fitting Process XXVIII
C.5 Fitting Process XXX
C.6 Periodicity of the Torsional Functions Outside of the Fitted Domain XXXV
C.7 Evolution of the SCCS Dihedral Angle XXXV
C.8 Infrared Spectra XXXVI

D Structure of the Oxyluciferin Emitting Species in Aqueous and Luciferase Media XXXVIII
D.1 Root Mean Square Displacement of the Protein XXXVIII
D.2 Root Mean Square Displacement of the OLU Chromophores in Aqueous Media XXXIX
D.3 Change in the SCCS Dihedral of the Phenol-Keto OLU in the S_0 XL

E Characterization of the Nuclear Coordinate for the Phenolate-Keto Oxyluciferin Red-Shift Emission from Quantum Mechanical Calculations XLIV
E.1 Convergence of the Sampling of the Principal Components . . . XLIV
E.2 Convergence of the Electronic Properties LXVI
E.3 Variances of the Principal Components LXVII
E.4 Frequencies of the Principal Components LXVIII

F Classical and QM/MM Energy Decomposition Analysis of Oxyluciferin/Luciferase in the Electronic Excited State LXX
F.1 Benchmark of Functionals and Basis Sets LXX
F.2 Number of Amino Acids in the QM Region LXXII
F.3 Charge Transfer along the MD trajectories LXXIV
F.4 Convergence with the Number of Frames of Phenol-keto, Phenol-enolate and Phenolate-enolate Oxyluciferins LXXV

G Structure of the Phenolate-Keto Oxyluciferin Emitting Species inside the SER286ASN Mutated Luciferase LXXVI
G.1 RMSD of the Protein LXXVI
G.2 Increased Temperature Simulation LXXVII
G.3 SCCS Angle of the Oxyluciferin LXXVIII

List of Figures

1.1 Schematic representation of complex formation from the luciferin chromophore (pink) and the luciferase (orange). 29
1.2 Representation of the nine natural luciferines with a representative BL organism: (A) coelenterazine, (B) cypridina luciferine, (C) bacterial luciferin (D) tetrapyrrole-based luciferine with X=H (dinoflagellates) and X=OH (krill), (E) fungal luciferin, (F) Latia luciferin, (G) Diplocardia luciferin, (H) Fridericia luciferin and (I) D-luciferin. 32
1.3 (A) *Photinus pyralis* life cycle. (B) Phenotypical differentiation between female (B.i) and male (B.ii) lanterns of *Photinus pyralis*. (C) Schematic representation of lampyrid lantern. (C).i-iv are described in text. 38
1.4 Graphical representation of *in vivo* BL of mice adapted from Figure 1 of reference [169]. 53
1.5 Schematic representation of different BL bioimaging techniques: A) BL Protein–Fragment Complementation Assay, B) circular permutation upon i) small ligand attachment and ii) protease activity and C) using cyclic luciferases. D-luciferin based luciferase is represented in orange and external peptide sequences, in pink. 54

2.1 Jacob's ladder of density functional approximations. 82
2.2 Graphical representation of the effect of the response function. . 88
2.3 Graphical representation of a general optimization procedure. . 92
2.4 Graphical representation of (A) the Steepest Descent and (B) Conjugate Gradient methods. 93
2.5 Graphical representation of the HO model. 96
2.6 Graphic representation of the main kinds of interactions employed in FFs. (A - D) Represent the bonding terms (bond stretching, angle bending, proper dihedral and improper dihedral, respectively) and (E), the non-bonding terms. 104

2.7 Schematic 2D representation of the replication of the primary cell (dark orange) in virtual cells (black) performed when periodic boundary conditions are applied to the MD simulation. . . 112

2.8 Schematic representation of the van der Waals (dark orange) and coulomb (dark pink) pair potentials as a function of the interatomic distance. 113

2.9 Schematic representation of the Ewald summation. 114

2.10 Graphical representation of the QM/MM schemes and embeddings. Light orange and dark pink represent, respectively, the classic and QM treatments of the inner (I) and outer (O) parts. Dashed lines represent the point charges of the outer region. . . 121

2.11 Schematic representation of the thermodynamic cycle used for the MM-GBSA calculation. The ligand is represented in dark pink; the receptor, in orange and the continuum solvent, in light pink. 124

2.12 Graphical representation of the (A) BSSE and (B) CP correction; where solid circles represent the presence of monomers A (orange) and B (pink) and blurred circles, their corresponding basis set. 128

3.1 Scheme of the computational protocol followed in Chapter III. . 141

3.2 S_0 (blue) and the S_1 (pink) rigid potential energy curves calculated with the density functionals: (A) B3LYP, (B) CAM-B3LYP, (C) M06-2X, (D) PBE0 and (E) ωB97XD. The secondary x-axis shows the angle difference with respect to the planar energy minimum found at a SCCS angle of 180^{o}. 146

3.3 (A) Variation of the oscillator strength (orange) and CT number (green) for the S_0 - S_1 electronic transition along the dihedral angle calculated with the CAM-B3LYP density functional. The secondary x-axis shows the angle difference with respect to the energy minimum. (B) HOMO and LUMO orbitals for the planar and perpendicular geometrical situations. 148

3.4 (A) Absorption and (B) emission spectra computed from the classical trajectories, together with their respective contributions from the S_0 - S_1 band depending on the SCCS dihedral angle. Additionally, the emission spectra obtained from the classical MD with the AMP inside the cavity is shown with a dashed line. 150

3.5 Probability distributions of the SCCS dihedral angles for the S_0 (blue) and the S_1 (pink) trajectories evolved at (A) classical and (B) QM/MM level. The secondary x-axis shows the angle difference with respect to the energy minimum. 152

3.6 (A) Absorption and (B) emission spectra computed from the QM/MM trajectories, together with their respective contributions from the S_0 - S_1 band depending on the SCCS dihedral angle. 154

3.7 Variation of the CT number with the the SCCS dihedral angle difference from planarity for the (A) absorption and (B) emission spectra computed from snapshots taken from the classical (orange) and QM/MM (green) trajectories. 156

3.8 (A) Centroid geometries of the two clusters with the 20 snapshots with smallest (blue) and largest (pink) CT number values extracted from the QM/MM trajectory of the S_0 state. Clusters of the 20 geometries with (B) smallest and (C) largest CT number values. 157

3.9 Total (blue) and electrostatic (pink) binding free energy of the chromophore with the closest residues of the protein along the S_0 QM/MM selected snapshots. 159

3.10 (A) CT number variation with the distance between the centres of mass of the chromophore and different amino-acids along the 100 S_0 QM/MM MD selected snapshots. (B) RMSD of the OLU inside the cavity. For both representations colours blue, green and orange represent clusters 1 to 3. 160

4.1 (A) Time evolution and (B) probability distribution of the SCCS dihedral angle upon the mutation of GLY317 (blue) and SER349 (pink) for ALA. 169

4.2 Average correlation coefficient (pink) and number of fittings m with a R^2 higher than 0.7 with respect to the number of terms in the torsion potential achieved during the initial 100 fittings of the SCCS dihedral angle parameters for the (A, C, E, G) S_0 and (B, D, F, H) S_1 states of the OLU (A, B) phenolate-keto, (C, D) phenol-keto, (E, F) phenol-enolate and (G, H) phenolate-enolate. In the plot, also the highest R^2 appear for each of the OLU chromophores. 177

4.3 Rigid PESs (solid lines) and probability distributions (points) of the SCCS dihedral angles of the (A, C, E, G) S_0 and (B, D, F, H) S_1 states of the OLU (A, B) phenolate-keto, (C, D) phenol-keto, (E, F) phenol-enolate and (G, H) phenolate-enolate. Energies are relative to the minimum of each surface. 180
4.4 QM (violet) and OLUFF (black) minimized structures of the OLU (A, B) phenol-keto, (C, D) phenol-enolate and (E, F) phenolate-enolate of the (A, C, E) S_0 and (B, D, F) S_1 states with their RMSD values. 182
4.5 Infrared spectra of the OLU (A, B) phenol-keto, (C, D) phenol-enolate and (E, F) phenolate-enolate of the (A, C, E) S_0 and (B, D, F) S_1 states calculated from the QM geometry optimization and the MD simulations using the GAFF and the OLUFF. . . . 184
4.6 Comparison between the QM and OLUFF frequencies of the OLU (A, B) phenol-keto, (C, D) phenol-enolate and (E, F) phenolate-enolate of the (A, C, E) S_0 and (B, D, F) S_1 states calculated from the QM optimized geometries and from a PC analysis, respectively. 188

5.1 Probability distributions of the SCCS dihedral angles of each of the OLU chromophores in the (A) S_0 and (B) S_1 states in aqueous media. 198
5.2 RDF of the O atoms from water surrounding each of the four electronegative sites of each of the OLU chromophores in the (A,C,E,G) S_0 and (B,D,F,H) S_1 states in aqueous media. 199
5.3 Average number of hydrogen bonds (HB) of each of the OLU chromophores with water and with each of the four electronegative atoms in the (A) S_0 and (B) S_1 states in aqueous media. . 200
5.4 Probability distributions of the SCCS dihedral angles of each of the OLU chromophores in the (A) S_0 and (B) S_1 states in the enzymatic environment. 201
5.5 RDF of the O atoms from water surrounding each of the four electronegative sites of each of the OLU chromophores in the (A,C,E,G) S_0 and (B,D,F,H) S_1 states in the enzymatic environment. 203
5.6 Average number of hydrogen bonds (HB) of each of the OLU chromophores with water and with each of the four electronegative atoms in the (A) S_0 and (B) S_1 states sites in the enzymatic environment. 204

5.7 Percentage of time with contacts between the four OLU chromophores in the (A) S_0 and (B) S_1 states and specific amino acids of the enzyme along the MD trajectory. Only those amino acids with contacts for at least the 10% of the simulation time have been included in the representation. 206

5.8 Average of hydrogen bonds (HB) between each of the four OLU chromophores in the (A) S_0 and (B) S_1 states and specific amino acids of the enzyme along the MD trajectory 207

5.9 Percentage of time with a water molecule acting as a hydrogen bonds (HB) bridge between each of the four OLU chromophores in the (A) S_0 and (B) S_1 states and specific amino acids of the enzyme along the MD trajectory. 208

5.10 (A,B) Distance between the OLU chromophore to the AMP cofactor, (C,D) average of hydrogen bonds (HB) between each of the four OLU chromophores and the AMP and (E,F) percentage of time with a water molecule acting as a HB bridge between each of the four OLU chromophores and the AMP for the MD trajectories in the (A,C,D) S_0 and (B,D,F) S_1 states of each chromophore. 210

6.1 Fitting coefficients together with the resulting energy shift dimensions as linear combinations of the PCs with expansion coefficients obtained from fit for the (A) 90^{o}, (B) s-trans and (C) 270^{o} conformations. 221

6.2 (A-C) PES and (D-E) emission energy (ΔE), oscillator strengths (f) and CT for geometries on (A,D) the 90^{o}, (B,E) s-trans and (C,F) 270^{o} conformations along the scanned energy shift dimension. The dashed horizontal line represents the the 600 nm threshold for the ΔE emission energy. 223

6.3 (A-C) Emission energy, (D-F) oscillator strengths and G-H) CT for the sampled geometries on (A,D,G) the 90^{o}, (B,E,H) s-trans and (C,F,I) 270^{o} conformations along the energy shift dimension. The dashed horizontal line represents the 600 nm threshold for the ΔE emission energy. 224

7.1 (A) Phenolate-keto OLU/luciferase complex with the 15 amino acids with the largest pairwise total binding free energy in pink. (B) Pairwise residue decomposition based on the 1A-MM-GBSA approach of total binding free energy (ΔG^{tot}) and the van der Waals (ΔG^{vdW}), electrostatic (ΔG^{elec}), polar solvation (ΔG^{pol}), non-polar (ΔG^{np}) solvation terms including the 40 residues that are located at shortest distances from OLU along the simulation. At the bottom of the the plot, between parenthesis, the total binding free energy of the chromophore appears and, at the right-hand side, the binding Gibbs free energy per residue of the 15 amino acids with the largest ΔG is represented. 232

7.2 (A) Convergence of the interaction energy and the EDA components as the number of amino acids in the QM region increases. (B) Number of transferred electrons between the two monomers (green) and averaged position of the orbital before and after the electronic transition (orange). 238

7.3 Convergence of the interaction energy and each of the EDA components with respect to the number of sampled geometries. 239

7.4 Energy distributions of the contributions to the total interaction energy calculated from the QM/MM-EDA on 250 geometries. . 240

7.5 Pairwise residue decomposition based on the MM-GBSA approach of total binding free energy (ΔG^{tot}) and the van der Waals (ΔG^{vdW}), electrostatic (ΔG^{elec}), polar solvation (ΔG^{pol}), non-polar (ΔG^{np}) solvation terms including the 40 residues that are located at shortest distances from the (A) phenol-keto, the (B) phenol-enolate and the (C) phenolate-enolate OLUs along the simulation. At the bottom of the the plot, between parenthesis, the total binding free energy of the chromophore appears. 244

7.6 Energy distributions of the contributions to the total interaction energy calculated from the QM/MM-EDA sampled geometries from the trajectories of the (A) phenol-keto, (B) phenol-enolate and (C) phenolate-enolate OLUs. 246

8.1 Scheme of the computational protocol followed along Chapter VIII. 254

8.2 Probability distributions of the SCCS dihedral angles of the phenolate-keto OLU in the WT (blue) and SER286ASN (pink) systems. 257

8.3 RDF of the O atoms from water surrounding each of the X four electronegative sites of phenolate-keto OLU (A-D) in the WT and SER286ASN systems and (E-H) in the different SCCS conformations in the SER286ASN system. 258
8.4 Average number of hydrogen bonds (HB) of the phenolate-keto OLU chromophore with water for each of the four electronegative atoms (A) in the WT and SER286ASN systems and (B) in the different SCCS conformations in the SER286ASN system. . 259
8.5 Percentage of time with contacts between (A) the phenolate-keto OLU and the WT and the SER286ASN luciferases and (B) the phenolate-keto OLU in the different SCCS conformations and the SER286ASN luciferase along the MD trajectory. Only those amino acids with contacts for at least the 1% of the simulation time have been included in the representation. 261
8.6 Average of hydrogen bonds (HB) between the phenolate-keto OLU with the enzyme for each of the four electronegative atoms (A) in the WT and SER286ASN systems and (B) in the different SCCS conformations in the SER286ASN system. Only those amino acids with at least an average of 0.1 of hydrogen bonds withe the OLU in at least one of the studied systems have been included in the representation. 262
8.7 Percentage of time with a water molecule acting as a hydrogen bond (HB) bridge between OLU chromophore an the enzyme (A) in the WT and SER286ASN systems and (B) in the different SCCS conformations in the SER286ASN system. Only those amino acids with a bridging water molecule for at least a 20 % of the simulation time of one of the studied systems have been included in the representation. 264
8.8 (A,B) Distance between the OLU chromophore to the AMP cofactor, (C,D) average of hydrogen bonds (HB) between the OLU chromophore and the AMP and (E,F) percentage of time with a water molecule acting as a HB bridge between the OLU chromophore and the AMP. Results for the MD trajectories in the (A,C,D) WT and SER286ASN systems and (B,D,F) the different SCCS conformations in the SER286ASN system. Vertical dashed lines in (A) represent the separation into the three different replicas. Secondary x axis in (B) applies to 90 and 270º conformations. 265

A.1 Amino acids with hydrophobic side chains. II

A.2 Amino acids with polar uncharged side chains. III
A.3 Special amino acids. III
A.4 Amino acids with negatively charged side chains. III
A.5 Amino acids with positively charged side chains. IV
A.6 Possible protonation states of histidine. IV

B.1 Structure of the OLU with labelled atoms. VI
B.2 Variation of the oscillator strength (orange) and CT number (green) for the S_0 - S_1 transition along the dihedral angle calculated with different density functionals: (A) B3LYP, (B) CAM-B3LYP, (C) M06-2X, (D) PBE0 and (E) ωB97XD. XIV
B.3 (A,C) Absorption and (B,D) emission spectra computed from the classical (A,B) and QM/MM (C,D) trajectories, together with their respective contributions from the S_0 to the S_1 band depending on the SCCS dihedral angle. XV
B.4 Probability distributions of the SCCS dihedral angles for the S_0 (blue) and the S_1 (pink) trajectories evolved at classical level for 500 ns. XVI
B.5 Representative structure with the protein environment of the two clusters with the 20 snapshots with smallest (A) and largest (B) CT number values extracted from the QM/MM trajectory of the S_0 state. XVII
B.6 RDF between (A) the O atom of the tz fragment and the O atom of water and (B) the O atom of the btz fragment and the O atom of water for the two clusters with the 20 snapshots with smallest (blue) and largest (pink) CT number values extracted from the QM/MM trajectory of the S_0 state. XVIII
B.7 Total binding free energy per residue of the classical S_1 trajectories without (orange) and with AMP (green) for the residues that are closer for at least 1% of the simulation time. XIX
B.8 Representation of the three clusters of the analysed snapshots of the S_0 QM/MM trajectory XX
B.9 CT number variation with the distance between the centres of mass of the chromophore and different amino-acids for the clusters of the 20 geometries with smallest (blue) and largest (pink) CT number values XXI

C.1 Comparison of the rigid PESs of the SCCS dihedral angles of the OLU phenolate-keto in the S_0 at CAM-B3LYP/6-311G(2d,p) and CAM-B3LYP/6-31G* levels of theory. XXVII

C.2 Participation of the HOMO-LUMO transition on the S_1 state of the OLU phenolate-keto. XXVII
C.3 Convergence of the fitted V_n' (A, C, E, G, I) and γ_n' (B, D, F, H, J) terms with respect to the number of terms in the fitted expression for the SCCS dihedral angle for five different cases of the fitting of the OLU phenolate-keto S_1. XXIX
C.5 Fitted torsional phases of the n SCCS dihedral angle parameters for each of the 100 m fittings for the (A, C, E, G) S_0 and (B, D, F, H) S_1 states of the OLU (A, B) phenolate-keto, (C, D) phenol-keto, (E, F) phenol-enolate and (G, H) phenolate-enolate. XXXI
C.4 Fitted torsional barriers of the n SCCS dihedral angle parameters for each of the 100 m fittings for the (A, C, E, G) S_0 and (B, D, F, H) S_1 states of the OLU (A, B) phenolate-keto, (C, D) phenol-keto, (E, F) phenol-enolate and (G, H) phenolate-enolate. XXXII
C.6 Fitted torsional barriers of the n SCCS dihedral angle parameters with respect to their respective initial guesses for the (A, C, E, G) S_0 and (B, D, F, H) S_1 states of the OLU (A, B) phenolate-keto, (C, D) phenol-keto, (E, F) phenol-enolate and (G, H) phenolate-enolate. XXXIII
C.7 Fitted torsional phases of the n SCCS dihedral angle parameters with respect to their respective initial guesses for the (A, C, E, G) S_0 and (B, D, F, H) S_1 states of the OLU (A, B) phenolate-keto, (C, D) phenol-keto, (E, F) phenol-enolate and (G, H) phenolate-enolate. XXXIV
C.8 Periodicity of the torsional functions of the (A) S_0 and (B) S_1 OLU tautomers outside of the fitted domain. XXXV
C.9 Evolution of the SCCS dihedral angle of the OLU phenol-enolate in the S_0 along a QM MD simulation starting from two different dihedral angle values. XXXV
C.10 Infrared spectra of the OLU (A, B) phenol-keto, (C, D) phenol-enolate and (E, F) phenolate-enolate of the (A, C, E) S_0 and (B, D, F) S_1 states calculated from the QM geometry optimization and the MD simulations using the GAFF and the parametrised FF. XXXVI

D.1 RMSD of the protein for the MD trajectories with the four OLU chromophores in the (A) S_0 and (B) S_1 states. XXXVIII
D.2 RMSD of each of the OLU chromophores in the (A) S_0 and (B) S_1 states in aqueous media. XXXIX

D.3 (A) SCCS dihedral angle of the phenol-keto OLU during the first 300 ns of the MD simulation in the S_0 state. (B-H) Different distances along the simulation. (B) Distances from the O atom of one water molecule to the O atoms of OLU, SER349 and AMP. (C) Distances from the hydrogen bond (HB) donor of SER349 to different HB acceptors of the chromophore. (D,E) Distances from the O atom of two different water molecules to the N atoms of the OLU and the HB acceptors of SER349. (F) Distances from the O atom of of the tz moiety to the lateral chain of GLY318 and GLY343. (G) Distances from the O atom of of the btz moiety to different atoms of SER316 and LYS531 labelled as in the FF. (H) Distances from the C_α of the tz moiety of the OLU to the P atom of the AMP. Vertical-doted lines stress the relevant regions along the simulation and the darkest one represents the conformational change. Horizontal-doted lines represent the HB threshold of 3 Å. Taking into account that for some distances there is a H-O bond in between the O-O distance, this line is shifted to 3.98 Å when required. XLI

D.4 Representative snapshots of the phenol-keto OLU MD simulation in the S_0 during the first 300 ns. XLII

E.1 Convergence of the sampling of PCs in the 90º conformation. The colour bar represents the probability distribution of the projection. Part 1 of 7. XLV

E.2 Convergence of the sampling of PCs in the 90º conformation. The colour bar represents the probability distribution of the projection. Part 2 of 7. XLVI

E.3 Convergence of the sampling of PCs in the 90º conformation. The colour bar represents the probability distribution of the projection. Part 3 of 7. XLVII

E.4 Convergence of the sampling of PCs in the 90º conformation. The colour bar represents the probability distribution of the projection. Part 4 of 7. XLVIII

E.5 Convergence of the sampling of PCs in the 90º conformation. The colour bar represents the probability distribution of the projection. Part 5 of 7. XLIX

E.6 Convergence of the sampling of PCs in the 90º conformation. The colour bar represents the probability distribution of the projection. Part 6 of 7. L

E.7 Convergence of the sampling of PCs in the 90º conformation. The colour bar represents the probability distribution of the projection. Part 7 of 7. LI
E.8 Convergence of the sampling of PCs in the s-trans conformation. The colour bar represents the probability distribution of the projection. Part 1 of 7. LII
E.9 Convergence of the sampling of PCs in the s-trans conformation. The colour bar represents the probability distribution of the projection. Part 2 of 7. LIII
E.10 Convergence of the sampling of PCs in the s-trans conformation. The colour bar represents the probability distribution of the projection. Part 3 of 7. LIV
E.11 Convergence of the sampling of PCs in the s-trans conformation. The colour bar represents the probability distribution of the projection. Part 4 of 7. LV
E.12 Convergence of the sampling of PCs in the s-trans conformation. The colour bar represents the probability distribution of the projection. Part 5 of 7. LVI
E.13 Convergence of the sampling of PCs in the s-trans conformation. The colour bar represents the probability distribution of the projection. Part 6 of 7. LVII
E.14 Convergence of the sampling of PCs in the s-trans conformation. The colour bar represents the probability distribution of the projection. Part 7 of 7. LVIII
E.15 Convergence of the sampling of PCs in the 270º conformation. The colour bar represents the probability distribution of the projection. Part 1 of 7. LIX
E.16 Convergence of the sampling of PCs in the 270º conformation. The colour bar represents the probability distribution of the projection. Part 2 of 7. LX
E.17 Convergence of the sampling of PCs in the 270º conformation. The colour bar represents the probability distribution of the projection. Part 3 of 7. LXI
E.18 Convergence of the sampling of PCs in the 270º conformation. The colour bar represents the probability distribution of the projection. Part 4 of 7. LXII
E.19 Convergence of the sampling of PCs in the 270º conformation. The colour bar represents the probability distribution of the projection. Part 5 of 7. LXIII

E.20 Convergence of the sampling of PCs in the 270º conformation. The colour bar represents the probability distribution of the projection. Part 6 of 7. LXIV
E.21 Convergence of the sampling of PCs in the 270º conformation. The colour bar represents the probability distribution of the projection. Part 7 of 7. LXV
E.22 Convergence of the emission energy, the oscillator strength f and the CT number of (A) the 90º, (B) s-trans and (C) 270º conformations. LXVI
E.23 Variances of the PC of (A) the 90º, (B) s-trans and (C) 270º conformations. LXVII

F.1 Percentages of contribution to the total interaction energy of the EDA components relative to the sum of the absolute values of the three represented components calculated on the representative geometry. LXXI
F.3 (A-C) Convergence of the interaction energy and the EDA components and (D-F) error of the total interaction energy with respect to the calculation with twenty amino acids in the QM region as the number of amino acids in the QM region increases with the M062X functional and (A,D) 6-31+G(2d,p), (B,E) 6-31G(2d,p) and (C,F) 6-31G(2df,p) basis sets. LXXII
F.2 Percentages of contribution to the attractive terms of the interaction energy of the EDA components relative to the three represented attractive components calculated on the representative geometry. LXXIII
F.4 Number of transferred electrons between the two monomers (green) and averaged position of the orbital before and after the electronic transition (orange) along the MD trajectory of the (A) phenolate-keto, (B) phenol-keto, (C) phenol-enolate and (D) phenolate-enolate OLU. LXXIV
F.5 Convergence of the interaction energy and each of the EDA components with respect to the number of sampled geometries along the MD trajectory of the (A) phenol-keto, (B) phenol-enolate and (C) phenolate-enolate OLU. LXXV

G.1 RMSD of the (A) WT and (B) SER286ASN luciferases for each of the replicas of the MD trajectories with the phenolate-keto OLU in the S_1 state. LXXVI

G.2 (A) Temperature ramp and (B) RMSD of the SER286ASN luciferases during the simulation. (C) Probability distributions of the SCCS dihedral angles of the phenolate-keto OLU at two different temperatures. LXXVII
G.3 SCCS angle of the phenolate-keto OLU in the S_1 state along time inside (A) the WT and (B) the SER286ASN luciferases. Vertical dashed lines represent the separation into the three different replicas. LXXVIII

List of Tables

7.1 Averages and standard deviations of the 1A-MM-GBSA and EDA components distributions calculated from the FF and the QM/MM-EDA on 250 geometries. Percentages are relative to the sum of electrostatic and van der Waals terms (1A-MM-GBSA) and electrostatic and pauli+polarization (EDA). Energies are in kcal/mol. 241

7.2 Averages and standard deviations of the EDA components distributions calculated from 100 frames out of the 250 EDA calculations in which the amino acids PHE249 and SER349 were closest or farthest to OLU. Energies are in kcal/mol. 242

7.3 Averages and standard deviations of the MM-GBSA distributions calculated from the FF. Percentages are relative to the sum of electrostatic and van der Waals terms. Energies are in kcal/mol. 243

7.4 Averages and standard deviations of the EDA distributions calculated from the QM/MM calculations. Percentages are relative to the sum of electrostatic and pauli+polarization terms. Energies are in kcal/mol. 245

B.1 Atom types and charges of the OLU chromophore as represented in Figure B.1. VII

B.2 Stretching parameters for the OLU chromophore. Force constants appear in kcal/mol and bond lengths in Å. VIII

B.3 Bending parameters for the OLU chromophore. Force constants appear in kcal/mol/radian^2 and equilibrium angles in degrees. Part 1. IX

B.4 Bending parameters for the OLU chromophore. Force constants appear in kcal/mol/radian^2 and equilibrium angles in degrees. Part 2. X

B.5 Dihedral parameters for the OLU chromophore in the S_0. Torsional barriers and dividers appear in kcal/mol and phases in degrees. XI
B.6 Dihedral parameters for the OLU chromophore in the S_1. Torsional barriers and dividers appear in kcal/mol and phases in degrees. XII
B.7 Improper torsional parameters for the OLU chromophore. Torsional barriers and dividers appear in kcal/mol. The torsional barrier height is NOT divided by the divider. XIII
B.8 Non-bonding parameters for the OLU chromophore van der Waals radii appear in Å and 6-12 potential well depth in kcal/mol. XIII
B.9 Number of hydrogen bonds between the O atom of the tz and btz fragments of the OLU and the surrounding water molecules. XVIII
B.10 Number of hydrogen bonds between the O atom of the tz and btz fragments of the OLU and the surrounding protein. XVIII
B.11 Number of hydrogen bonds between the O atom of the tz and btz fragments of the OLU and its environment. XIX

E.1 Frequencies, in cm^{-1}, calculated from the eigenvalues of the PC analysis. LXVIII

List of Schemes

1.1 Decomposition of a four-membered ring dioxetanone. 35
1.2 Difference between the CIEEL and the CTIL mechanisms. . . . 36
1.3 Proposed biosynthetic route of D-luciferin. 41
1.4 Proposed storage mechanism of D-luciferin. 42
1.5 D-luciferin BL reaction. 43
1.6 Chemical structure of all the possible OLU tautomers. Those structures that are not possible emitters appear in orange. . . . 44
1.7 D-luciferin dark reaction pathway. 45
1.8 Proposed recycling mechanism of D-luciferin. 46
1.9 Schematic representation of the btz and tz fragments, respectively in blue and in pink, on the structure of the phenolate-keto OLU. The SCCS dihedral angle is highlighted in orange. 47
1.10 Synthetic derivatives of D-luciferin and their emission wavelengths in the WT luciferase enzyme. 50

2.1 Calculation of the dipole derivative autocorrelation function. t represents the internal time of each of the trajectory fragments while τ is the time displacement of each of the trajectory fragments with respect to the beginning of the simulation. 119

4.1 Computational steps followed during the parametrisation process. 176

C.1 Convergence process. XXVIII

List of Listings

C.1 Example of the input file to generate the parameters and initial structure to perform the clasical MD simulations. In this example, *leaprc.custom* contains the information about the atom types of the system and *OLU.mol2*, the structure and charges of the OLU chromophore. XXII
C.2 Example of the input file to the initial minimisation of the structure to perform the clasical MD simulations. XXIII
C.3 Example of the input file for the heatting of the minimised structure. XXIV
C.4 Example of the input file to the production of the MD simulation.XXV

List of Abbreviations

3A three-average

1A one-average

4CL 4-Coumarate-CoA Ligase

ABC-D ATP-Binding Cassette protein D subfamily

ACOT Acyl-CoA Thioesterase

ADP adenosine diphosphate

AIMD *Ab Initio* Molecular Dynamics

ALA Alanine

AMP adenosine monophosphate

ARG Arginine

ASN Asparagine

ASP Aspartic Acid

ATP adenosine triphosphate

BCE Before Common Era

BL Bioluminescence

BO Born–Oppenheimer

BSSE Basis Set Superposition Error

btz benzothiazole

CCD Charge-Coupled Device

CGTO Contracted Gaussian Type Orbitals

CIEEL Chemically Induced Electron-Exchange Luminescence

CoA Coenzyme A

CP Counterpoise

CT Charge Transfer

CTIL Charge Transfer-Induced Luminescence

CYS Cysteine

DFT Density Functional Theory

DOF degrees of freedom

DZ Double Zeta

EDA Energy Decomposition Analysis

ESPT Excited State Proton Transfer

FF Force Field

FWHM Full Width at Half Maxima

FMN flavin mono-nucleotide

GAFF General Amber Force Field

GB Generalized Born

GDIIS Geometry Direct Inversion in the Iterative Subspace

GGA Generalized Gradient Approximation

GLN Glutamine

GLU Glutamic Acid

GLY Glycine

GRCTIL Gradually Reversible Charge Transfer Induced Luminescence

GTO Gaussian Type Orbitals

HEI High-Energy Intermediate

HIE neutral Histidine protonated at N-ϵ

HIS Histidine

HO Harmonic Oscillator

HOMO Highest Occupied Molecular Orbital

HF Hartree-Fock

HPPD 4-Hydroxyphenylpyruvate Dioxygenase

ILE Isoleucine

IR Infrared

KS Kohn-Sham

LEU Leucine

LCAO Linear Combination of Atomic Orbitals

LCI Luciferase Complementation Imaging

LDA Local Density Approximation

LR Linear Response

LRE Luciferin Regenerating Enzyme

LSDA Local Spin Density Approximation

LST luciferin sulfotransferase

LUMO Lowest Unoccupied Molecular Orbital

LYS Lysine

MC Monte Carlo

MD Molecular Dynamics

MET Methionine

MM Molecular Mechanics

MO Molecular Orbital

MP2	Møller–Plesset second-order Perturbation theory
mRNA	messenger ribonucleic acid
MSK	Merz–Singh–Kollman
MW	Mass-Weighted
NMR	Nuclear Magnetic Resonance
NO	nitric oxide
OLU	Oxyluciferin
OLUFF	OxyLUciferin Force Field
PAPS	3'-phosphoadenosine 5'-phosphosulfate
PAPSS	PAPS Synthase
PBC	Periodic Boundary Conditions
PC	Principal Components
PCA	Protein–Fragment Complementation Assay
PDB	Protein Data Bank
Pxmp2	Peroxisomal Membrane Protein 2
PES	Potential Energy Surface
PGTO	Primitive Gaussian Type Orbitals
PHE	Phenylalanine
PO	Phenoloxidase
PPi	Pyrophosphate
PRO	Proline
QM	Quantum Mechanics
QM/MM	Quantum-Mechanics/Molecular-Mechanics
R^2	Least Squares Difference
RAE	Real Academia Española
RDF	Radial Distribution Functions
RHF	Restricted Hartree Fock
RMSD	Root Mean Square Deviation
RMSF	Root Mean Square Fluctuation
RNA	ribonucleic acid
S_0	ground electronic state
S_1	first electronically excited state
S_3	Third Electronically Excited State
SA	Surface Area
SASA	Solvent-Accessible Surface Area
SCF	Self Consistent Field
ScpX	Sterol carrier protein-X
SD	Slater determinant
SE	Schrödinger Equation

SEC Selenocysteine

SER Serine

SET Single Electron Transfer

S_N2 bimolecular nucleophilic substitution

SP Single Point

STO Slater Type Orbitals

SULF Sulfatase

TD Time Dependent

TDA Tamm–Dancoff Approximation

THR Threonine

tz thiazole

TICT Twisted Intramolecular Charge Transfer

TISE Time Independent Schrödinger Equation

TRP Tryptophan

TS Transition State

TYR Tyrosine

TAT Tyrosine Aminotransferase

TZ Triple Zeta

VAL Valine

WT Wild-Type

XC Exchange Correlation

Abstract

Nature presents a wide variety of bioluminescent systems resulting from a substrate/enzyme interaction, luciferin/luciferase, leading to a derivative of the substrate in an electronically excited state that is able to emit light in the visible region (350-750 nm) during the relaxation process. Nowadays, nine natural luciferines have been documented together with their complementary luciferases giving rise to more than thirty bioluminescent systems. The first one of them to be discovered was D-luciferin, present in some species of insects, as fireflies, and first described in 1957 by Bitler and McElroy.

D-luciferin bioluminescent system is present in different species of animals. For each of them, the luciferase enzyme that catalyses the bioluminescence reaction presents some structural differences that induce shifts in the emission wavelength. In this line, understanding the factors that lead to the colour modulation can allow for the development of more specialized tools for research, medicine and analytic chemistry. For this reason, the present thesis focuses on modelling at an atomistic scale this bioluminiscent system, formed by firefly's oxyluciferin/luciferase complex. Firstly, classical molecular dynamics based on parameters already present in literature were tested against simulations employing quantum mechanics/molecular mechanics (QM/MM) potentials in order to determine whether or not these parameters are suitable for the sampling of the system based on structural features and electronic properties. Attending to these results, the OxyLUciferin Force Field (OLUFF) has been parametrised to generate samplings that reproduce the QM properties of the system in two electronic states, the ground and first excited states, of the four possible oxyluciferin tautomers.

Then, the OLUFF was employed to generate samplings of the system in different media by means of molecular dynamics simulations. These trajectories were analysed and compared to each other and to experimental results taken from literature. The four emitting oxyluciferin species were studied in aqueous media and inside the active site of the wild-type luciferase enzyme in order to characterise the structural differences between them in the ground

and first excited electronic states. Moreover, for each of them, the oxyluciferin/luciferase interaction energy was computed and characterised at both classical and QM/MM level in the excited state. Furthermore, the phenolate-keto emitter, which is considered to be the red-emitting chromophore, was employed to characterise the nuclear motion that leads to a red-shifted emission. In addition, it was also used to study the structural differences of the oxyluciferin complexed with wild-type luciferase and the SER286ASN mutant, a mutation that has been reported to emit in the red region of the electromagnetic spectrum. Upon the analysis of these calculations, it can be concluded that the torsion around the single C-C' bond of the chromophore plays a crucial role in the red-shift mechanism of D-luciferin bioluminescent system. Although typically disregarded, twisted conformations of this degree of freedom can be sampled in the electronically excited state and, therefore, they should be considered as plausible emitting states of the oxyluciferin. Moreover, the oxyluciferine/luciferase interaction in the electronically excited state is governed by the electrostatic term. Nonetheless, non-electrostatic contributions are not negligible due to the presence of amino acids with π systems inside the luciferase pocket. Suggesting protein mutations depending on the nature of the lateral chain of the amino acids is a challenging task, as it would introduce changes in the oxyluciferin binding but also in the structure and size of the active site and in the number of water molecules participating in the stabilization of the pocket.

To summarise, this PhD thesis employs a dynamical approach to provide a thorough computational characterization of the oxyluciferin/luciferase complex in the electronically excited state, describing not only the structure of the complex but also characterising the electronic properties of the emission phenomena and the interaction energy. Moreover, the OLUFF has been proposed, which can be used to sample both the ground and the first electronically excited state of the chromophore in order to further investigate D-luciferin bioluminescent system from a dynamical perspective, which has been shown to play a crucial role on the behaviour of the system.

Keywords: bioluminescence, oxyluciferine, computational modelling, molecular dynamics, density functional theory, structure, energy decomposition analysis, force fields and photochemistry.

Resumen

La naturaleza presenta una gran variedad de sistemas bioluminiscentes. Éstos resultan de la interacción enzima-sustrato, luciferasa/luciferina, que da como resultado la oxidación del sustrato a la molécula de oxiluciferina en estado electrónico excitado que, al relajarse al estado fundamental, emite luz en la región del visible (350-750 nm). Hasta la fecha, se han documentado 9 luciferinas naturales y decenas de luciferasas complementarias, cuyas diferentes combinaciones dan lugar a más de 30 sistemas bioluminiscentes. La primera luciferina, la D-luciferina, fue descrita por McElroy en 1957.

Concretamente, el sistema bioluminiscente de la D-luciferina está presente en diferentes especies animales. En cada una de ellas, la enzima luciferasa que cataliza la reacción bioluminiscente presenta algunas diferencias estructurales que provocan cambios en la longitud de onda de emisión. En este sentido, comprender los factores que provocan la modulación del color puede permitir el desarrollo de herramientas más especializadas para la investigación, la medicina y la química analítica. Por esta razón, la presente tesis se centra en modelar a escala atomística este sistema bioluminiscente, formado por el complejo oxiluciferina/luciferasa de las luciérnagas. En primer lugar, se compararon las dinámicas moleculares clásicas basadas en parámetros ya presentes en la literatura con simulaciones que empleaban potenciales de mecánica cuántica/mecánica molecular (QM/MM) con el fin de determinar si estos parámetros son adecuados para el muestreo del sistema basándose en características estructurales y propiedades electrónicas. Atendiendo a estos resultados, se parametrizó un campo de fuerzas para la oxiluciferina (OLUFF, por sus siglas en inglés) para generar muestreos que reprodujeran las propiedades cuánticas del sistema en dos estados electrónicos, el fundamental y el primer estado excitado, de los cuatro tautómeros posibles de la oxiluciferina.

A continuación, se utilizó el OLUFF para generar muestreos del sistema en diferentes medios mediante simulaciones de dinámica molecular, que se analizaron y compararon entre sí y con los resultados experimentales obtenidos de la bibliografía. Las cuatro especies emisoras de oxiluciferina se estudiaron

tanto en medio acuoso como dentro del sitio activo de la enzima luciferasa natural con el fin de caracterizar las diferencias estructurales entre ellas en los estados electrónicos fundamental y primer estado excitado. Además, para cada una de ellas, se estudió la interacción oxiluciferina/luciferasa tanto en el estado excitado tanto a nivel clásico como QM/MM. Finalmente, se empleó el emisor fenolato-ceto, considerado el cromóforo emisor de la luz roja, para caracterizar el movimiento nuclear que conduce a una emisión desplazada hacia el rojo. Este cromóforo también se utilizó para estudiar las diferencias estructurales de la oxiluciferina complejada con la luciferasa natural y el mutante SER286ASN que, según se ha descrito en la literatura, emite en la región del rojo del espectro electromagnético. Tras analizar estos cálculos, se pudo concluir que la torsión alrededor del enlace C-C' del cromóforo desempeña un papel crucial en el mecanismo de desplazamiento al rojo del sistema bioluminiscente de la D-luciferina. Las conformaciones deformadas de este grado de libertad pueden muestrearse en el estado electrónico excitado y, por lo tanto, deben considerarse como estados emisores plausibles de la oxiluciferina. Además, la interacción entre la oxiluciferina y la luciferasa en el estado de electrónico excitado está gobernada por el término electrostático. No obstante, las interacciones no electrostáticas también representan una contribución relevante debido a la presencia de aminoácidos con sistemas π dentro del sitio activo de la luciferasa. Sugerir mutaciones proteicas en función de la naturaleza de la cadena lateral de los aminoácidos es una tarea difícil, ya que no solo se introducen cambios en la unión de la oxiluciferina sino que también se alteraran la estructura y el tamaño del sitio activo y en el número de moléculas de agua que participan en la estabilización del centro activo.

En resumen, esta tesis doctoral proporciona, mediante un enfoque dinámico, una caracterización computacional exhaustiva del complejo oxiluciferina/luciferasa en el estado de electrónico excitado, describiendo no solo la estructura del complejo, sino también caracterizando las propiedades electrónicas de los fenómenos de emisión y la energía de interacción. Además, se ha propuesto el OLUFF, que puede utilizarse para muestrear tanto el estado fundamental como el primer estado electrónico excitado del cromóforo con el fin de investigar más en profundidad el sistema bioluminiscente D-luciferina desde una perspectiva dinámica, que ha demostrado desempeñar un papel crucial en el comportamiento del sistema.

Palabras clave: bioluminiscencia, oxiluciferina, modelización computacional, dinámica molecular y teoría del funcional de la densidad, estructura, análisis de descomposición de la energía, campos de fuerza y fotoquímica

Chapter I

Introduction

1.1 Bioluminescence

It is no secret that human beings have always been fascinated by shining physical processes. Even thought the most clear example of this is the fast control of fire, it is not the only light-emitting process that has dazzled humanity from its very beginning. Bioluminescence (BL) is a spectacular phenomenon involving light emission produced by living organisms as a result of a biochemical reaction. This process is generally caused by interactions between an enzyme protein called luciferase and its substrate, luciferin, (Figure 1.1). The former catalyses the oxidation of luciferin by molecular oxygen, leading to the transformation of the luciferase into an Oxyluciferin (OLU) molecule in an excited state which, upon radiative relaxation, emits light in the visible region.[1, 2]

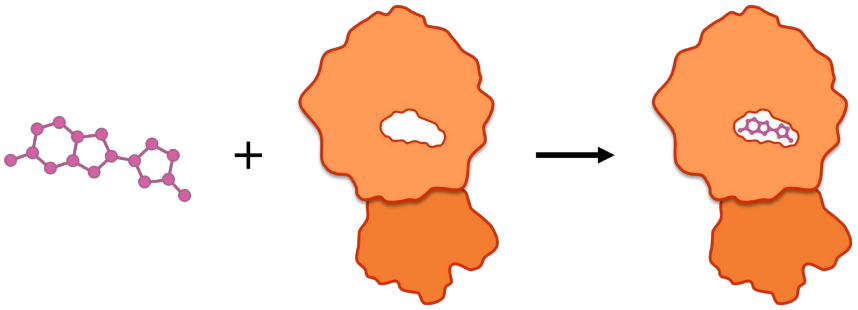


Figure 1.1: Schematic representation of complex formation from the luciferin chromophore (pink) and the luciferase (orange).

Already in the 15$^{\text{th}}$ century Before Common Era (BCE), BL phenomena were reported in Chinese literature.[3] This phenomenon did not only impress this specific oriental civilization but also many different cultures around the globe, as reflected by many folklore-based and artistic works like poems and paintings from the people of ancient Japan, India, Africa as well as Central and South America. Nonetheless, there is no mention of any effort being directed at understanding or applying the knowledge of such phenomena until the 3$^{\text{th}}$ century BCE, when Aristotle not only described up to 180 BL animals, but also first identified the so-called "cold light" after realizing that, unlike the light from a flame, the self-luminosity of BL organisms was not accompanied by the heat of combustion or incandescence.

During the following centuries, more and more BL creatures were reported, but a real desire of understanding BL only appeared when a more scientific approach was promulgated by 17$^{\text{th}}$ century philosophers.[1] Most probably, the scientific understanding of this dazzling phenomenon began in 1668, when Robert Boyle[4] identified that the BL of wood and glow-worms required air

to take place (oxygen was yet to be identified as an element and a significant component of air).[5] Furthermore, only after the invention of the microscope, Marcello Malpighi was able to first associate the light emitting material in fireflies with particular granules in the cell in 1688.[6] In the following two centuries and thanks to the development of more sophisticated equipment and techniques, BL processes began to be studied in a rational and enlightened way, leading to the corroboration in 1885 by Raphael Dubois of the idea that BL was the result of a chemical reaction.[7] This scientist also coined the current luciferin and luciferase generic terminology, elucidating that the heat stable part was probably an organic molecule, which he named *luciferine*, while the active ingredient was heat labile since it had to be kept cool for an effective reaction. This last part was labelled *luciferase* upon the discovery of enzymes.

In 1888, Eilhard Wiedemann proposed the use of the term "luminescence" to designate light emission that was more intense than would be expected from the source's temperature.[7] Moreover, he proposed a classification depending on the source of excitation, which was quickly adopted by other scientists. In this classification, light emission was divided into (i) fluorescence, which is the fast spin-allowed emission from an electronically excited state to the ground state; (ii) phosphorescence, the slow spin-forbidden emission from an electronically excited state to the ground state; (iii) electroluminescence, in which the electronic excitation is produced by an electric current; (iv) chemical luminescence, in which the excitation is due to a chemical reaction; (v) thermoluminescence, which is produced by heating but whose characteristics are not temperature dependent; (vi) triboluminescence, produced by friction; and (vii) crystal luminescence, a type of luminescence that accompanies crystallization processes.

Finally, during the last century, a scientific explanation was provided for the biochemical mechanisms of BL systems, together with the identification of numerous chemical partners involved in BL systems and the elucidation of the three-dimensional structure of the luciferase protein. Most of the contributions that were achieved during the first 20 years of the 20$^{\text{th}}$ century were almost exclusively those of Edmond N. Harvey at Princeton University, who is considered to be the father of the modern BL.[7–9] During the second half of the 20$^{\text{th}}$ century, the research on BL was carried out mainly, but not exclusively, by his students Frank H. Johnson, who began his activity in 1949; William D. McElroy and Marlene DeLuca in 1951 and Milton J. Cormier and Frederick I. Tsuji in 1954 and his collaborator Emil H. White.[3] Furthermore, from that moment on, a tremendous outburst in the understanding of BL systems has taken place. Among many others, it has been possible to synthesize and characterize the structure[10, 11] and properties[12] of differ-

ent luciferines; to extract, purify and isolate the three-dimensional crystalline structures of several BL enzymes;[13–17] to gain deeper insights on the BL reaction mechanisms[12, 18–22] and the essential requirement of different cofactors like adenosine triphosphate (ATP);[23] to determine luciferin analogues with the aim of switching the light emission colours[24, 25] and to clone and express different luciferases.[26–30]

1.1.1 Emitting Species

It is clear that the 180 animals described by Aristotle in the IV century BCE only represents a very small fragment of the known BL species. Specially because he focused mostly on marine species and some insects. In fact, in 2019 about 10,000 species from 800 genera with the ability to emit light were already described, although this number might be as well an underestimation.[31] Furthermore, despite their wide range of applications, which will be further discussed in Section 1.2.4 for fireflies' BL system, most of these BL systems are largely understudied since neither the molecular components of the light-emitting reactions nor the full biochemical pathway leading to light emission is understood.

Nowadays, nine natural luciferines have been documented together with their complementary luciferases leading to more than forty different BL systems.[2] As luciferin chromophores are the principal determinants of colour and properties of light-emitting reactions, BL systems are typically categorized by them, as shown in Figure 1.2.

The coelenterazine chromophore,[32] represented in Figure 1.2(A), is a modified tripeptide formed by one PHE and two Tyrosine (TYR) amino acids with an unknown biosynthetic route. This luciferin is the substrate for a large amount of luciferases found in marine ecosystems. However, most of these organisms do not synthesise coelenterazine themselves. Instead, they obtain it from their diet. This factor leads to the convergence of independently evolved luciferases in groups of organisms coming from very distant branches of the phylogenetic tree, resulting on very diverse organisms with system-dependant characteristics such as molecular weight, pH-sensitivity, thermostability and catalysis rate. Nonetheless, all natural systems based on the coelenterazine chromophore emit blue light between 450–500 nm, which lies within the visible region of the electromagnetic spectrum. Furthermore, they do not typically require any cofactors for the BL process to occur. In addition, some systems present a fluorescent protein that interacts with the luciferase, altering the colour of BL.[31] Coelenterazine-based BL systems are very popular for biomedical applications such as bioimaging and drug-screening[33, 34], mon-

itoring tumour progression[35] and drug response and BL imaging *in vivo* of either single-cell or whole-body.[36]

Like in the case of the coelenterazine, the cypridina chromophore shown in Figure 1.2(B), a blue-emitting (445–465 nm)[2] modified tripeptide formed by Tryptophan (TRP), Isoleucine (ILE) and Arginine (ARG),[37] is a marine

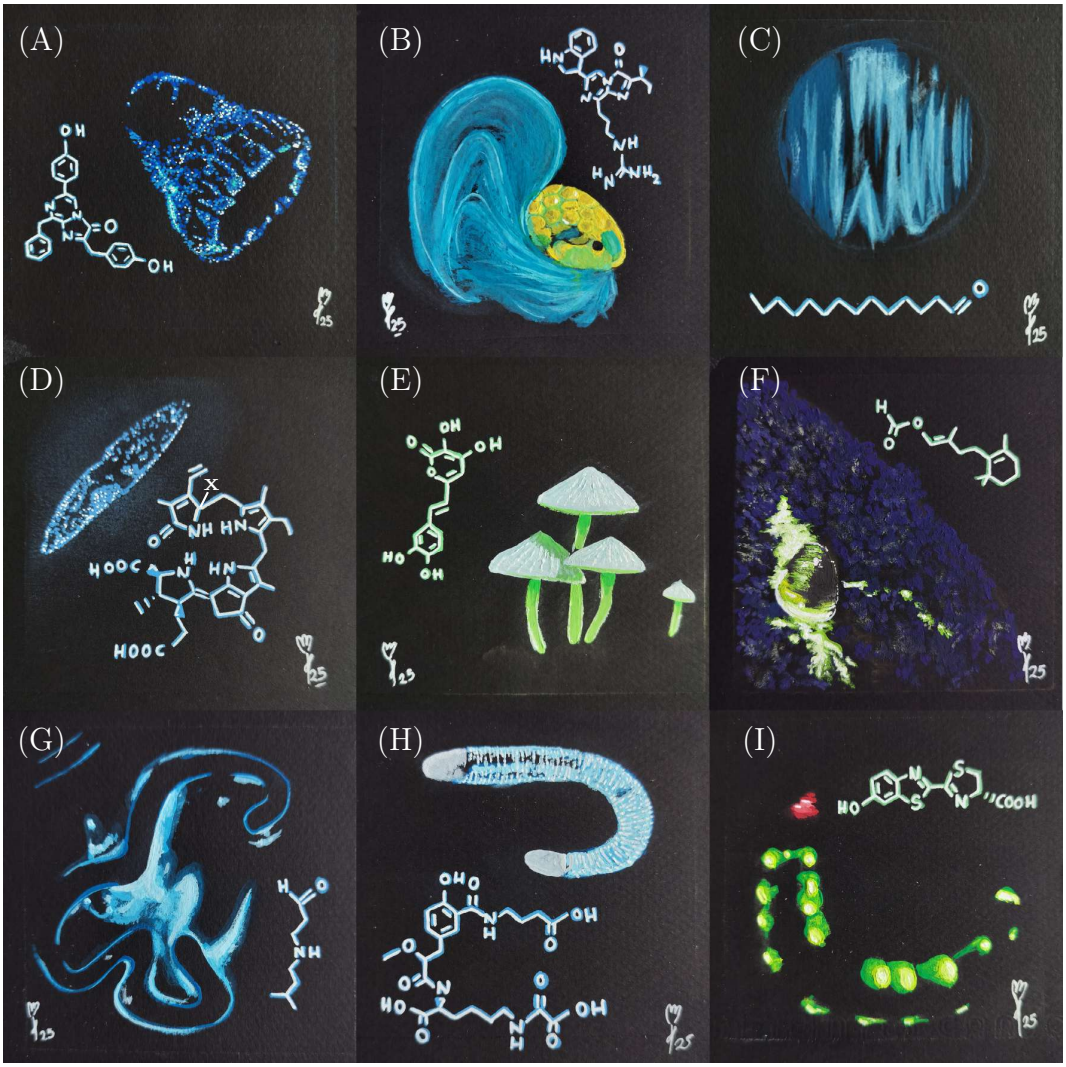


Figure 1.2: Representation of the nine natural luciferines with a representative BL organism: (A) coelenterazine, (B) cypridina luciferine, (C) bacterial luciferin (D) tetrapyrrole-based luciferine with X=H (dinoflagellates) and X=OH (krill), (E) fungal luciferin, (F) Latia luciferin, (G) Diplocardia luciferin, (H) Fridericia luciferin and (I) D-luciferin.

luciferin found in the ostracod Cypridina and BL midshipman fish, *Porichthys* genera. This BL system has been widely used in studies of circadian rhythms [38–40] and in immunoassays.[41–43]

In the case of the bacterial BL system, the light source comes from a flavin mono-nucleotide (FMN) derivative. Nonetheless, the luciferin, represented in Figure 1.2(C), is considered to be the myristic aldehyde, which is oxidised during the BL reaction catalysed by a luciferase formed by complex of two polypeptide chains. In most cases, the colour of the emitted light is blue (around 490 nm). However, it is possible to find natural BL with red-shifted wavelengths.[44] The study of antimicrobial drugs, bacterial infections and environmental monitoring are among the main applications of bacterial BL system, which employ living bacteria as a light source.[45]

Dinoflagellates, a group of single-celled eukaryotes considered protists (term which refers to all eukaryotes that do not belong to the three traditional eukaryotic kingdoms: animals, land plants or fungi), and *Euphausiids*, marine crustaceans of the order *Euphausiacea* generally know as krill, present a BL system based on two very similar tetrapyrrole luciferines (Figure 1.2(D)),[46–51] whose mechanism have not been fully studied.[2] In the former, BL occurs inside specific cytoplasmic organelles, the scintillons,[52] as localised flashes of light triggered by electrical or mechanical stimulation while, in the latter, the light emission takes place in complex and specialized organs with lenses and the ability to focus. This particular BL system is rarely employed in research due to the unavailability of a synthetic luciferin.[31]

Fungi utilise the simple α-pyrone 3-hydroxyhispidin[53] shown in Figure 1.2(E), that is oxidised by an insoluble luciferase in a reaction that results in the emission of green light (around 530 nm). This BL system, which is preserved for all the 70 fungal BL species distributed across 9 genera,[54] still presents a significant amount of unknowns like the luciferase cellular localization *in vivo*. However, the protein has a predicted N-terminal transmembrane helix, consistent with its insolubility.[55, 56] Furthermore, very recently, it was found that caffeic acid is the starting substrate in fungal luciferin synthesis,[56] thus, broadening the range of possible applications since it is naturally found in all plants.

A least spread BL system is that of the New Zealand limpet *Latia neritoides* represented in Figure 1.2(F). The Latia luciferin is a formate of the terpene aldehyde in the enolic form that emits yellow-green light (536 nm) during the BL process.[57, 58] Although it is known that the presence of both oxygen and the so-called "purple protein", a red fluorescent protein acting as a cofactor, is required for the Latia BL to take place, there are still many unknowns about this system. For instance, the Latia luciferin only acts as an energy source for

the BL reaction. However, the actual component responsible for light emission has not yet been detected. This fact, in addition to the unresolved structure of the purple protein, limits the development of methods for its application.[2]

Another atypical BL system can be found in the annelid *Diplocardia longa*, whose luciferin structure consists on a N-isovaleryl-3-aminopropanal, as shown in Figure 1.2(G). This BL system is also present in other 13 earthworm species belonging to six genera with emission wavelengths between 500 and 570 nm.[59] Interestingly, the BL reaction includes hydrogen peroxide as a reactant[60] and, therefore, it can be used as an analytical method to detect it.[61, 62] However, its utilization is challenging due to the limited availability and low stability of the luciferase and the inhibition produced by strong chelating agents and organic buffers.[2]

A luciferin from another worm, the *Fridericia heliota*,[63, 64] has an unusual peptide consisting of a γ-aminobutyric acid, a modified TYR amino acid, a LYS amino acid and oxalic acid residues (Figure 1.2(H)). This luciferin, which produces a bright and constant emission at 478 nm, is non-toxic and very stable. Furthermore, and similarly to firefly luciferin as will be discussed in Section 1.2, it requires ATP for its BL reaction. Therefore, it could be a suitable candidate for a wide range applications, once the *Fridericia* luciferase is purified, sequenced and cloned.

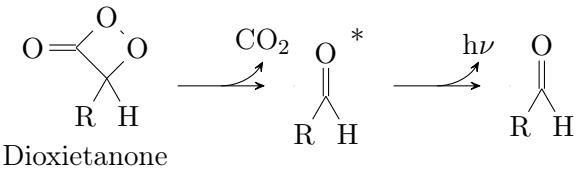
Ultimately, the most thoroughly studied BL system, and the one to which this thesis is dedicated, is that based on (S)-2-(6'-hydroxy-2'-benzothiazolyl)-2-thiazoline-4-carboxylic acid, in Figure 1.2(I). This chromophore, usually called D-luciferin, consists on a benzothiazole (btz) moiety with a hydroxy group at the 6' position covalently bonded to a thiazole (tz) group with a carboxylic acid at position 4, analogously to its emitter shown in Scheme 1.9. This structure was reported by E. H. White, F. McCapra, G. Field, and W. D. McElroy in 1961.[3, 10] D-luciferin BL system is present in several lineages of beetles, including fireflies, click beetles and railroad worms. The chain of BL reactions starting from this stable and non-toxic chromophore leads to the emission of yellow, orange and in some cases red light, depending on the structure of the luciferase and pH. Attending to this, it is possible to distinguish three big groups of natural luciferases:[31] (i) the one present in fireflies (*Lampyridae* family), which is the most known luciferase and emits between 540 and 570 nm; (ii) luciferases produced by click beetles (*Elateridae* family), whose emission lies between 540 and 600 nm, and (iii) luciferases expressed by railroad worms of the *Phengodidae* family, characterised by the expression of two different luciferases in different areas of their bodies allowing them to emit in two different colours: red (630 nm) and green-yellow (540 nm).[2] The existing dependence of light emission on the concentration of ATP

has allowed to use the system for a wide variety of applications that will be further discussed in Section 1.2.4.

1.1.2 Light Generation in Bioluminescent Systems

As mentioned before, the BL emission is the final process of a multistep chemical reaction, in which the last three steps consists on (i) the formation of a High-Energy Intermediate (HEI), (ii) the generation of a molecule in an electronically excited state with a lifetime in the range of nanoseconds and (iii) the emission of a photon in the visible region.

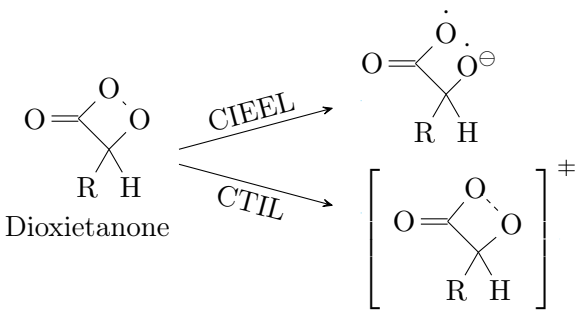
A very important and interesting step of this biochemical reaction is, precisely, the generation of the electronically excited intermediate. It happens upon the decomposition of the HEI four-membered ring dioxetanone to form CO_2 and an excited carbonyl compound,[52] as represented in Scheme 1.1. During the so-called chemiexcitation process, the cleavage of the O–O bond of this strained and energy-rich peroxide releases a large chemical reaction energy, which leads to the generation of the carbonyl compound in an excited state as consequence of a fast cascade of processes.



Scheme 1.1: Decomposition of a four-membered ring dioxetanone.

Originally, it was believed that this process followed a Chemically Induced Electron-Exchange Luminescence (CIEEL) mechanism,[65–70] which is initiated by the electron transfer from an electron donor to the O–O bond of the cyclic peroxide, forming a neutral radical/anionic radical pair (Scheme 1.2). Then, a subsequent back electron transfer between the radicals would release enough energy to promote the chromophore to its first electronically excited state (S_1). This mechanism, which was proposed in the 1970's, is, however, in doubt since it is not clear whether or not the back electron transfer can take place.[71–75] Moreover, the efficiency of this process between the radicals in a solvent cage is highly questioned.[73, 75]

To overcome this problem, the Charge Transfer-Induced Luminescence (CTIL) mechanism was proposed as a variation of the CIEEL.[71, 72] As shown in Scheme 1.2, this newer mechanism consists on the rate-determining



Scheme 1.2: Difference between the CIEEL and the CTIL mechanisms.

endothermic formation of a Transition State (TS) with Charge Transfer (CT) characteristics leading to the excited-state product via a conical intersection.[65] The main two differences between these two mechanisms are: (i) the emphasis on the partial inter- or intra-CT of the CTIL mechanism rather than complete electron transfer to O–O leading to its cleavage and (ii) the formation of a molecular CT complex in the CTIL mechanism,[71] while the CIEEL mechanism assumes the generation of cation and anion radical fragments. Hence, the CIEEL requires a back electron transfer, which results in the annihilation of the generated radical ions and, simultaneously, an excited-state product. However, in the CTIL mechanism, the back CT process is not absolutely needed and the extent of the back CT occurrence depends on the micro-environment of the O–O bond.[76]

More recent studies employing theoretical chemistry calculations have determined that the decomposition mechanism is neither the CIEEL nor the CTIL.[5, 65, 77] In fact, depending on the nature of the chromophore and its protonation state, it can take place via an entropic trap or the Gradually Reversible Charge Transfer Induced Luminescence (GRCTIL) mechanism.[77–80] The former consists on the vibrational energy redistribution leading to a change in the momentum of the molecule from the O–C–C–O dihedral angle twist to a dissociative C–C stretch. As a result, the cleavage of the C–C bond only happens after the complete breakage of the O–O bond, thus involving two TSs separated by a biradical intermediate.[5, 81] During the decomposition process, the Potential Energy Surface (PES) of both the ground electronic state (S_0) and the S_1 state become close in energy, intersecting twice along the reaction path. As a result, the excited-state intermediates and the corresponding BL products are formed. On the flip side, in the GRCTIL mechanism the CT and back CT occur gradually and in a concerted manner with the ruptures of the O–O and C–C bonds, respectively, leading to a cross of between the

PESs of the S_0 and the S_1 electronic states.[78]

1.2 Fireflies Bioluminescence

Each of the BL systems discussed in the previous sections present their own specific characteristics and, although they are all interesting and worth studying, the present thesis focusses on D-luciferin BL in the Wild-Type (WT) luciferase of fireflies. Hence, in this section more specific details about this system will be discussed. Specifically, Section 1.2.1 focusses on firefly's life cycle and the biological processes that controls its BL, while Section 1.2.2 describes the biochemical cycle of D-luciferin from its *in vivo* biosynthesis and storage to the BL reaction and its regeneration. Moreover, Section 1.2.3 addresses the factors that influence the colour of emission and Section 1.2.4 presents the main biomedical applications of D-luciferin based BL systems.

1.2.1 Firefly's Life Cycle and Bioluminescent Phenotype

Nearly 2000 lampyrid species of fireflies are distributed across all continents except Antarctica, half of them occurring in tropical America and Asia.[82] However, it is possible that new undescribed species might be encountered due to the extreme geographical localization of some of them. Nonetheless, it is also possible that they become extinct before they are even discovered because of the large amount of industries and human settlements poisoning firefly's habitats. Sadly, the human presence is not only risking this insect's survival, but also that of many other species around the globe.

As represented in Figure 1.3(A), fireflies are holometabolous insects, meaning that they undergo a complete metamorphosis and, therefore, present four life stages: egg, larva, pupa, and adult. Eggs are deposited either a few at a time over several days or weeks or together at one time, depending on the species. In some cases, eggs present a dim ability to emit light. At the beginning, the glow comes from luminous material smeared in the mother's egg canal. As the light organs of the embryos develop, the glow can also come from within the eggs. Once the eggs hatch, firefly larvae start to explore their surroundings. During this stage, characterized by the specialization on hunting and eating soft-bodied preys, all fireflies are presumed to be luminescent, though this phenomenon has only been observed for a a few species. The biological function of egg and larval luminescence is believed to be a warning of unpleasant taste, with sometimes emetic or lethal consequences to prospective predators. When ready to pupate, larvae seek or make sheltered places, where

they transform into the adult form and breeding specialist. This pupae stage is a period of phenotype reconstruction, resulting in a body form adapted to a completely different lifestyle and range of activities. Adult fireflies are concerned only with reproduction. In nocturnal species, mating rituals involve light emission with controlled flash pattern specificity.

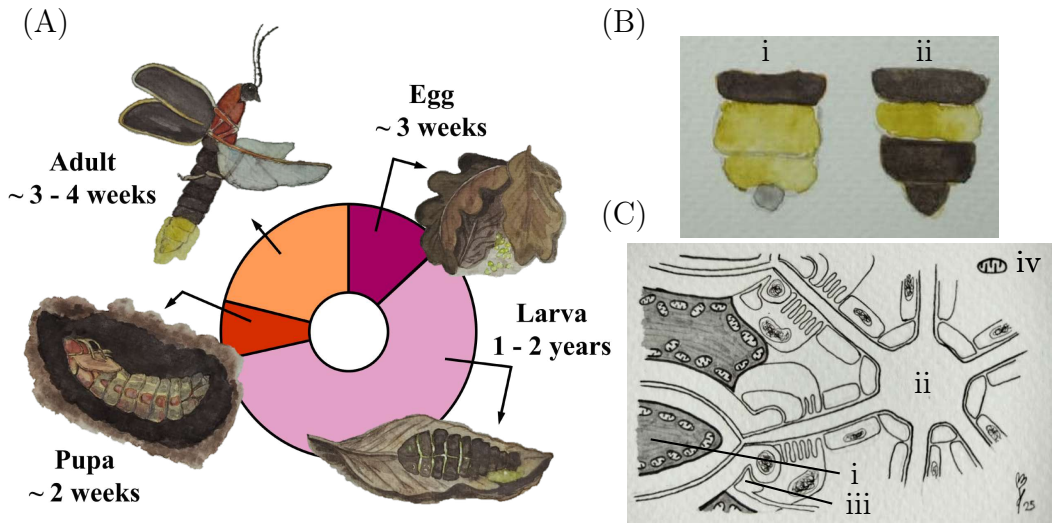


Figure 1.3: (A) *Photinus pyralis* life cycle. (B) Phenotypical differentiation between female (B.i) and male (B.ii) lanterns of *Photinus pyralis*. (C) Schematic representation of lampyrid lantern. (C).i-iv are described in text.

Firefly's BL is completely localized within their anatomy.[83] Though they present phenotypical differentiation between female and male sexes as shown in Figure 1.3(B), for both of them the light emission is produced at the lanterns, localized at the final abdominal segments. The physiology of the lanterns, represented in Figure 1.3(C), consists on rosettes of photocytes (Figure 1.3(C).i in grey), the specialized cells responsible for the BL surrounding a specialized tracheal system (Figure 1.3(C).ii). This tracheal system is part of the insect respiratory system and is in charge of the oxygen supply necessary for the oxidation step of the BL reaction, which is controlled by nerves (Figure 1.3(C).iii), thus leading to light flashes.

From a biological perspective, firefly's lanterns behave in the following summarized way.[83] In a first step, the lantern would receive the neural signal, switching the lantern on leading to the clearing off the tracheae and tracheoles of the tissue fluid that fills them during the inactive stage. This slow change in the media allows for a faster diffusion of oxygen to the photocytes. Nonetheless, this step along is not enough for the oxygen to efficiently reach

the luciferase enzyme. The photocytes present an unique structure in which, as represented in Figure 1.3(C), the mitochondria (Figure 1.3(C).iv) are deployed at the interface with the tracheal system thus acting as "gatekeepers" by absorbing and using up any incoming oxygen to synthesise ATP via oxidative phosphorylation. It is not until gaseous nitric oxide (NO) is quickly produced by the enzyme nitric oxide synthetase in response to calcium gradients at the cell surface that oxygen can actually overcome this mitochondrial barrier.[84, 85] NO is able to complex and inactivate the mitochondrial enzyme cytochrome c oxidase, responsible of the oxygen consumption; thus inhibiting the whole mitochondrial function. Therefore, the NO release allows oxygen to diffuse into the photocyte and reach the peroxisomes, in which the luciferase is located[86] and, therefore, where the BL reaction takes place. Apart from NO being labile and easily degraded, it has been observed that the inhibitory effect of NO on respiration can be reversed by competition with oxygen and by shining a bright light, thus the light emission itself could lead to the reactivation of the mitochondrial barrier and so to the production of short flashes of light.

1.2.2 D-Luciferin Biochemical Cycle

As mentioned before, the *in vivo* BL reaction takes place inside the peroxisomes of the photocytes situated in the insect's lantern. However, each of the individual components required for the biochemical reaction is synthesised elsewhere. For example, the luciferase enzyme, which is folded into a large N-terminal domain and a small C-terminal domain joined by a flexible linker peptide, is synthesized by free polysomes, groups of ribosomes bound to a messenger ribonucleic acid (mRNA) molecule, in the cytoplasm and, later, it is transported and fused into peroxisomes. This transport process is directed by the tripeptide motif SER-LYS-LEU, that is usually responsible for peroxisome targeting.[87–89]

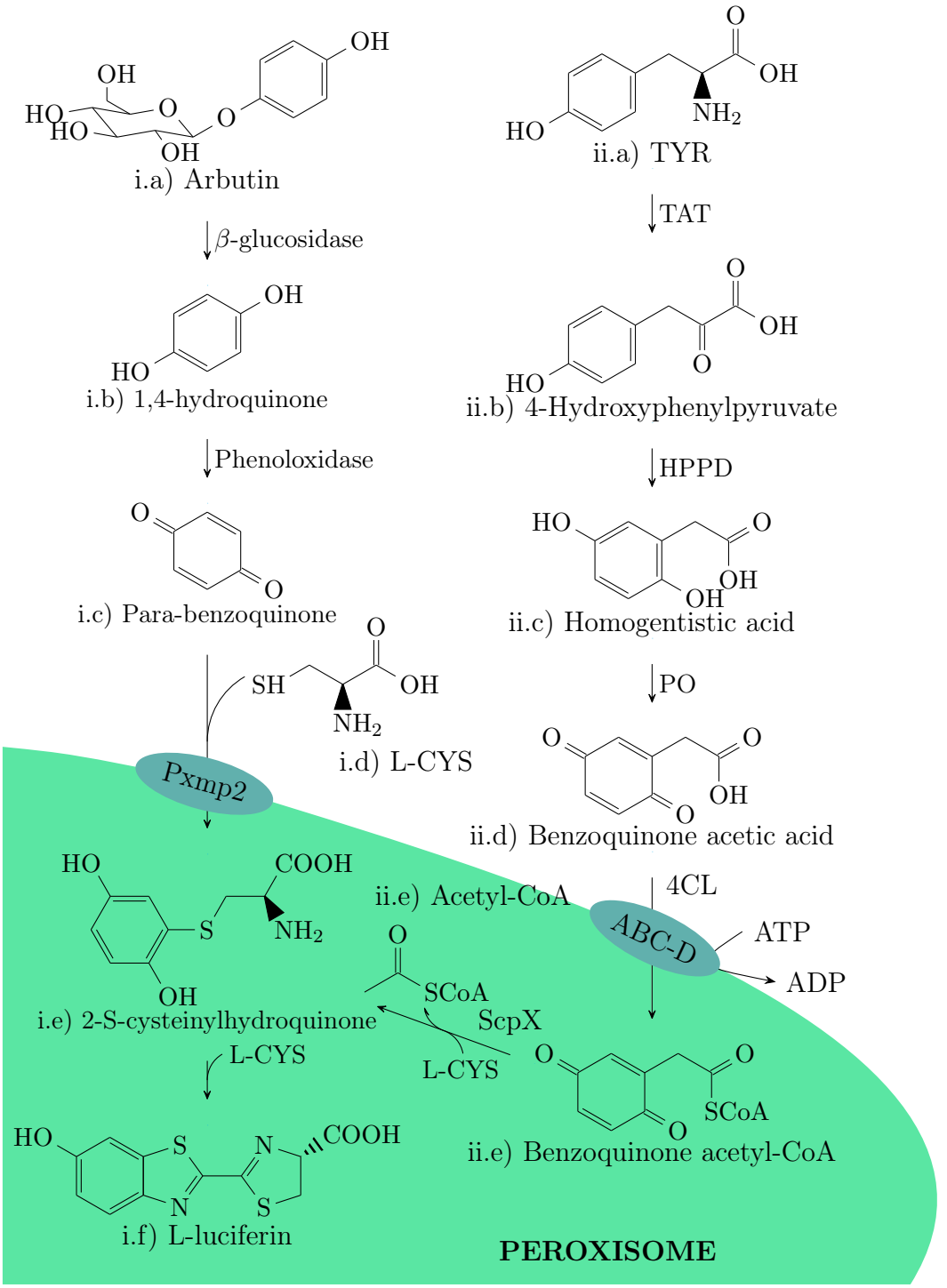
Biosynthesis

Although little is known of the biosynthetic pathways in Elateridae and Phengodidae lineages, at least for the Lampyridae family thorough investigations have shed some light on the metabolic biochemistry of firefly's BL.[90]

Originally, it was suggested that the *in vivo* biosynthesis might proceed via condensation of 2-cyano-6-hydroxybenzothiazole (structure c in Scheme 1.8) with D-CYS.[10] Nonetheless, more recent studies[90] suggest that the biosynthetic pathway starts with the hydrolysis of arbutin (structure i.a in

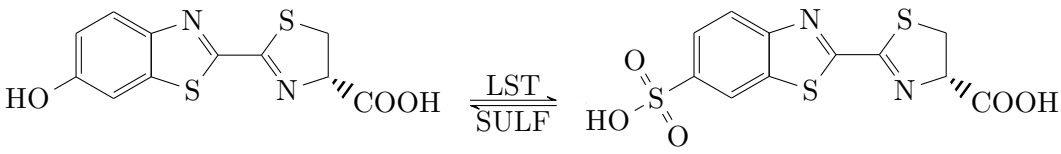
Scheme 1.3) catalysed by a β-glucosidase.[91] The resulting 1,4-hydroquinone (structure i.b) is then oxidized to para-benzoquinone (structure i.c). These two reactions take place in the cytoplasm of the photocytes. [92] Then, para-benzoquinone is transported into the peroxisomes through the Peroxisomal Membrane Protein 2 (Pxmp2) channel, which may also contribute to the diffusion of L-CYS (structure i.d) inside the peroxisome.[93] Inside the photocyte organelle, a 1,4-addition of L-CYS to the benzoquinone moiety, followed by decarboxylation and carbon-sulphur rearrangement of CYS, leads to the formation of 2-S-cysteinylhydroquinone (structure i.e)[94, 95] which, upon the addition of a second L-CYS, yields L-luciferin (structure i.f). At the final step of the biosynthetic route, L-luciferin undergoes Coenzyme A (CoA) esterification followed by epimerization and thioester hydrolysis catalysed, respectively, by the luciferase and Acyl-CoA Thioesterase (ACOT) enzymes, thus generating the D-form (structure a in Scheme 1.4).[91, 96, 97]

Other possible D-luciferin precursor is TYR (structure ii.a in Scheme 1.3),[92] that is present in large amounts in the photogenic layer of firefly's lantern.[98] This amino acid can undergo a series of three oxidative steps followed by a chemical transformation and another oxidation step leading to the formation of 2-S-cysteinylhydroquinone (structure i.e), thus converging into the previously mentioned mechanism. In the first step, TYR is oxidized to 4-hydroxyphenylpyruvate (structure ii.b) by means of the Tyrosine Aminotransferase (TAT) enzyme. Then, the enzyme 4-Hydroxyphenylpyruvate Dioxygenase (HPPD) oxidizes it into homogentisic acid (structure ii.c) which, upon binding to a Phenoloxidase (PO), is further oxidized to benzoquinone acetic acid (structure ii.d).[99–101] Consecutively, this molecule can react with acetyl-CoA to form benozoquinone acetyl-CoA (structure ii.e), possibly catalysed by the 4-Coumarate-CoA Ligase (4CL). Then, after the consumption of an ATP molecule and the consequent release of adenosine diphosphate (ADP), the benozoquinone acetyl-CoA is actively transported into the peroxisome by a protein of the ATP-Binding Cassette protein D subfamily (ABC-D). Once inside the peroxisome, the acetyl group can be removed via the thiolase reaction mechanism of β-oxidation catalysed by thiolase activity of Sterol carrier protein-X (ScpX).[102, 103] After removing an acetyl group, the group-sulfhydryl (SH) of a L-CYS (instead of that of acetyl-CoA in usual β-oxidations) might again react with the terminal carbon of the benzoquinone derivative to form 2-S-cysteinylhydroquinone (structure i.e).[102]



Scheme 1.3: Proposed biosynthetic route of D-luciferin.

As a final remark on D-luciferin biosynthesis, it has been suggested[92] that it can be stored inside the peroxisomes in the form of sulfo-D-luciferin (structure b in Scheme 1.4), which is inactive for luciferase. This derivative can be obtained from the luciferin sulfotransferase (LST) catalysed reaction of D-luciferin with the sulfo-donor 3'-phosphoadenosine 5'-phosphosulfate (PAPS), produced from ATP and inorganic sulfates under the catalysis of PAPS Synthase (PAPSS).[104] This reaction can be reversed by means of a Sulfatase (SULF) enzyme, thus making the D-luciferin available again for the BL reaction.



a) D-luciferin b) Sulfo-D-luciferin

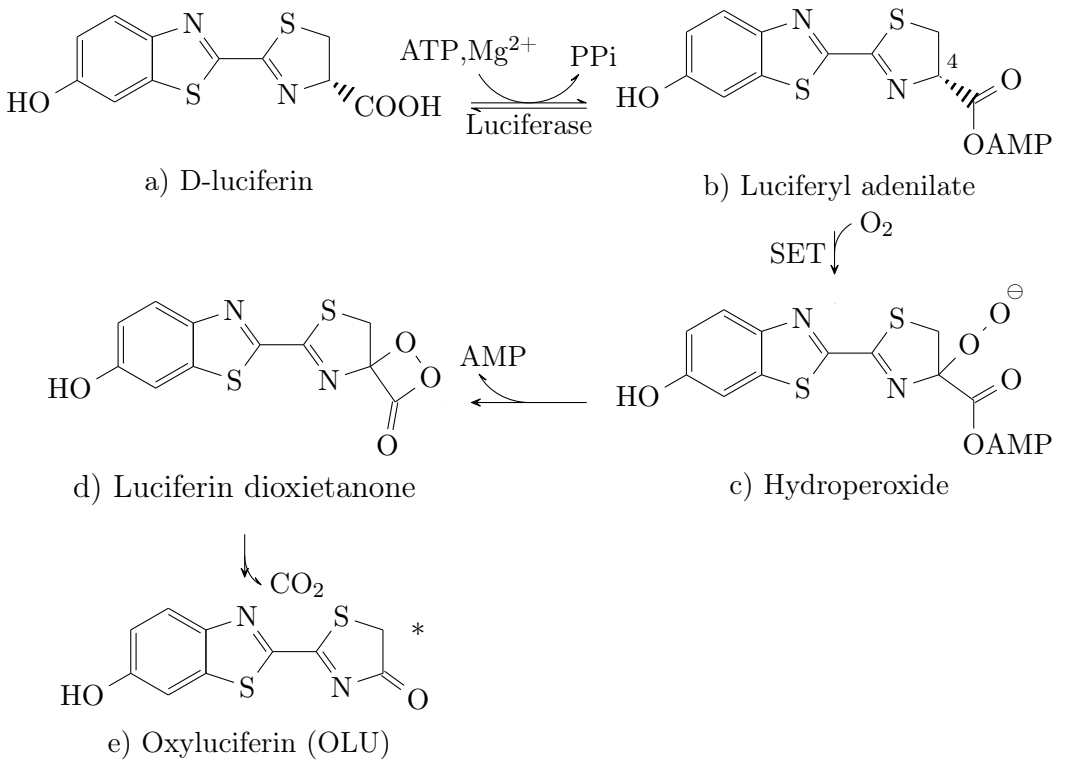
Scheme 1.4: Proposed storage mechanism of D-luciferin.

Bioluminescent Reaction

Once inside the peroxisome, D-luciferin can diffuse into the active site of the open conformation of the luciferase enzyme to undergo the BL reaction with a quantum yield of 41% in aqueous solution.[105] This reaction presents a two steps mechanism consisting on a reversible adenylation followed by an oxydation.

The mechanism starts with the adenylation reaction of D-luciferin with ATP, which is an essential requirement for the *in vitro* reaction,[23] leading to the formation of luciferyl adenylate (structure b in Scheme 1.5), facilitated by a magnesium dication via a bimolecular nucleophilic substitution (S_N2) nucleophilic displacement. The nucleophilic oxygen atom in the carboxyl group at the C_4 carbon of the tz ring is able to attack to the electrophilic phosphorus at the α-phosphoryl group of ATP, thus displacing the inorganic Pyrophosphate (PPi), an excellent leaving group, and transferring the 5'-adenosine monophosphate (AMP) to D-luciferin.[106–109] The presence of the AMP moiety, instead of acting as an energy donor,[106] increases the acidity of the C_4 carbon,[110, 111] favouring the consecutive oxidation step through the formation of a C_4 carbanion.[112] This proton loss is the rate determining step of the Single Electron Transfer (SET) of the oxidation mechanism in which the carbanion undergoes a nucleophilic attack of molecular oxygen, resulting in a linear hydroperoxide (structure c).[110, 113] After the

SET, the α-hydroperoxy group undergoes an intramolecular nucleophilic attack displacing the AMP towards the formation of a strained four-membered cyclic intermediate, the luciferin dioxetanone (structure d).[52, 114, 115] As previously mentioned in Section 1.1.2, this HEI spontaneously decomposes into OLU (structure e) in the S_1 state, generating CO_2. In the case of firefly's luciferin, this step follows the already explained entropic trap mechanism. Finally, the resulting OLU molecule in the excited state decays to the ground state releasing a photon.



a) D-luciferin

b) Luciferyl adenilate

d) Luciferin dioxietanone

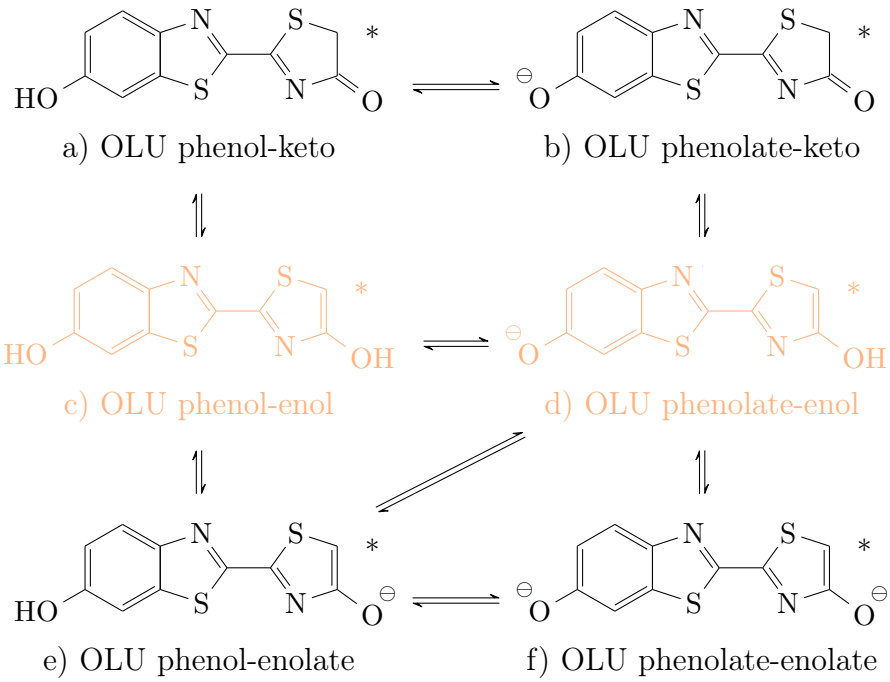
c) Hydroperoxide

e) Oxyluciferin (OLU)

Scheme 1.5: D-luciferin BL reaction.

Although it is well known that OLU is the emitting chromophore of the BL reaction, and due to its chemical instability, the structure of the final product of the BL reaction remained unsolved until 2009 when it was finally crystallized.[116] This instability is explained by accounting for the coexistence of the enol and keto tautomers, which can undergo an autodimerization by a Mannich-type reaction.[117] Nonetheless, it is still unclear which of its six possible tautomeric forms, represented in Scheme 1.6, is the real emitter or whether the measured spectra is obtained from combination of different OLU species since they can be easily interconverted. For instance, the OLU

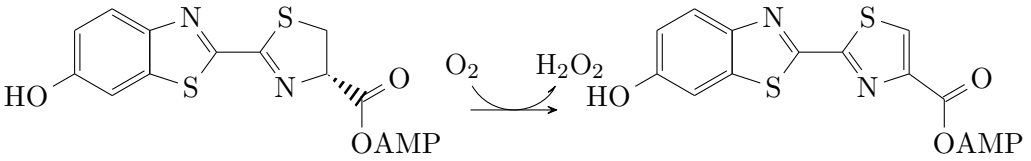
phenol-keto (structure e in Scheme 1.5 and structure a in Scheme 1.6), directly obtained from the dioxietanone decomposition, can rapidly lose a proton via Excited State Proton Transfer (ESPT) to form the phenolate-keto OLU (structure b in Scheme 1.6).[118] Furthermore, it is also possible that both OLU phenol-keto and phenolate-keto species undergo a keto-enol tautomerization, leading to the formation of the OLU phenol-enol and phenolate-enol chromophores (structures c and d), respectively. Nonetheless, it is known that none of these enol forms are responsible for the emission of the system.[119] In the case of the phenol-enol OLU, this is due to the fact that not only it emits in the blue region (445 nm),[120] which is not observed in nature, but also because it can undergo a very fast ESPT to its monoanionic forms, namely, phenol-enolate (structure e) and phenolate-enol, which, at the same time, present a very fast relaxation to the OLU phenolate-enolate (structure f). The species phenol-enolate and phenolate-enolate are also likely emitters of the process. Therefore, when studying the OLU emission it is important to consider the four possible emitters instead of just one of them.



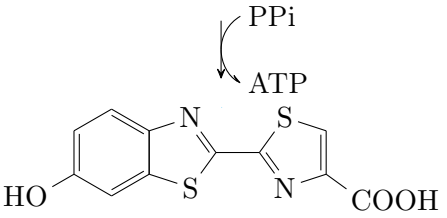
Scheme 1.6: Chemical structure of all the possible OLU tautomers. Those structures that are not possible emitters appear in orange.

Besides the light-producing pathway, the luciferase enzyme also catalyses

the so called dark reaction pathway with a 20% yield (in contrast to the 80% yield of the bright reaction path). In a first step, the luciferyl adenylate (structure b in Scheme 1.5 and structure a in Scheme 1.7), produced after the adenylation of D-luciferin, reacts with molecular oxygen to produce dehydroluciferyl adenilate (structure b in Scheme 1.7), a very powerful tight-binding inhibitor of the BL reaction, and hydrogen peroxide.[121] In a consecutive step, the inhibitor reacts with the previously released PPi producing dehydroluciferin (structure c) and regenerating ATP, favouring the BL path since it is not as strong inhibitor as dehydroluciferyl adenilate and thus, can be more easily removed from the active site of luciferase. Therefore, the catalysing activity of the protein is recovered.[87]



a) Luciferyl adenilate b) Dehydroluciferyl adenilate



c) Dehydroluciferin

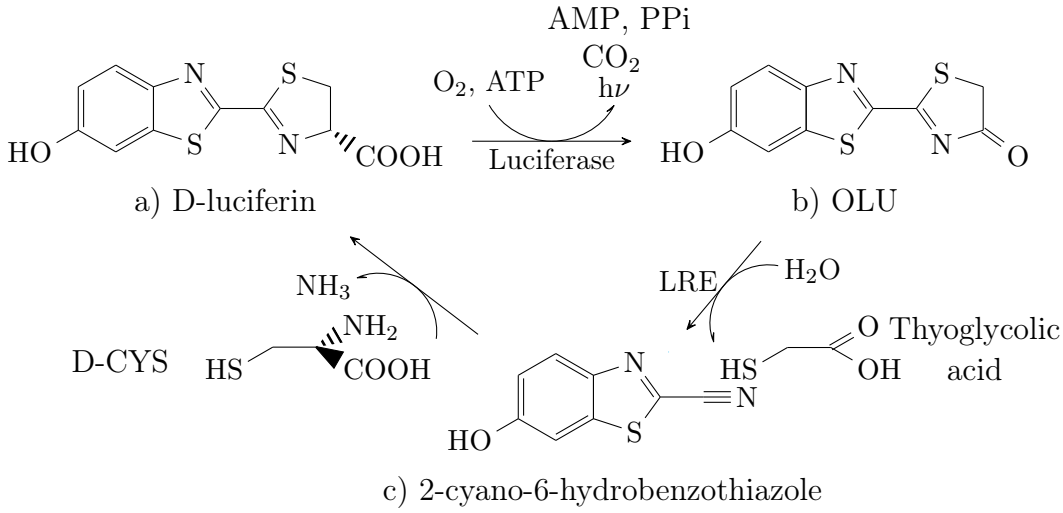
Scheme 1.7: D-luciferin dark reaction pathway.

During the BL reaction, it has been observed that native luciferase undergoes several conformational changes.[88] After the adenylation step, a rotation of around 140º of the C-terminal domain takes place to facilitate the diffusion of the oxygen molecule inside the active site.[122, 123] In this new structure, GLY446 opens a tunnel to the active site while HIS245 further expands it for access to C_4 via side chain rotation.[112] This conformational change could explain the difference in yield between the two reaction paths since the first step for the bright reaction, the deprotonation, does not require the presence of O_2 and, therefore, the intermediate should already be present when oxygen reaches the active site, while the dark reaction path can only take place after the conformational change is completed and oxygen diffuses into the active site. Moreover, luciferase adopts a "closed form" during the formation of the HEI, thus creating a hydrophobic pocket around the active site, and an "open

form" in complex with the reactants and products.[124]

Recovery of D-Luciferin

After the light emission, the last step in the luciferin metabolic pathway is the recycling of the OLU (structure b in Scheme 1.8). It has been reported that the regeneration mechanism of D-luciferin, catalysed by the Luciferin Regenerating Enzyme (LRE),[87, 91, 96, 125–130] takes place via the formation of the 2-cyano-6-hydrobenzothiazole (structure c) which, upon reacting (non-enzymatically) with both L- and D-CYS, results into either L- or D-luciferin (structure a), respectively. Then, L-luciferin (structure i.f in Scheme 1.3) can be interconverted to D-luciferin, as shown in Scheme 1.3. In contrast to this findings, it has also been reported[130] that the regeneration of D-luciferin can also be achieved in the absence of the LRE.



Scheme 1.8: Proposed recycling mechanism of D-luciferin.

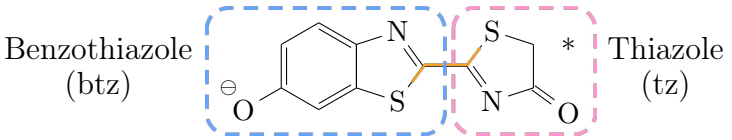
1.2.3 Colour Modulation

A very interesting, and yet unsolved, property of the D-luciferin BL system is the wide range of colours in which it can emit light. This range, which in nature lies between 540 (green-yellow) and 630 nm (red), but can be artificially widened,[2] is known to be correlated to several environmental factors, such as pH, and the microenvironment surrounding the chromophore. Over the years many colour-tuning mechanisms have been proposed.[119, 129,

131] Nonetheless, the exact mechanism of the colour tuning is still unclear. These mechanisms are typically divided into: i) single-emitter mechanisms, ii) emitter-protein modulation mechanisms and iii) the multiple-emitter mechanism.

Initially and until 1991, it was believed that the relationship between the light emitter of OLU and the colour of BL could be explained by the keto–enol tautomerism, which could explain the red shift observed upon acidification (from 565 nm at pH 7.6 to 620 nm pH 5.4).[132] Within this mechanism, red and green-yellow colours of emission would correspond, respectively, to the keto and the enol forms of OLU that are present in the media in different concentrations depending on their acid-base equilibrium.[133–136] However, on the light of new evidence this mechanism was abandoned. This was due to the fact that experiments showed red emission upon: (i) luciferase mutation[137] and (ii) blocking the keto-enol tautomerization in the keto form by means of a 5,5-dimethylation on the tz moiety of the OLU with two WT luciferases naturally emitting in green-yellow and red colours.[138]

Consecutively, it was suggested that the colour modulation corresponded to the presence of a twisted structure around of the C_2-C_2' bond that connects both the btz and the tz moieties (see Scheme 1.9).[139, 140] The so-called Twisted Intramolecular Charge Transfer (TICT) mechanism is usually disregarded attending to both experimental[88, 116, 123, 124, 141–153] and computational evidence.[154, 155] Nonetheless, all these results only offer an static perspective of the problem by looking at either PESs or crystalline structures without accounting for the dynamic nature of the system. Along this thesis, specifically in Chapters III, VI and VIII, the feasibility of this mechanism has been further analysed from a dynamic perspective leading to the conclusion that the red emission can only be reached in the twisted region of the SCCS dihedral. Within this conformation, the brightness of the electronic transition is reduced due to an increase of its CT character but it is not quenched as a



Scheme 1.9: Schematic representation of the btz and tz fragments, respectively in blue and in pink, on the structure of the phenolate-keto OLU. The SCCS dihedral angle is highlighted in orange.

consequence of the vibrational motion.

Relying exclusively on the emission of the phenolate-keto OLU, the resonance structure mechanism[156] proposes that there exists a resonance-stabilized structure so that the resulting charge distribution dictates the colour of emitted light. However, it is still unclear how a CT can result in differences in colour emission upon the interaction with a specific amino acid or a water molecule.

The second category of mechanisms, can be subdivided into ii.a) the micro-environment mechanism[105, 157–161] and ii.b) the open-closed mechanism.[124, 162] The former initially explained the colour shift in terms of proton transfer between the phenolate group or keto–enol–enolate equilibrium.[163–167] However, it was later generalized to include the effect of the polarity of the pocket.[142, 168] The main drawback to this model is that it is unclear which are the relevant interactions from the micro-environment referred to in the related literature. On the flip side, the later mechanism states that the colour of the emitted light depends on the conformation of the luciferase enzyme which, at the same time, influences the active pocket. On the one hand, the motion of the OLU chromophore in the overall closed conformation is restricted by the pocket, which holds it more tightly thus restricting the energy transfer to the environment leading to a yellow-green emission. On the other hand, the open conformation allows a larger freedom of movement on the OLU emitting specie, which, at the same time, leads to a larger energy transfer to the environment and a lower energy of emission. These two mechanisms can be considered complementary as both of them explain the colour differences in terms of the structural configurations of the active site of the different enzymes and the changes on the interactions of the emitter with the surrounding amino acid residues and/or water molecules. Based on this criteria, Chapter V aims to give a full description of the movement of the four possible emitting OLU tautomers inside the active pocket of the WT luciferase in order to gain a deeper understanding on the micro-environment of the system. Furthermore, these results have been compared in Chapter VIII to those of the mutation of a SER by a ASN amino acid in position 286 (SER286ASN), that has been reported to yield red colour of emission (605 nm), in order to better understand the differences on the emitted colour.[124]

The last of the proposed mechanism, the multiple-emitter mechanism,[105, 157–161] accounts for the possibility of having different emitting species at the same time, whose intrinsic emission is only partially affected by intermolecular interactions and polarity. These emitting species, already presented in Scheme 1.6, exist simultaneously and, therefore, the observed colour of emission depends on the relative contribution of the intensity of each emitter. Originally,

any of the OLU tautomers was pointed out as a possible emitter.[105] However, the scientific consensus now assigns, respectively, the phenolate-enol and phenolate-keto OLUs to green and orange/red colours of emission, which is in agreement with mutational analysis demonstrating the existence of at least two species with distinct emission properties.[124, 159]

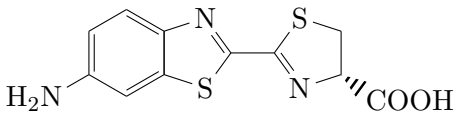
When discussing about different mechanisms, it is straightforward to think that they should, in principle, oppose each other and thus one of them would lead to a full picture of the problem.[119] Nonetheless, as evidenced in the previous discussion, it is not the case of the D-luciferin BL system. In this system, each of the proposed mechanism can help to understand the factors that affect the colour differences from different perspectives and should be taken into account when analysing the results.

Colour Modulation via Functionalization

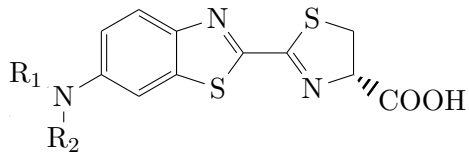
Apart from mutations on the sequence of amino acids of the luciferase enzyme, it has also been demonstrated that the wavelength of the emitted photon can be shifted by functionalization of the D-luciferin ligand, without affecting the brightness of the BL system.[2] This functionalization has been proven to increase the applicably of this system not only by widening the range of the emitted energy, but also by improving the stability of the D-luciferin in solution.

The first synthetic derivatives, the aminoluciferins (structures a-c in Scheme 1.10), were developed by White and McElroy in the 60's from the substitution of the 6'-hydroxy group on the btz moiety by an amino group. Even though this substitution does not result in a complete loss of BL, it does significantly reduce the quantum yield of the reaction. Fortunately, this issue can by easily overcome by using synthetic luciferases. These analogues present larger wavelengths of emission, a bathochromic shift, due to the inductive effect produced by the electron-donating amino group.[24] This shift correlates with the degree of substitution of the amino group. Lower degrees of substitution, like in structure a in Scheme 1.10, result in smaller red shifts than larger ones, like those of alkylaminoluciferins (structure b) and cyclic aminoluciferins (structure c).[169] Furthermore, upon the substitution of the S atom of the tz moiety by a Se atom, it is possible to further increase the red shift of the amino derivatives (structure d).[170]

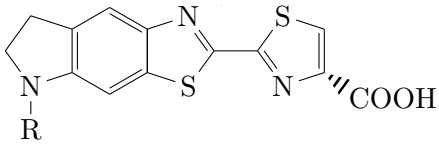
Other type of synthetic D-luciferin analogue is obtained from the substitution of the btz fragment of the D-luciferin by an aromatic moiety like a naphthalene (structure e) or a quinoline (structure f).[171, 172] These derivatives, first synthesised in 1989, present a wavelength of emission dependent



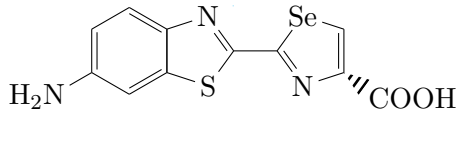
a) Aminoluciferin
(594 nm)



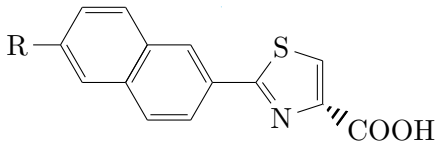
b) Alkylaminoluciferin
(623 nm for R_1=R_2=Me)



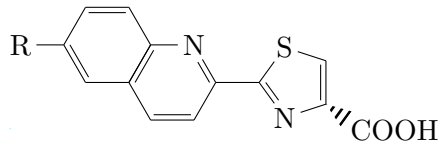
c) Cyclic aminoluciferin
(599 nm for R=H)



d) Se aminoluciferin
(600 nm)



e) Naphthalene derivatives
(524 nm for R=OH)



f) Quinoline derivatives
(608 nm for R=OH)

Scheme 1.10: Synthetic derivatives of D-luciferin and their emission wavelengths in the WT luciferase enzyme.

on the R substituent. On the one hand, if the R substituent is an hydroxy group, the emission presents a hypsochromic shift (to shorter wavelengths) for the naphthalene derivative and a bathochromic shift for the quinoline specie. Moreover, the emission stability of these hydroxyl chromophores is highly affected by the pH of the media, being necessary a high pH (9.3) for them to present a stable emission. On the other hand, if the R substituent is an amino group, the emission presents a bathochromic shift for the quinoline derivative, while for the naphthalene species the emission wavelength remains very similar to that of OLU. Furthermore, the methylation of the amino group of the quinoline derivative produces a further increase in the emission wavelength in contrast to the naphthalene species, for which no significative differences are observed.

1.2.4 Applications in Medicine

Apart from the nocturnal light spectacle performed by animal species provided with D-luciferin based BL, which can obviously trigger anyone's curiosity; this BL system also presents a wide variety of applications in very different fields. All these applications rely, to some extent, on the fact that ATP is the limiting factor for the amount and intensity of light radiated from a sample, which was discovered in the 1950's by McElroy and Hastings.[11] Thus, each application owes at least some debt to them. Nonetheless, and although BL methods were first introduced into clinical chemistry in 1979,[173] it was not until the 1980's, when Marlene DeLuca began to work on the field, that the range of applications of D-luciferin based BL systems began to widen.[174] During her short carrier, she paved the way to modern molecular biology using report gene technology and revolutionized the study of molecular biology.[3] Specifically, she was able to synthesize an active firefly luciferase by *in vitro* translation of ribonucleic acid (RNA) from the lantern of the insects[175] and to obtain a transient and stable expression of the firefly luciferase gene first in plant cells and transgenic plants[176] and then in mammalian cells.[177]

Classical applications of this system rely on McElroy and Hastings' discovery to determine the ATP concentration, which can be employed in many different analytical studies.[2] This molecule is the universal energy carrier within cells. In the case of eukaryotic cells, its intracellular concentration is always maintained at a high level. Thus, its analysis allows for the evaluation of the cellular metabolic potential. While healthy cells maintain normal levels of ATP, these levels fall rapidly upon cellular death. Therefore, evaluation of intracellular concentration of ATP permits the evaluation of the effects of different agents on cell proliferation or cytotoxicity.[178] Moreover, and given that 90% of the ATP in these type of cells is synthesized in the mitochondria; intracellular quantification of this molecule can be be used to asses the mitochondrial activity levels,[179, 180] whose reduction is related to ageing processes.[181]

Furthermore, high sensitivity methods for the *in vivo* determination of ATP concentrations in blood by means of the D-luciferin/luciferase BL system have also been developed.[38] These methods have been clinically tested in patients with, among other conditions, chronic renal failure, cystic fibrosis and cancer with the hope that ATP measurement will become routine during diagnosis and medical treatment in the future. Additionally, BL assays can also be employed to evaluate the contamination of surface probes and other samples that contact various microorganisms because ATP molecules are found in every living organism. As a matter of fact, ATP-induced BL is one of the

most rapid and sensitive methods to improve cleaning procedures and reduce the spread of infections in intensive care units of hospitals, if employed in conjunction with education and training of cleaning staff,[182, 183] and to assess[184] and identify (when combined with immunoassays)[185] bacterial infections of the urinary tract. Nonetheless, the main disadvantages of this method are substrate instability and relatively high cost.[2]

A more ingenious approach of BL applications is bioimaging, which allows to obtain, in a non-invasive manner, images of live tissues or processes within cells. The application of BL techniques is very suitable for this purpose because of i) assay procedures are potentially simple, ii) the high sensitivity of BL due to the absence of autofluorescence and a low background, and iii) assays are broadly applicable to various organisms, from bacteria to living subjects. This advantages facilitate BL applications, which present a higher sensitivity than those utilizing fluorescent proteins, in studies of various processes both *in vitro* and *in vivo* employing either luminometers or Charge-Coupled Device (CCD) cameras.[186] BL bioimaging applications mostly focus on molecular imaging on single-cell-based analyses without the destruction of the cells themselves for the understanding of the dynamics of gene products including RNAs, proteins, lipids, sugars, and small molecules.[187] This approach has already been used for intracellular detection of bioactive small molecules, the study of protein conformational changes, protein localization and dynamics, and measurement of protein–protein interactions in real-time with high spatial and temporal resolution. The study of these cellular events in the physiological environment of live cells and animals can lead to the discovery of new drugs against different diseases.

Among these BL bioimaging techniques, an increasingly popular method is the Protein–Fragment Complementation Assay (PCA), which can be used to monitor the activities of small molecules, protein–protein interactions, protein–nucleic acid interactions, and protein conformational changes. The use of a luciferase as a reporter protein presents a strong advantage for this type of assays since it allows for real-time analysis in live cells and animals (see Figure 1.4). Moreover, the association process of the required luciferase fragments is, in general, reversible and with shorter detection times (from several seconds to minutes) in contrast to fluorescent PCA, which presents an irreversible protein reconstitution and detection times from several minutes to hours.[187]

In the general PCA method, the reporter protein is rationally dissected into two complementary fragments, each of them genetically fused to the proteins of interest. Upon association of the two interacting proteins under study, the unfolded reporter protein fragments are brought into proximity, thus allowing them to complement and fold into the unique three-dimensional structure of

Figure 1.4: Graphical representation of *in vivo* BL of mice adapted from Figure 1 of reference [169].

the reporter protein and to reconstitute its activity.[187]

In the case of BL-PCA (Figure 1.5(A)), also called Luciferase Complementation Imaging (LCI); the method is based on the hypothesis that the N- and C-terminal domains of the luciferase, mentioned in Section 1.2.2, do not have a dramatic effect on each other's stability and that they fold independently. Hence, a separation of the two luciferase domains and their attachment to other proteins is probable and should not have any disruptive effect on their respective structures.[188] The determination of the optimal dissection sites of luciferase is crucial for guaranteeing temporal inactivation and conditional reconstitution of the luciferase activity. Fortunately, green and red emitting luciferase fragments have already been developed from firefly's and click beetle's luciferases.[189, 190] These so-called split firefly's luciferase fragments were initially used for the detection of protein–protein interactions[191] and since then, the potential applications of BL-PCAs were explored to monitor caspase activities, protein translocation, protein conformational changes, drug–protein interaction, and protein–protein interactions non-invasively in living mice.[191–193] Compared to other methods, luciferase complementation strategies provide detection of reversible kinetics of protein–protein interactions within minutes due to rapid association and moderate dissociation rates[190] and general applicability for protein interactions with considerable spatial and temporal resolution in opaque or strongly autofluorescent living subjects.

Apart form the PCA approach, it is also possible to exploit circular permutation, which consists on the rearrangement of the protein sequence, to perform molecular imaging. In this type of essays , the original N-terminal

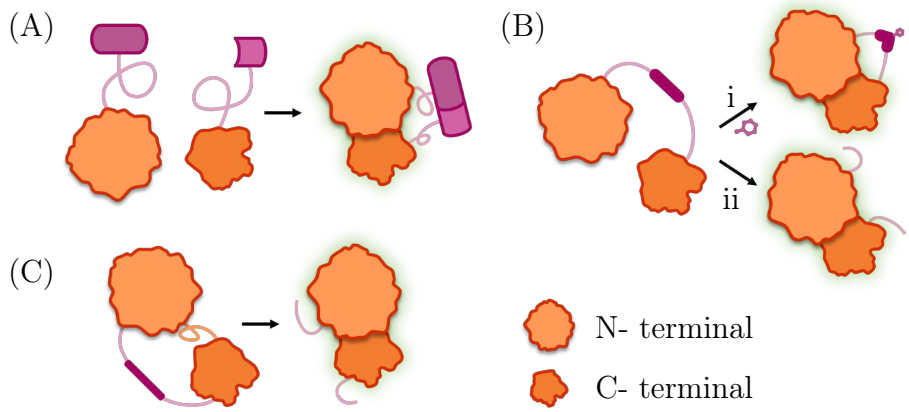


Figure 1.5: Schematic representation of different BL bioimaging techniques: A) BL Protein–Fragment Complementation Assay, B) circular permutation upon i) small ligand attachment and ii) protease activity and C) using cyclic luciferases. D-luciferin based luciferase is represented in orange and external peptide sequences, in pink.

and C-terminal regions of a luciferase are linked with a peptide linker, which may have sites for specific ligand attachment that would induce conformational changes and alter the BL signals (Figure 1.5(B)). Then, new N-terminals and C-terminals are randomly created.[188] This peptide insertion temporarily disrupts fluorescence due to solvent penetration within the protein core. Upon the action of an external actor like, for instance, a small molecule (Figure 1.5(B.i)) or a protease enzyme (Figure 1.5(B.ii)), the inhibitory action is relieved. Although the fundamental concept of circular permutation was originally introduced for coelenterazine based luciferases,[194] it can also be applied to D-luciferin based luciferases to detect protease activity.[195] Nonetheless, its application is limited owing to the optimization of enzyme activity in most studies.[188, 196]

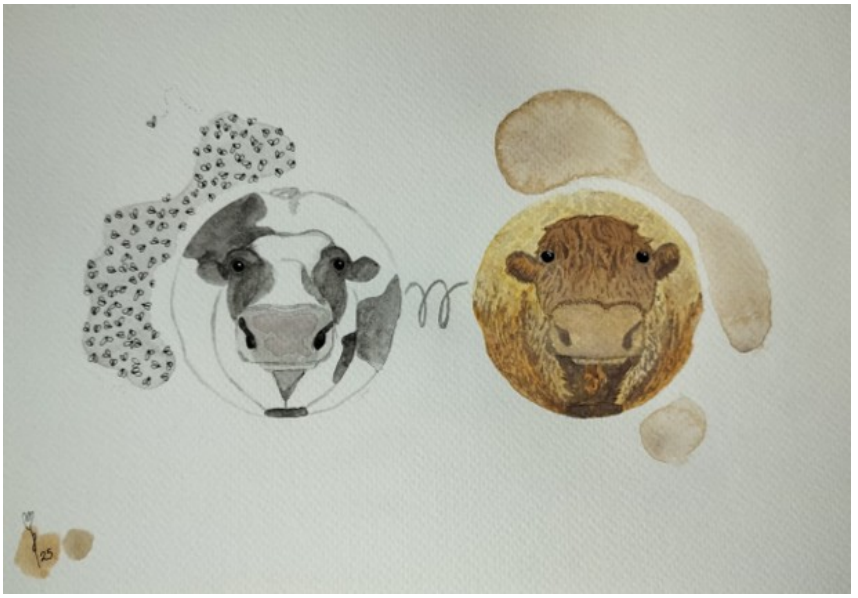
Another approach consists on generation of cyclic luciferases by ligating the N-terminal and C-terminal ends by protein splicing, which produces a closed circular polypeptide chain.[197, 198] After this modification, the structure of the luciferase is distorted, resulting in the lost of its BL activity. When the substrate sequence is digested by a protease, the luciferase changes into an active form and restores its activity, as graphically depicted in Figure 1.5(C). An important drawback of cyclic luciferases is that their applicability is limited to screening just a few proteases and is not suitable for investigating many other biochemical analytes.[188]

As mentioned before, these methodologies present a wide variety of possi-

bilities for research in many different fields,[2, 188] which go from the study of neurodegenerative deceases such as Alzheimer, Parkinson and amyotrophic lateral sclerosis; tumour and infections cells to immunological research and intracellular communication and signalling.[199, 200] Thus, a further understanding of the BL system itself could allow for the development of more specialized tools for the research of these serious illnesses. For example, *in vivo* applications would benefit from red shifted emissions since the presence of hemoglobin, melanin, and other pigmented macromolecules in tissues decrease the intensity of the measured signal because they absorb strongly in the blue-green region of the visible spectrum (400–600 nm) where the OLU emits.[201, 202]

Chapter II

Computational Methods



As defined by the Real Academia Española (RAE), physical chemistry means

"Rama de la ciencia que estudia la interrelación entre las propiedades físicas y químicas de una sustancia."

That is, physical chemistry is the branch of science that studies the relationship between physical and chemical properties of a substance. For this purpose, different experimental technics and theoretical frameworks have been developed. Concurrently to the development of more complex experimental methodologies that increase the amount of measurable observables, theoretical perspectives that allow a coherent and consistent comprehension of the results have been formulated. These theories are the basis behind the so-called computational chemistry, which employs the software implementation of the different equations resulting from these theories to predict the physicochemical properties of the system under study. While it is true that experimental methodologies offer a measurement of real life observables, very often fail to provide a full picture of the chemical problem since only certain properties are accessible. In this regard, computational chemistry offers a complementary approach to deepen into the chemical problem at hand.

Within this chapter, the foundations of computational chemistry will be discussed. Although there exist many other approaches for the calculations, this chapter only describes those that have been employed during this thesis: Section 2.1 focuses on the electronic structure theory by describing the Time Independent Schrödinger Equation (TISE), a pillar of Quantum Mechanics (QM), and different approximations for its solution for systems with multiple particles, namely, the Hartree-Fock (HF) method and the Density Functional Theory (DFT). Furthermore, this section deepens on the most typical QM calculations in order to provide a broader picture of the applications of the solutions of the Schrödinger equation. On the flip side, Section 2.2 discusses different flavours of nuclear motion simulations under the classical framework of Newton's equations of motion and the key aspects to consider when producing the trajectories. Finally, Sections 2.3 and 2.4 focus, respectively, on the description of big systems employing both QM and Molecular Mechanics (MM) theories, and on interaction energies at different levels of theory.

2.1 Electronic Structure Theory

In 1926 and inspired by wave-particle duality postulated by the Nobel laureate in Physics Louis-Victor de Broglie during his PhD thesis,[203] Erwin Schrödinger published a theory that allowed to describe the state of a system as a wavefunction[204–207] that resulted in the simultaneous deduction of the laws of motion and the quantum conditions from one simple Hamiltonian principle:

$$i\hbar\frac{\partial}{\partial t}\Psi(\boldsymbol{r},t) = \hat{\mathcal{H}}\Psi(\boldsymbol{r},t) \tag{2.1}$$

where $\hbar$ is the reduced Planck constant ($\hbar = h/2\pi$ with h the Planck constant, allowing conversion from energy to frequency units); $\Psi(\boldsymbol{r},t)$, the state function and $\boldsymbol{r}$ and t are the spatial and temporal coordinates of the system, respectively. Moreover, the term $\hat{\mathcal{H}}$ of the non-relativistic Time Dependent (TD) Schrödinger Equation (SE) (Equation 2.1) corresponds to the energy operator of the system, the Hamiltonian, that can be constructed from a kinetic and a potential energy terms:

$$\hat{\mathcal{H}} = -\frac{\hbar^2}{2m}\nabla^2 + V(\boldsymbol{r},t) \tag{2.2}$$

with m being the mass of the particle; ∇^2, the Laplace operator defined as the sum of all the unmixed second partial derivatives in the Cartesian coordinates $\boldsymbol{r}$ and $V(\boldsymbol{r},t)$, the TD potential. This theory granted Erwin Schrödinger the Nobel Prize in Physics 1933 *for the discovery of new productive forms of atomic theory.*

The state function $\Psi(\boldsymbol{r},t)$, also known as the wavefunction, contains all the information of the system and, while it lacks a physical meaning itself, Max Born postulated[208] that the probability density:

$$|\Psi(\boldsymbol{r},t)|^2 d\boldsymbol{r} \tag{2.3}$$

corresponds to the probability at time t for finding the particle within the interval of space $d\boldsymbol{r}$.

2.1.1 The Time Independent Schrödinger Equation

While the resolution of the TDSE leads to a full description of the chemical system at hand, many applications of QM to chemistry do not require from this equation and therefore, a more simple version of Equation 2.1, the TISE, can be employed. This equation can be easily derived from the TDSE.

For simplicity, let's consider a one-particle, one-dimensional system under a time independent external field $V(x)$.[209] Then, the TDSE can be written as:

$$i\hbar\frac{\partial}{\partial t}\Psi(x,t) = -\frac{\hbar^2}{2m}\frac{\partial^2}{\partial x^2}\Psi(x,t) + V(x)\Psi(x,t) \tag{2.4}$$

One possible solution for this equation, although not the only one, would be a state function in which the spatial and temporal components of the wave-function are not dependent from one another, namely a state function in which the spatial and temporal components are separable. Thus, $\Psi(x,t)$ can be expressed as the product of a function of time, $f(t)$ and a function of x, $\psi(x)$.

$$\Psi(x,t) = f(t)\psi(x) \tag{2.5}$$

Taking partial derivatives of Equation 2.5, we have:

$$\frac{\partial}{\partial t}\Psi(x,t) = \frac{df(t)}{dt}\psi(x) \qquad\qquad \frac{\partial^2}{\partial x^2}\Psi(x,t) = f(t)\frac{d^2\psi(x)}{dx^2} \tag{2.6}$$

Thus, substituting in Equation 2.4:

$$\begin{aligned} i\hbar\frac{df(t)}{dt}\psi(x) &= -\frac{\hbar}{2m}f(t)\frac{d^2\psi(x)}{dx^2} + V(x)f(t)\psi(x) \\ i\hbar\frac{1}{f(t)}\frac{df(t)}{dt} &= -\frac{\hbar}{2m}\frac{1}{\psi(x)}\frac{d^2\psi(x)}{dx^2} + V(x) \end{aligned} \tag{2.7}$$

This way, the left-hand side is exclusively dependent on time and the right-hand side, on the spatial coordinate. Thus, each side of Equation 2.7 is completely independent of either t or x and so, since the function does not depend on any of the two variables, x and t, it must be a constant. Let's call it E. Then, equating the left side of Equation 2.7 to E and integrating both sides of this equation with respect to t:

$$\begin{aligned} \frac{df(t)}{f(t)} &= -\frac{iE}{\hbar}dt \\ \int\frac{df(t)}{f(t)} &= -\int\frac{iE}{\hbar}dt \\ ln\left(f(t)\right) &= -\frac{iE}{\hbar}i + \mathcal{C} \\ f(t) &= e^{-\frac{iEt}{\hbar}}e^{\mathcal{C}} = Ae^{-\frac{iEt}{\hbar}} \end{aligned} \tag{2.8}$$

where $\mathcal{C}$ is an arbitrary constant of integration and $e^{\mathcal{C}}$ can be substituted by a constant A, which can be omitted from $f(t)$ since it can be considered as a factor in the $\psi(x)$ function that multiplies $f(t)$ in Equation 2.5. Thus:

$$f(t) = e^{-\frac{iEt}{\hbar}} \tag{2.9}$$

Substituting now Equations 2.8 and 2.9 in Equation 2.7 then, it can be shown that:

$$-\frac{\hbar^2}{2m}\frac{d^2}{dx^2}\psi(x) + V(x)\psi(x) = E\psi(x) \tag{2.10}$$

Equation 2.10 is the non-relativistic time-independent Schrödinger equation for a single particle of mass m moving in one dimension. While most of the terms in this equation have been properly described, the E constant, which appeared when stating that the left-hand side of Equation 2.7 had to be a constant, has not been properly defined. In fact, the selection of the E letter is not completely arbitrary. E has the same dimensions as $V(x)$ and so E has the dimensions of energy. In fact, it is postulated that E is the energy of the system. Hence, for cases where the potential energy is only a function of x, there exist wavefunctions of the form:

$$\Psi(x,t) = e^{-\frac{iEt}{\hbar}}\psi(x) \tag{2.11}$$

This wavefunction is a TD function nonetheless, any measurable quantity is still time independent, as shown in Equation 2.12 for the probability density. Thus, the states calculated from it are typically called stationary states, and neglecting their time-evolution is a correct approach when studying the properties of the systems they describe.

$$\begin{aligned} |\Psi(x,t)|^2 &= \Psi(x,t)^*\Psi(x,t) = \left[e^{-\frac{iEt}{\hbar}}\psi(x)\right]^* e^{-\frac{iEt}{\hbar}}\psi(x) \\ &= e^{\frac{iEt}{\hbar}}\psi^*(x)e^{-\frac{iEt}{\hbar}}\psi(x) = e^0\psi^*(x)\psi(x) \\ |\Psi(x,t)|^2 &= |\psi(x)|^2 \end{aligned} \tag{2.12}$$

2.1.2 The Born-Oppenheimer Approximation

Up to this point, the state function and the energy operator have been discussed entirely in a general way except for the non-relativistic TISE, which can be expressed for a system of multiple particles moving in a three-dimensional space as:

$$\hat{\mathcal{H}}\Psi = E\Psi \tag{2.13}$$

Moreover, although the Hamiltonian operator has already been described in Equation 2.2, the form of the time independent potential has not been discussed. In the case of a molecular system[210, 211] containing M nuclei and N electrons, the energy operator is composed by the kinetic energy of the nuclei (T_n), the kinetic energy of the electrons (T_e), the nuclear (V_{nn}) and

electronic (V_{ee}) repulsion energy and the coulomb attraction between electrons and nuclei (V_{ne}):

$$\hat{\mathcal{H}} = T_n + T_e + V_{nn} + V_{ee} + V_{ne} \tag{2.14}$$

Each of these terms can be expressed, in atomic units, as a function of the nuclear ($\boldsymbol{R}$) and electronic ($\boldsymbol{r}$) coordinates:

$$\begin{aligned}\hat{\mathcal{H}}(\boldsymbol{r},\boldsymbol{R}) = &-\sum_{k=1}^{M}\frac{1}{2m_k}\nabla_k^2 - \sum_{i=1}^{N}\frac{1}{2}\nabla_i^2 \\ &+\sum_{k<l}^{M}\frac{Z_k Z_l}{|\boldsymbol{R_k}-\boldsymbol{R_l}|} + \sum_{i<j}^{N}\frac{1}{|\boldsymbol{r_i}-\boldsymbol{r_j}|} - \sum_{i=1}^{N}\sum_{k=1}^{M}\frac{Z_k}{|\boldsymbol{r_i}-\boldsymbol{R_k}|}\end{aligned} \tag{2.15}$$

where k and l run over the M nuclei, i and j run over the N electrons, m_k is the mass of nucleus k and Z_k is the atomic number of nucleus k. From this expression, it can be observed that the two kinetic energy terms only depend on one of the particles of the system, while potential attraction and repulsion terms present a pairwise interaction between two particles of the system. Thus, the movement of any of the particles of the system is affected by the presence of all the other particles, therefore the movement of all the particles of the system is correlated. This picture presents a great challenge for the resolution of the TISE since there is no known analytic solution for a system containing three (or more) particles. To facilitate the resolution of the TISE, Max Born and J. Robert Oppenheimer proposed in 1927 the so-called Born–Oppenheimer (BO) approximation.[212] This approximation allows the separation of the motion of the nuclei and the motion of the electrons by considering that the motion of the nucleus is much slower than that of the electrons due to their mass difference (protons and neutrons are about 1800 times more massive than electrons). Withing this framework, the electrons of a molecule are considered to present an instantaneous relaxation with respect to the nuclear motion or what it is equivalent, the electrons are considered to move in a field of fixed nuclei. Under this assumption, the second term of Equation 2.15, the kinetic energy of the nuclei, can be neglected and the third term, the nuclear repulsion energy, can be considered to be a constant as the electronic energy presents a parametrical dependence on the nuclear coordinates. Thus, it is possible to describe an electronic Hamiltonian operator which depends only on the electronic coordinates:

$$\hat{\mathcal{H}}_{el}(\boldsymbol{r};\boldsymbol{R}) = -\sum_{i=1}^{N}\frac{1}{2}\nabla_i^2 + \sum_{i<j}^{N}\frac{1}{|\boldsymbol{r_i}-\boldsymbol{r_j}|} - \sum_{i=1}^{N}\sum_{k=1}^{M}\frac{Z_k}{|\boldsymbol{r_i}-\boldsymbol{R_k}|} \tag{2.16}$$

Moreover, the total molecular wavefunction $\Psi(\boldsymbol{r}, \boldsymbol{R})$ can be separated into a nuclear wavefunction $\Omega(R)$ that describes the vibrational, rotational and translational movements of the molecular system and an electronic wavefunction $\Phi(\boldsymbol{r}; \boldsymbol{R})$, which depends explicitly on the electronic coordinates but parametrically on the nuclear positions meaning that, for different arrangements of the nuclei, $\Phi(\boldsymbol{r}; \boldsymbol{R})$ is a different function of the electronic coordinates:

$$\Psi(\boldsymbol{r}, \boldsymbol{R}) = \Phi(\boldsymbol{r}; \boldsymbol{R})\Omega(\boldsymbol{R}) \tag{2.17}$$

Under the BO approximation, it is possible to obtain the electronic energy of the system, which depends parametrically on the nuclear coordinates ($\mathcal{E}_{el} = \mathcal{E}_{el}(\boldsymbol{R})$) by solving the electronic TISE:

$$\hat{\mathcal{H}}_{el}(\boldsymbol{r}; \boldsymbol{R})\Phi(\boldsymbol{r}; \boldsymbol{R}) = \mathcal{E}_{el}\Phi(\boldsymbol{r}; \boldsymbol{R}) \tag{2.18}$$

And so, the total energy $\mathcal{E}_{tot}$ is given by the addition of the $\mathcal{E}_{el}$ and the nuclear repulsion energy:

$$\mathcal{E}_{tot}(\boldsymbol{r}; \boldsymbol{R}) = \mathcal{E}_{el} + \sum_{k<l}^{M} \frac{Z_k Z_l}{|\boldsymbol{R_k} - \boldsymbol{R_l}|} \tag{2.19}$$

The BO approximation is entirely justified in most cases and, without it, the concept of PES, that is the surface defined by $\mathcal{E}_{el}$ over all the possible arrangements of the nuclear coordinates,[211] would not have a clear definition and so, many pillars of chemistry would lack of theoretical support. For instance, it would not be possible to determine the critical points on the PES which are typically associated to equilibrium and transition state geometries and, instead, it would only be possible to discuss high-probability regions of the nuclear wavefunctions. However, and even though the use of the BO approximation is broadly used in computational chemistry, it completely breaks down when two or more electronic states come close together energetically due to the strong coupling between nuclear and electronic motions and, thus, the motion of the nuclei can trigger changes such as an electronic transition[213, 214] or the nature of the character of the state. A description of the methods that are able to overcome this is beyond the scope of the thesis, and several accounts in that regard can be found elsewhere.[214, 215] Next sections will be dedicated to the resolution of the TISE via different approximate numerical techniques (within the framework of the BO approximation).

2.1.3 The Hartree-Fock Approximation

While the electronic TISE in Equation 2.18 depends only on the spatial coordinates of the electrons of the system (and parametrically on those of the nuclei), a complete description of the electronic problem can only be achieved when the whole nature of the fundamental particle is explicitly included. Thus, for the description of each electron, it is necessary to describe not only its three spatial coordinates $\boldsymbol{r}$ but also a spin coordinate ω:

$$\boldsymbol{x} = \{\boldsymbol{r}, \omega\} \tag{2.20}$$

Therefore, the wavefunction of a general molecular system of N interacting electrons can be then expressed as a function of the $\boldsymbol{x}$ collective variable of each electron:

$$\Phi(\boldsymbol{x_1}, \boldsymbol{x_2}, ..., \boldsymbol{x_N}) \tag{2.21}$$

Furthermore, electrons are fermions, meaning that they present half-integer spin values, and as so, they obey the antisymmetry principle. This principle, also known as Pauli exclusion principle, was formulated by Wolfgang Pauli in 1940[216–218] and states that:

"under the exchange of two identical particles, the total (many-particle) wavefunction is antisymmetric for fermions."

This means that, if the space and spin coordinates of two identical particles ($\boldsymbol{x_i}$ and $\boldsymbol{x_j}$) are interchanged, then the total wavefunction changes sign:

$$\Phi(\boldsymbol{x_1}, ..., \boldsymbol{x_i}, ..., \boldsymbol{x_j}, ..., \boldsymbol{x_N}) = -\Phi(\boldsymbol{x_1}, ..., \boldsymbol{x_j}, ..., \boldsymbol{x_i}, ..., \boldsymbol{x_N}) \tag{2.22}$$

One of the simplest ways to build this antisymmetric spin orbital wavefunction is by means of a Slater determinant (SD), a mathematical construct that arranges p one-electron spin orbitals wavefunctions $\chi(\boldsymbol{x})$, into a determinant form in which rows are labelled by electrons and columns by spin orbitals:

$$\Phi(\boldsymbol{x}_1, ..., \boldsymbol{x}_N) = \frac{1}{\sqrt{N!}} \begin{vmatrix} \chi_a(\boldsymbol{x}_1) & \chi_b(\boldsymbol{x}_1) & \cdots & \chi_p(\boldsymbol{x}_1) \\ \chi_a(\boldsymbol{x}_2) & \chi_b(\boldsymbol{x}_2) & \cdots & \chi_p(\boldsymbol{x}_2) \\ \vdots & \vdots & \ddots & \vdots \\ \chi_a(\boldsymbol{x}_N) & \chi_b(\boldsymbol{x}_N) & \cdots & \chi_p(\boldsymbol{x}_N) \end{vmatrix} \tag{2.23}$$

where $(N!)^{-1/2}$ is just a normalization factor. By applying the properties of determinants, this formulation ensures both requirements of Pauli exclusion principle: (i) that the overall wavefunction changes sign when any two electrons

are exchanged, since upon the exchange of two rows the determinant is multiplied by -1, and (ii) that two electrons cannot occupy the same spin orbital, which would mean a SD with two identical columns and so it would make the determinant equal to 0, meaning that the wavefunction does not exist. Moreover, when a spin orbital wavefunction is described by a single SD, it can be demonstrated[210] that the motion of electrons with parallel spin is correlated while the motion of electrons with opposite spin is not and so, this description allows for the incorporation of some exchange correlation effects. Nonetheless, single determinantal wavefunction methods are considered to be uncorrelated as the motion of electrons with opposite spins remains non-correlated.

Once the wavefunction is completely defined, the different properties of the multi-electronic system can be calculated by different operations that involve the wavefunction. Nonetheless, it is not possible to solve exactly the TISE. Withing the HF approximation, the complicated many-electron problem described in Equation 2.18 is reformulated by a sum of one-electron operators[210] in which electron-electron repulsion is treated in an averaged way, thus avoiding the two-electron term:

$$\hat{f}(\boldsymbol{x}_1)\chi_p(\boldsymbol{x}) = \mathcal{E}_p\chi_p(\boldsymbol{x}) \quad \forall \{\chi_p\} \tag{2.24}$$

where $\{\mathcal{E}_p\}$ are the spin-orbital energies. The $\hat{f}(\boldsymbol{x})$ operator, denominated the Fock operator, is defined as the sum of the core Hamiltonian $\hat{h}(\boldsymbol{x}_1)$ and the potential energy $\hat{\mathcal{V}}^{HF}(\boldsymbol{x}_1)$, which represents the potential of mean field created by the rest of the electrons of the system on electron one:

$$\hat{f}(\boldsymbol{x}_1) = \hat{h}(\boldsymbol{x}_1) + \hat{\mathcal{V}}^{HF}(\boldsymbol{x}_1) \tag{2.25}$$

The core Hamiltonian $\hat{h}(\boldsymbol{x}_1)$, as expressed in atomic units in Equation 2.26, not only contains the kinetic energy of the electron but also the coulomb attraction between the electron and the nuclei.

$$\hat{h}(\boldsymbol{x}_1) = -\frac{1}{2}\nabla_1^2 + \sum_{k=1}^{M}\frac{Z_k}{|\boldsymbol{r}_1 - \boldsymbol{R}_k|} \tag{2.26}$$

Furthermore, $\hat{\mathcal{V}}^{HF}(\boldsymbol{x}_1)$ can be expressed as a function of the two-electron integrals $\hat{\mathcal{J}}_b(\boldsymbol{x}_1)$ and $\hat{\mathcal{K}}_b(\boldsymbol{x}_1)$, which represent the classical Coulomb repulsion between electrons and the exchange interactions between electrons arising from the antisymmetric character of the wavefunction, respectively:

$$\hat{\mathcal{V}}^{HF}(\boldsymbol{x}_1) = \sum_{b=1}^{N}\left(\hat{\mathcal{J}}_b(\boldsymbol{x}_1) - \hat{\mathcal{K}}_b(\boldsymbol{x}_1)\right) \tag{2.27}$$

where $\hat{\mathcal{J}}_b(\boldsymbol{x}_1)$ and $\hat{\mathcal{K}}_b(\boldsymbol{x}_1)$ are given by:

$$\hat{\mathcal{J}}_b(\boldsymbol{x}_1)\chi_a(\boldsymbol{x}_1) = \left[\int d\boldsymbol{x}_2 \chi_b^*(\boldsymbol{x}_2)\frac{1}{|\boldsymbol{r}_1 - \boldsymbol{r}_2|}\chi_b(\boldsymbol{x}_2)\right]\chi_a(\boldsymbol{x}_1) \tag{2.28}$$

and:

$$\hat{\mathcal{K}}_b(\boldsymbol{x}_1)\chi_a(\boldsymbol{x}_1) = \left[\int d\boldsymbol{x}_2 \chi_b^*(\boldsymbol{x}_2)\frac{1}{|\boldsymbol{r}_1 - \boldsymbol{r}_2|}\chi_a(\boldsymbol{x}_2)\right]\chi_b(\boldsymbol{x}_1) \tag{2.29}$$

In practice, the HF approximation requires a more specific treatment of the spin orbitals, which can be divided into two types: (i) restricted spin orbitals with constrained spatial functions for up (α) and down (β) spin functions and (ii) unrestricted spin orbitals with different spatial functions for α and β spins. For a Restricted Hartree Fock (RHF) wavefunction of a closed-shell system, $N/2$ spatial orbitals $\psi(\boldsymbol{r})$ are doubly occupied with the even N number of total electrons of the system and the restricted set of spin orbitals has the form:

$$\{\chi_{2m}(\boldsymbol{x})\} = \begin{cases}\psi(\boldsymbol{r})_m\alpha(\omega) \\ \psi(\boldsymbol{r})_m\beta(\omega)\end{cases} \qquad m = 1, 2, ..., N/2 \tag{2.30}$$

Hence, the wavefunction of the ground state, Ψ_0, can be expressed as a SD of the doubly occupied spatial molecular orbitals $\{\psi_i \mid a = 1, 2, ..., N/2\}$, that will be described in the following paragraphs. Analogously to Equation 2.24 and once the spin dependency is eliminated, the calculation of the molecular orbitals is equivalent to the problem of solving the spatial integro-differential RHF equations:

$$\hat{f}^{RHF}(\boldsymbol{r}_1)\psi_i(\boldsymbol{r}_1) = \mathcal{E}_i^{RHF}\psi_i(\boldsymbol{r}_1) \;\; \forall \; \{\psi_i\} \tag{2.31}$$

where the mean field of the restricted Fock operator $\hat{\mathcal{V}}^{RHF}$ presents a factor 2 on the Coulomb operator with respect to that of the spin orbitals in Equation 2.27.

$$\hat{\mathcal{V}}^{RHF}(\boldsymbol{r}_1) = \sum_b^{N/2}\Big(2\hat{\mathcal{J}}_b(\boldsymbol{r}_1) - \hat{\mathcal{K}}_b(\boldsymbol{r}_1)\Big) \tag{2.32}$$

The resolution of these equations can be numerically achieved in atoms, but not in molecules. To overcome this problem, Clemens C. J. Roothaan and George G. Hall independently demonstrated in 1951[219, 220] that it is possible to introduce a set of K known spatial basis functions $\{\phi_\nu(\boldsymbol{r}_1) \mid \nu = 1, 2, ..., K\}$, typically called atomic orbitals, that allow the transformation of the differential

equation into a set of algebraic equations which can be easily solved by standard matrix techniques. Each unknown Molecular Orbital (MO) $\psi_i(\boldsymbol{r}_1)$ can therefore be expressed as a Linear Combination of Atomic Orbitals (LCAO):

$$\psi_i(\boldsymbol{r}_1) = \sum_{\nu=1}^{K} C_{\nu i}\phi_\nu(\boldsymbol{r}_1) \qquad\qquad i = 1, 2, ..., K \tag{2.33}$$

When performing a HF calculation, the $\{\phi_\nu(\boldsymbol{r}_1)\}$ set of known atomic orbitals is always restricted to a finite set of K basis function due to practical computational reasons, and thus the selection of the basis plays a crucial role in the QM calculation. As the basis set becomes more and more complete, the expansion of the molecular orbitals becomes a more accurate description, converging to the exact eigenfunction of the Fock operator in Equation 2.31.

Hence, Equation 2.31 reduces to the problem of calculating the set of expansion coefficients $\{C_{\nu i}\}$:

$$\hat{f}^{RHF}(\boldsymbol{r}_1)\sum_{\nu=1}^{K} C_{\nu i}\phi_\nu(\boldsymbol{r}_1) = \mathcal{E}_i^{RHF}\sum_{\nu=1}^{K} C_{\nu i}\phi_\nu(\boldsymbol{r}_1) \tag{2.34}$$

Which, multiplying on the left by $\phi_\mu^*(\boldsymbol{r}_1)$ and integrating, becomes:

$$\begin{aligned}\sum_{\nu=1}^{K} C_{\nu i}\int d\boldsymbol{r}_1\phi_\mu^*(\boldsymbol{r}_1)\hat{f}^{RHF}(\boldsymbol{r}_1)\phi_\nu(\boldsymbol{r}_1) &= \mathcal{E}_i^{RHF}\sum_{\nu=1}^{K} C_{\nu i}\int d\boldsymbol{r}_1\phi_\mu^*(\boldsymbol{r}_1)\phi_\nu(\boldsymbol{r}_1)\\ \sum_{\nu=1}^{K} F_{\mu\nu}^{RHF}C_{\nu i} &= \mathcal{E}_i^{RHF}\sum_{\nu=1}^{K} C_{\nu i}S_{\mu\nu}\end{aligned} \tag{2.35}$$

where $F_{\mu\nu}^{RHF}$ and $S_{\mu\nu}$ represent the matrix elements of the restricted Fock operator and the overlap K x K Hermitian matrices, respectively. Although the basis functions $\{\phi_\nu(\boldsymbol{r}_1)\}$ are assumed to be orthonormal, in general they are not orthogonal to each other and thus the overlap matrix is not equivalent to the identity matrix.[210] In fact, the off-diagonal elements present values between -1 and 1, where the sign only depends on the relative sign of the two basis functions and the value, in magnitude, represents the linear dependency between them. The set of equations in Equation 2.35 is the so called Roothaan-Hall equations, which can be written in a more compact form as a single matrix equation:

$$F^{RHF}C = SC\mathcal{E}^{RHF} \tag{2.36}$$

where $\mathcal{E}^{RHF}$ is a diagonal matrix of the orbital energies $\mathcal{E}_i^{RHF}$ and C is a K x K square matrix of the expansion coefficients $C_{\nu i}$ whose columns describe the molecular orbitals as described by Equation 2.33:

$$\mathcal{E}^{RHF} = \begin{pmatrix} \mathcal{E}_1^{RHF} & 0 & \cdots & 0 \\ 0 & \mathcal{E}_2^{RHF} & \cdots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \cdots & \mathcal{E}_K^{RHF} \end{pmatrix} \tag{2.37}$$

$$C = \begin{pmatrix} C_{11} & C_{12} & \cdots & C_{1K} \\ C_{21} & C_{22} & \cdots & C_{2K} \\ \vdots & \vdots & \ddots & \vdots \\ C_{K1} & C_{K2} & \cdots & C_{KK} \end{pmatrix} \tag{2.38}$$

Having defined C, S and $\mathcal{E}^{RHF}$ the only missing piece is the explicit expression of the Fock matrix F^{RHF}. Considering the definition of the restricted Fock matrix elements $F_{\mu\nu}^{RHF}$ in Equation 2.35 and the expression for the Fock operator in Equations 2.25 and 2.32:

$$\begin{aligned} F_{\mu\nu}^{RHF} = & \int d\boldsymbol{r}_1 \phi_\mu^*(\boldsymbol{r}_1) \hat{h}(\boldsymbol{r}_1) \phi_\nu(\boldsymbol{r}_1) \\ & + \int d\boldsymbol{r}_1 \phi_\mu^*(\boldsymbol{r}_1) \left[2\hat{\mathcal{J}}_i(\boldsymbol{r}_1) - \hat{\mathcal{K}}_i(\boldsymbol{r}_1) \right] \phi_\nu(\boldsymbol{r}_1) \end{aligned} \tag{2.39}$$

The first integral corresponds to the elements of the core Hamiltonian matrix $\hat{\mathcal{H}}_{\mu\nu}^{core}$. Its elements involve the one-electron operator $\hat{h}(\boldsymbol{r}_1)$ as described in Equation 2.26:

$$\hat{\mathcal{H}}_{\mu\nu}^{core} = T_{\mu\nu} + V_{\mu\nu}^{nucl} \tag{2.40}$$

Thus, its calculation involves calculating the kinetic and nuclear attraction integrals; $T_{\mu\nu}$ and $V_{\mu\nu}^{nucl}$, respectively:

$$T_{\mu\nu} = -\frac{1}{2} \int d\boldsymbol{r}_1 \phi_\mu^*(\boldsymbol{r}_1) \nabla_1^2 \phi_\nu(\boldsymbol{r}_1) \tag{2.41}$$

$$V_{\mu\nu}^{nucl} = -\int d\boldsymbol{r}_1 \phi_\mu^*(\boldsymbol{r}_1) \left[\sum_{k=1}^{M} \frac{Z_k}{|\boldsymbol{r}_1 - \boldsymbol{R}_k|} \right] \phi_\nu(\boldsymbol{r}_1) \tag{2.42}$$

Since this $\hat{\mathcal{H}}_{\mu\nu}^{core}$ matrix does not depend on the expansion coefficients, and so on the molecular orbitals, it remains constant during the iterative procedure required for solving the Roothaan-Hall equations. Hence, it can be evaluated

only once during the HF calculation. On the flip side, the two-electron integrals $\hat{\mathcal{J}}_i(\boldsymbol{r}_1)$ and $\hat{\mathcal{K}}_i(\boldsymbol{r}_1)$ can be calculated analogously from Equations 2.28 and 2.29 but for spatial coordinates. Thus, substituting Equations 2.28 and 2.29 in the second integral of Equation 2.39:

$$\begin{aligned}\int d\boldsymbol{r}_1 \phi_\mu^*(\boldsymbol{r}_1) \left[2\hat{\mathcal{J}}_i(\boldsymbol{r}_1) - \hat{\mathcal{K}}_i(\boldsymbol{r}_1)\right] \phi_\nu(\boldsymbol{r}_1) = \\ \sum_{i=1}^{N/2} \left[2 \int d\boldsymbol{r}_1 d\boldsymbol{r}_2 \phi_\mu^*(\boldsymbol{r}_1)\phi_\nu(\boldsymbol{r}_1) \frac{1}{|\boldsymbol{r}_1 - \boldsymbol{r}_2|} \psi_i^*(\boldsymbol{r}_2)\psi_i(\boldsymbol{r}_2) \right. \\ \left. - \int d\boldsymbol{r}_1 d\boldsymbol{r}_2 \phi_\mu^*(\boldsymbol{r}_1)\psi_i(\boldsymbol{r}_1) \frac{1}{|\boldsymbol{r}_1 - \boldsymbol{r}_2|} \psi_i^*(\boldsymbol{r}_2)\phi_\nu(\boldsymbol{r}_2)\right]\end{aligned} \tag{2.43}$$

and then solving for the set of known atomic orbitals in Equation 2.33:

$$\begin{aligned}\int d\boldsymbol{r}_1 \phi_\mu^*(\boldsymbol{r}_1) \left[2\hat{\mathcal{J}}_i(\boldsymbol{r}_1) - \hat{\mathcal{K}}_i(\boldsymbol{r}_1)\right] \phi_\nu(\boldsymbol{r}_1) = \\ \sum_{i=1}^{N/2} \sum_{\sigma\lambda}^{K} C_{\sigma i}^* C_{\lambda i} \left[2 \int d\boldsymbol{r}_1 d\boldsymbol{r}_2 \phi_\mu^*(\boldsymbol{r}_1)\phi_\nu(\boldsymbol{r}_1) \frac{1}{|\boldsymbol{r}_1 - \boldsymbol{r}_2|} \phi_\sigma^*(\boldsymbol{r}_2)\phi_\lambda(\boldsymbol{r}_2) \right. \\ \left. - \int d\boldsymbol{r}_1 d\boldsymbol{r}_2 \phi_\mu^*(\boldsymbol{r}_1)\phi_\lambda(\boldsymbol{r}_1) \frac{1}{|\boldsymbol{r}_1 - \boldsymbol{r}_2|} \phi_\sigma^*(\boldsymbol{r}_2)\phi_\nu(\boldsymbol{r}_2)\right]\end{aligned} \tag{2.44}$$

This expression can be written in a more compact way by considering the definition of density matrix $P_{\lambda\sigma}$ in Equation 2.45. This matrix specifies the charge density function $\rho(\boldsymbol{r})$, whose integral over space gives the probability of finding any electron at $\boldsymbol{r}$ within the interval $d\boldsymbol{r}$.

$$P_{\lambda\sigma} = \sum_{i=1}^{N/2} 2C_{\sigma i}^* C_{\lambda i} \tag{2.45}$$

Hence, the Fock matrix elements $F_{\mu\nu}^{RHF}$ in Equation 2.39 can be expressed as:

$$\begin{aligned}F_{\mu\nu}^{RHF} = \hat{\mathcal{H}}_{\mu\nu}^{core} \\ + \sum_{\sigma\lambda}^{K} P_{\lambda\sigma} \left[\int d\boldsymbol{r}_1 d\boldsymbol{r}_2 \phi_\mu^*(\boldsymbol{r}_1)\phi_\nu(\boldsymbol{r}_1) \frac{1}{|\boldsymbol{r}_1 - \boldsymbol{r}_2|} \phi_\sigma^*(\boldsymbol{r}_2)\phi_\lambda(\boldsymbol{r}_2) \right. \\ \left. - \frac{1}{2} \int d\boldsymbol{r}_1 d\boldsymbol{r}_2 \phi_\mu^*(\boldsymbol{r}_1)\phi_\lambda(\boldsymbol{r}_1) \frac{1}{|\boldsymbol{r}_1 - \boldsymbol{r}_2|} \phi_\sigma^*(\boldsymbol{r}_2)\phi_\nu(\boldsymbol{r}_2)\right]\end{aligned} \tag{2.46}$$

Equation 2.46 shows that the Fock matrix depends on the density matrix which, at the same time, depends on the expansion coefficients as shown in

Equation 2.45. Hence, the Roothaan-Hall equation in Equation 2.36 is non-linear and needs to be solved iteratively. To do so, it is possible to perform a change of basis such that:

$$X^{\dagger}SX = S' = \mathbb{1} \tag{2.47}$$

where X is the transformation matrix obtained by diagonalization of S and $\mathbb{1}$ is the identity matrix. Within this basis:

$$C = XC' \tag{2.48}$$

Hence, Equation 2.36 can be written as:

$$F^{RHF}XC' = SXC'\mathcal{E}^{RHF} \tag{2.49}$$

which multiplying on the left by $X^{\dagger}$ and substituting Equation 2.47 becomes:

$$\begin{aligned} X^{\dagger}F^{RHF}XC' &= X^{\dagger}SXC'\mathcal{E}^{RHF} \\ F^{RHF\prime}C' &= C'\mathcal{E}^{RHF} \end{aligned} \tag{2.50}$$

This equation, the transformed Roothaan-Hall equation, presents an eigenvalue problem, where the eigenvectors C' and eigenvalues $\mathcal{E}^{RHF}$ can be obtained by diagonalizing the transformed restricted Fock matrix $F^{RHF\prime}$.

As mentioned before, the Roothaan-Hall equation needs to be solved in an iterative procedure denominated the Self Consistent Field (SCF) process. Once the basis set $\{\phi_{\nu}(\boldsymbol{r}_1)\}$ and the molecule (R_M, Z_M and N) are specified, it is possible to compute the S overlap and the $\hat{\mathcal{H}}^{core}$ core Hamiltonian matrices by means of Equations 2.35 and 2.40, respectively. Then, the S overlap matrix is diagonalized to obtain the X transformation matrix. Consecutively, an initial guess for the P density matrix is generated by means of an initial set of expansion coefficients C and then, the F^{RHF} Fock matrix is built from Equation 2.46. From this matrix and the X transformation matrix, it is possible to obtain the transformed Fock matrix $F^{RHF\prime}$ which, upon diagonalization, allows the calculation of the C' transformed expansion coefficients matrix and the $\mathcal{E}^{RHF}$ orbital energies. By transforming back C' into C by means of Equation 2.48, it is possible to calculate a new P density matrix and determine whether the SCF procedure is converged by comparing, for instance, the density matrices between iterations.

Once the SCF procedure is finished, it is possible to use the converged matrices to compute different properties of the system. For instance, assuming that the basis functions are centred on the atomic nuclei, then the corresponding number of electrons associated with a given atom A in a molecule are

obtained by summing over all basis functions centred on that atom by means of the Mulliken population analysis. Hence, the net q_A charges associated to that specific atom are given by:

$$q_A = Z_A - \sum_{\mu \in A} \sum_{\nu}^{N_{AOs}} P_{\mu\nu} S_{\mu\nu} \tag{2.51}$$

where Z_A represents the charge of the A atomic nucleus; $P_{\mu\nu}S_{\mu\nu}$, the number of electrons associated with ϕ_μ and the index of summation indicates that the summation is produced exclusively on the basis functions centred on atom A, leading to a strong dependence on the basis set when computing the Mulliken charges.

Furthermore, while the use of the Roothaan-Hall equations is restricted to doubly occupied spatial molecular orbitals, it is also possible to apply the HF approximation to a set of unrestricted spin orbitals in which the N^α α and the N^β β electrons are described by a different set of spatial orbitals ($\{\psi^\alpha(\boldsymbol{r}_1)\}$ and $\{\psi^\beta(\boldsymbol{r}_1)\}$). This formulation is described by the Pople-Nesbet equations, which are the unrestricted generalization of the Roothaan-Hall equations:[210, 221]

$$F^\alpha C^\alpha = S C^\alpha \mathcal{E}^\alpha \qquad F^\beta C^\beta = S C^\beta \mathcal{E}^\beta \tag{2.52}$$

where, analogously to the Roothaan-Hall formulation, S represents the overlap matrix as defined in Equation 2.35; $\mathcal{E}^\alpha$ and $\mathcal{E}^\beta$, the diagonal matrices of the orbital energies and C^α and C^β, the K x K square matrices of the expansion coefficients. Furthermore, F^α and F^β are the matrix representation of the Fock operators $\hat{f}^\alpha$ and $\hat{f}^\beta$ in the basis $\{\phi_\nu(\boldsymbol{r}_1)\}$:

$$F^\alpha_{\mu\nu} = \int d\boldsymbol{r}_1 \phi^*_\mu(\boldsymbol{r}_1) \hat{f}^\alpha(\boldsymbol{r}_1) \phi_\nu(\boldsymbol{r}_1) \quad F^\beta_{\mu\nu} = \int d\boldsymbol{r}_1 \phi^*_\mu(\boldsymbol{r}_1) \hat{f}^\beta(\boldsymbol{r}_1) \phi_\nu(\boldsymbol{r}_1) \tag{2.53}$$

where the Fock operators are given by:

$$\begin{aligned} \hat{f}^\alpha(\boldsymbol{r}_1) &= \hat{h}(\boldsymbol{r}_1) + \sum_{a=1}^{N^\alpha} [\hat{\mathcal{J}}^\alpha_a(\boldsymbol{r}_1) - \hat{\mathcal{K}}^\alpha_a(\boldsymbol{r}_1)] + \sum_{b=1}^{N^\beta} \hat{\mathcal{J}}^\beta_b(\boldsymbol{r}_1) \\ \hat{f}^\beta(\boldsymbol{r}_1) &= \hat{h}(\boldsymbol{r}_1) + \sum_{b=1}^{N^\beta} [\hat{\mathcal{J}}^\beta_b(\boldsymbol{r}_1) - \hat{\mathcal{K}}^\beta_b(\boldsymbol{r}_1)] + \sum_{a=1}^{N^\alpha} \hat{\mathcal{J}}^\alpha_b(\boldsymbol{r}_1) \end{aligned} \tag{2.54}$$

and the $\hat{h}(\boldsymbol{r}_1)$ core Hamiltonian, the $\hat{\mathcal{J}}^\alpha_a(\boldsymbol{r}_1)$ and $\hat{\mathcal{J}}^\beta_b(\boldsymbol{r}_1)$ Coulomb and $\hat{\mathcal{K}}^\alpha_a(\boldsymbol{r}_1)$ and $\hat{\mathcal{K}}^\beta_b(\boldsymbol{r}_1)$ exchange operators are analogous to those in Equation 2.26 and 2.43, respectively, for each $\{\psi^\alpha(\boldsymbol{r}_1)\}$ and $\{\psi^\beta(\boldsymbol{r}_1)\}$ set of spatial orbitals.

These equations can be evaluated following a SCF procedure, similarly to the Roothaan-Hall equations, solving the two matrix eigenvalue problems simultaneously, since F^{α} and F^{β} depend on both C^{α} and C^{β}. Even thought, unrestricted HF is typically employed to describe open-shell states of molecules with a different amount of α and β electrons, in the very special case where the number of α and β electrons is the same, the Pople-Nesbet equations might present two independent solutions: (i) the restricted solution of the Roothaan-Hall equations, which always exists and necessarily results if the initial guesses for the α and β density matrices are same, and, under certain conditions, (ii) the lower in energy unrestricted solution.

2.1.4 The Variational Principle

The SCF procedure satisfies the Variational Principle, which ensures the finding of the best possible approximate solutions for the TISE:[210, 222, 223]

"Given a normalized ground state wavefunction $\Phi_0(\boldsymbol{x_1}, ..., \boldsymbol{x_N})$ that satisfies the appropriate boundary conditions of the system, the expectation value of the Hamiltonian $\hat{\mathcal{H}}$ is an upper bound to the exact ground state energy E_0."

$$E_0 \leq \frac{\int d\boldsymbol{x_1} d\boldsymbol{x_2} ... d\boldsymbol{x_N} \Phi_0^* \hat{\mathcal{H}} \Phi_0}{\int d\boldsymbol{x_1} d\boldsymbol{x_2} ... d\boldsymbol{x_N} \Phi_0^* \Phi_0} \tag{2.55}$$

Moreover, even though the variational method is stated for the ground state energy, it is possible to generalise it to the excited states of state-specific methods so that the expectation values of applying the Hamiltonian operator to the trial wavefunction are always larger or equal to the electronic energy of the excited states of the system.

Furthermore, the Linear Variational Method demonstrates that the difference between the exact solution and the approximated one arises from the completeness of the basis set. Hence, for variational computational methods, the use of an infinite basis set would result on the exact representation of the wavefunction and the use of a finite basis set would imply an approximate solution. Nonetheless, the larger the basis set, the more accurate the result of the calculation but also the larger the computational time. Hence, the selection of the basis set plays a crucial role on the accuracy of the calculations, and it is an issue that needs to be carefully addressed before-hand.

2.1.5 The Molecular Orbitals Theory and Basis Sets

In previous sections, it was discussed how the MOs can be expanded in a set of known atomic orbitals, denominated basis set, by means of the LCAO in Equation 2.33. This approach only leads to an approximation of the exact MOs since the exact solution for a particular level of theory is only achieved in the complete basis set limit, which is given by an infinite basis set. Nonetheless, such a basis set is not reachable since it implies the use of a huge number of functions and an immense computational effort. This section will cover a mathematical description of these atomic orbitals together with some technical considerations for their use.[210, 213]

There are two main types of basis sets: Slater Type Orbitals (STO) and Gaussian Type Orbitals (GTO) given by:

$$\begin{aligned} &\text{STO} && \phi_{\zeta,n,l,m}(\boldsymbol{r},\theta,\varphi) = NY_{l,m}(\theta,\varphi)\,\boldsymbol{r}^{n-1}e^{-\zeta\boldsymbol{r}} \\ &\text{GTO} && \phi_{\zeta,n,l,m}(\boldsymbol{r},\theta,\varphi) = NY_{l,m}(\theta,\varphi)\,\boldsymbol{r}^{2n-2-l}e^{-\zeta\boldsymbol{r}^2} \end{aligned} \tag{2.56}$$

where ζ represents the number of functions included to describe an atom; n, l and m, the quantum numbers; $\boldsymbol{r}$, θ and φ, the polar coordinates; N, the normalization constant and $Y_{l,m}$, a spherical harmonic function. GTO have problems representing the proper behaviour near the nucleus as they fall-off too rapidly far from the nucleus compared with STO basis sets and need three times more functions to reach the same accuracy. However, they are preferred in terms of computational efficiency since the two-electron integrals can be more easily calculated.

When using either STO or GTO, the accuracy will be given by the number of functions used in the expansion. While there is always a minimum amount of them, referred to as minimum basis set, where only those functions needed to contain all the non-virtual electrons of the neutral atoms are used, the result given by them is rather poor and more functions need to be added to the basis set. This is achieved by doubling, Double Zeta (DZ) type basis; tripling, Triple Zeta (TZ) type basis; or even quadrupling or quintupling the functions of the minimum basis set. Nevertheless, this approach can increase significantly the computational time since it scales with the number of atoms. Therefore, a further approximation, in which only the valence orbitals are duplicated a few times, can be used. This split valence basis can be justified by the fact that core orbitals do not participate in the chemical bonding and are very similar to those of the isolated atoms. In addition, it is also possible to include extra polarization and diffuse functions in the basis set. The former implies the use of higher angular momentum functions to account for charge polarization effects and electronic correlation, while the latter describes loosely bound electrons

and properties dependent on the wavefunction tail by means of basis functions with small exponents.

Up to this point, it has been assumed that all the coefficients of the molecular orbital expansion need to be optimized to minimize the energy. However, the GTO present a functional behaviour different from the expected one for the MO since core-electrons account for a large part of the total energy and so this minimization will tend to increase the participation of core orbitals and decrease that of the valence ones, which are, in fact, the most relevant for the description of the properties of the system. Therefore, a bigger amount of functions would be needed to achieve an accurate result and a corresponding increase in the computational time. For this reason, and taking into account that the variation of core orbitals with respect to the chemical bonding situation is very small, and therefore the coefficients in front of them also change very little, it is a good approximation to describe this core orbitals as a linear combination of Primitive Gaussian Type Orbitals (PGTO), the full set of basis functions, in which the a_i expansion coefficients are fixed. This fixed linear combination of PGTO is known as a basis set contraction, resulting in Contracted Gaussian Type Orbitals (CGTO), and suppose a reduction of the computational time.

$$\phi(CGTO) = \sum_{i}^{k} a_i \phi_i(PGTO) \tag{2.57}$$

Along this thesis only the so-called Pople $k-spqG(h,H)$ basis sets have been employed, that are split valence basis sets in which k indicates the contraction of the core orbitals, how many PGTO are used to represent them. The *spq* values correspond to both: (i) the degree of splitting of the basis set for s- and p- functions, meaning *sp* a DZ basis set and *spq* a TZ basis set, and (ii) how many PGTO are used for their representation. Moreover, letters after the G letter, between parenthesis and separated by a comma, indicate the polarization functions (d- or f- type) added to heavy (h) and hydrogen atoms (H), respectively. Finally, it is possible to include diffuse functions, represented by one or two "+" signs before the G letter.

2.1.6 The Density Functional Theory

Within the Hohenberg-Kohn formulation, DFT is a QM method that reduces the computational time with respect to those based in the wavefunction by substituting the multi-electronic problem with $4N$ variables (three spatial and spin coordinates for each electron) by a one-particle problem with three spatial coordinates independently on the number of electrons of the system (N),

since all the properties can be obtained from the electronic density, given by Equation 2.58.[213]

$$\rho(\boldsymbol{r}) = N \int_0^\infty d\boldsymbol{r_1}...d\boldsymbol{r_N} |\Psi\left(\boldsymbol{r_1}, ..., \boldsymbol{r_N}\right)|^2 \tag{2.58}$$

The position of the local maxima electron density, which is positive definite, correspond to the nuclear positions and the intensity to their respective nuclear charges. Furthermore, the electron density must be such that its integral over the whole space provides the number of electrons N of the system:

$$\int_0^\infty d\boldsymbol{r}\rho(\boldsymbol{r}) = N \tag{2.59}$$

The Hohenberg-Kohn Theorems

DFT is based on two basic theorems given by Walter Kohn and Pierre Hohenberg in 1964.[224] The first of them, which can be demonstrated by reduction to the absurd[223] and that was generalized in 1979 by Mel Philip Levy[225] to degenerated states, establishes that the wavefunction and the total energy of the system is a functional of the electronic ground-state density or which is the same:

"Any observable of a non-degenerated stationary state can be written, exactly, as a functional of the electron density of the ground state."

As stated by Hohenberg and Kohn, this theorem only holds for non-degenerated ground stationary states with a N- and v-representable $\rho(\boldsymbol{r})$ electronic densities. This means that $\rho(\boldsymbol{r})$ must be a positive function in all the space whose integral is equal to the N total number of electrons of the system (N-representable). Moreover, there exists an external potential from which $\rho(\boldsymbol{r})$ can be derived (v-representable). Whenever this theorem can be applied:

$$E[\rho(\boldsymbol{r})] = T[\rho(\boldsymbol{r})] + V_{ne}[\rho(\boldsymbol{r})] + V_{ee}[\rho(\boldsymbol{r})] + (V_{nn}) \tag{2.60}$$

where $T[\rho(\boldsymbol{r})]$ represents the electronic kinetic energy; $V_{ne}[\rho(\boldsymbol{r})]$, the electrostatic attraction between nuclei and electrons; $V_{ee}[\rho(\boldsymbol{r})]$, the electronic repulsion energy; and V_{nn}, the nuclear repulsion energy which, within the Born-Oppenheimer framework, is just a constant. Moreover, since $T[\rho(\boldsymbol{r})]$ and $V_{ee}[\rho(\boldsymbol{r})]$ do not depend on the external potential they are typically gathered into the Hohenberg-Kohn functional $F_{HK}[\rho(\boldsymbol{r})]$, while, in isolated molecules,

$V_{ne}[\rho(\boldsymbol{r})]$ can be expressed in terms of the electronic density and the external potential $v^{ext}(\boldsymbol{r})$:

$$\frac{\delta V_{ne}[\rho(\boldsymbol{r})]}{\delta\rho(\boldsymbol{r})} = v^{ext}(\boldsymbol{r}) \tag{2.61}$$

Hence, Equation 2.60 can be rewritten as:

$$E[\rho(\boldsymbol{r})] = F_{HK}[\rho(\boldsymbol{r})] + \int d\boldsymbol{r}\rho(\boldsymbol{r})v^{ext}(\boldsymbol{r}) \tag{2.62}$$

The second Hohenberg-Kohn theorem represents a general version of the variation principle, and states that:

"The electronic density of a non-degenerate ground state can be calculated, exactly, by determining the density that minimizes the energy of the ground state."

Hence, for a trial density $\tilde{\rho}(\boldsymbol{r})$, given that it is N- and v-representable:

$$E[\tilde{\rho}(\boldsymbol{r})] \geq E_0 \tag{2.63}$$

And so, the exact ground state density can be obtained by minimizing the energy with respect to the electronic density:

$$\left[\frac{\delta E[\rho(\boldsymbol{r})]}{\delta\rho(\boldsymbol{r})}\right] = 0 \tag{2.64}$$

This minimization must ensure the N-representability condition, which can be done by apply the restriction in Equation 2.59 my means of Lagrange's Method of Undetermined Multipliers. This method leads to the formulation of the fundamental equation of the DFT:

$$\mu = \frac{\delta E[\rho(\boldsymbol{r})]}{\delta\rho(\boldsymbol{r})} = v^{ext}(\boldsymbol{r}) + \frac{\delta F_{HK}[\rho(\boldsymbol{r})]}{\delta\rho(\boldsymbol{r})} \tag{2.65}$$

where μ represents the Lagrange multiplier.

As happens with the HF approximation, the DFT lacks an analytic expression for the $F_{HK}[\rho(\boldsymbol{r})]$ functional of the exact density as the non-classical contribution to the electron-electron interaction, which includes self-interaction correction, exchange, and Coulomb correlation effects, is neglected. Therefore, approximate expressions have been proposed so that the problem becomes that of determining the density that minimizes $E[\rho(\boldsymbol{r})]$ for a given molecular system. At this point, it is important to emphasize that, although exact in

its formulation, DFT needs to be solved in an approximate way by means of empirically parametrised equations to describe the Exchange Correlation (XC) contribution that, in practice are responsible for the reduction of the computational time.[211] This implies that, when using approximate functionals, DFT formally no longer respects the variational principle.

Kohn-Sham Approach

As mentioned in the previous section, there is no analytical expression for the $F_{HK}[\rho(\boldsymbol{r})]$. This problem was addressed by Walter Kohn and Lu Jeu Sham in 1965[226] by reducing the intractable many-body problem of interacting electrons in a static external potential to a tractable problem of non-interacting electrons moving in an effective potential v^{eff}, which includes the external potential, the effects of the Coulomb interactions between the electrons and the exchange and correlation interactions. They proposed a methodology to calculate the electron density from the kinetic energy for a monodeterminant wavefunction, for which they were awarded the Nobel prize in 1998, together with the *development of computational methods in quantum chemistry* carried out by John Pople during the second half of the 20^{th} century.

Let's consider a reference system of N non-interacting electrons moving in an external potential v_r such that, when applied to the system, generates a reference wavefunction Ψ_r such that its electron density is the same that the electron density of the real system:

$$\rho[\Psi_r] = \rho^{real} \tag{2.66}$$

For this ideal system in which electrons only interact with the nuclei, the HF method, described in Section 2.1.3, presents an exact solution, that of a monoelectronic Hamiltonian:

$$\mathcal{H}_r = \sum_{i=1}^{N} \hat{h}(\boldsymbol{r}) = -\frac{1}{2}\sum_{i=1}^{N}\left(\nabla_i^2 + v_r(\boldsymbol{r})\right) \tag{2.67}$$

Furthermore, the exact Ψ_r wavefunction obeying the antisymmetry principle is given by a SD, as expressed in Equation 2.23, formed by one-electron spatial MOs instead of spin orbitals. Hence, analogously to the HF Equation 2.24, the (orthonormal) set of molecular orbitals $\{\psi_i(\boldsymbol{r}) \mid i = 1, 2, ..., N\}$ that provides the wavefunction Ψ_r is obtained by solving the one-electron TISE:

$$\left[-\frac{1}{2}\nabla^2 + v_r(\boldsymbol{r})\right]\psi_i(\boldsymbol{r}) = \mathcal{E}_i\psi_i(\boldsymbol{r}) \quad \forall\ \{\psi_i(\boldsymbol{r})\} \tag{2.68}$$

and so the exact density and the kinetic energy for the non-interacting system are given by:

$$\rho^{real}(\boldsymbol{r}) = \sum_{i=1}^{N} |\psi_i(\boldsymbol{r})|^2 \tag{2.69}$$

and

$$T_r[\rho(\boldsymbol{r})] = -\frac{1}{2}\sum_{i=1}^{N} \int d\boldsymbol{r}\psi_i^*(\boldsymbol{r})\nabla^2\psi_i(\boldsymbol{r}) \tag{2.70}$$

In this reference system, the fundamental equation of the DFT in Equation 2.65 becomes:

$$\mu = \frac{\delta E_r[\rho(\boldsymbol{r})]}{\delta\rho(\boldsymbol{r})} = \frac{\delta T_r[\rho(\boldsymbol{r})]}{\delta\rho(\boldsymbol{r})} + v_r(\boldsymbol{r}) \tag{2.71}$$

and its total energy is given by:

$$E_r[\rho(\boldsymbol{r})] = \sum_{i=1}^{N} \mathcal{E}_i = T_r[\rho(\boldsymbol{r})] + \int d\boldsymbol{r}\rho(\boldsymbol{r})v_r(\boldsymbol{r}) \tag{2.72}$$

This expression differs from that of the real system given by Equations 2.60 and 2.62 on (i) the kinetic energy term since the kinetic energy of the reference system is different form the real kinetic energy and (ii) the presence of the $V_{ee}[\rho(\boldsymbol{r})]$ electronic repulsion term, which can be approximated by the classical Coulomb's expression as a functional of the density:

$$J[\rho(\boldsymbol{r})] = \frac{1}{2}\int\int d\boldsymbol{r}_1 d\boldsymbol{r}_2 \frac{\rho(\boldsymbol{r}_1)\rho(\boldsymbol{r}_2)}{|\boldsymbol{r}_1 - \boldsymbol{r}_2|} \tag{2.73}$$

Hence, Equation 2.60 can be rewritten by adding and subtracting the kinetic energy functional of the non-interacting system $T_r[\rho(\boldsymbol{r})]$, and the coulomb repulsion term $J[\rho(\boldsymbol{r})]$:

$$\begin{aligned} E[\rho(\boldsymbol{r})] = T_r[\rho(\boldsymbol{r})] + \int d\boldsymbol{r}\rho(\boldsymbol{r})v_r(\boldsymbol{r}) + J[\rho(\boldsymbol{r})] \\ + \Big[T[\rho(\boldsymbol{r})] - T_r[\rho(\boldsymbol{r})]\Big] + \Big[V_{ee}[\rho(\boldsymbol{r})] - J[\rho(\boldsymbol{r})]\Big] \end{aligned} \tag{2.74}$$

where the last two terms in squared brackets correspond to:

$$\begin{aligned} \text{Correlation kinetic energy} \quad & T_C[\rho(\boldsymbol{r})] = T[\rho(\boldsymbol{r})] - T_r[\rho(\boldsymbol{r})] \\ \text{Exchange correlation potential} \quad & W_{xc}[\rho(\boldsymbol{r})] = V_{ee}[\rho(\boldsymbol{r})] - J[\rho(\boldsymbol{r})] \end{aligned} \tag{2.75}$$

These two expressions can be gathered in a single term, referred to as the XC energy (Equation 2.76), which arises from all the complicated many-body effects, all the energy contributions that are not considered by a non-interacting system of electrons and the classical electron-electron interaction: (i) the dependency of the probability density of each other electron on the probability density of a specific one, (ii) the self-interaction, (iii) the interacting parts of the kinetic energy not supplied by $T_r[\rho(\boldsymbol{r})]$, and (iv) exchange energy coming form the asymmetry of the wavefunction in Pauli principle.[213] This energetic term, that is further discussed in Section 2.1.6, needs to be obtained empirically with different approximate functionals.

$$E_{xc}[\rho(\boldsymbol{r})] = T_C[\rho(\boldsymbol{r})] + W_{xc}[\rho(\boldsymbol{r})] \tag{2.76}$$

Similarly to $V_{ne}[\rho(\boldsymbol{r})]$ in Equation 2.61, this energy can be expressed in terms of the electronic density and the XC potential $v_{xc}(\boldsymbol{r})$:

$$\frac{\delta E_{xc}[\rho(\boldsymbol{r})]}{\delta\rho(\boldsymbol{r})} = v_{xc}(\boldsymbol{r}) \tag{2.77}$$

Considering the definitions in Equations 2.73, 2.75 and 2.76, and within the Local Density Approximation (LDA), Equation 2.74 can be written as:

$$E[\rho(\boldsymbol{r})] = T_r[\rho(\boldsymbol{r})] + \int d\boldsymbol{r}\rho(\boldsymbol{r})v_r(\boldsymbol{r}) + \frac{1}{2}\int\int d\boldsymbol{r}_1 d\boldsymbol{r}_2 \frac{\rho(\boldsymbol{r}_1)\rho(\boldsymbol{r}_2)}{|\boldsymbol{r}_1 - \boldsymbol{r}_2|} + E_{xc} \tag{2.78}$$

And so, applying the fundamental equation of the DFT in Equation 2.65:

$$\frac{\delta E[\rho(\boldsymbol{r})]}{\delta\rho(\boldsymbol{r})} = \frac{\delta T_r[\rho(\boldsymbol{r})]}{\delta\rho(\boldsymbol{r})} + v^{ext}(\boldsymbol{r}) + \int d\boldsymbol{r}_2 \frac{\rho(\boldsymbol{r}_2)}{|\boldsymbol{r}_1 - \boldsymbol{r}_2|} + \frac{\delta E_{xc}[\rho(\boldsymbol{r})]}{\delta\rho(\boldsymbol{r})} = \mu \tag{2.79}$$

which, defining the Coulomb potential $\varphi(\boldsymbol{r})$ as:

$$\varphi(\boldsymbol{r}) = v^{ext}(\boldsymbol{r}) + \int d\boldsymbol{r}_2 \frac{\rho(\boldsymbol{r}_2)}{|\boldsymbol{r}_1 - \boldsymbol{r}_2|} \tag{2.80}$$

and taking into account the definition of $v_{xc}(\boldsymbol{r})$ in Equation 2.77, becomes:

$$\begin{aligned} \frac{\delta T_r[\rho(\boldsymbol{r})]}{\delta\rho(\boldsymbol{r})} + \varphi(\boldsymbol{r}) + v_{xc}(\boldsymbol{r}) &= \mu \\ \frac{\delta T_r[\rho(\boldsymbol{r})]}{\delta\rho(\boldsymbol{r})} + v_{eff}(\boldsymbol{r}) &= \mu \end{aligned} \tag{2.81}$$

Hence, it is possible to conclude, by comparison with Equation 2.71, that the equation for a system of interacting electrons is the same as that of a system

of non-interacting electrons under an effective potential $v_{eff}(\boldsymbol{r})$. Hence, the solution for Equation 2.81 corresponds to the expression in Equation 2.68, where $v_r(\boldsymbol{r})$ is substituted by $v_{eff}(\boldsymbol{r})$:

$$\begin{aligned} \hat{h}^{KS}\psi_i^{KS}(\boldsymbol{r}) &= \mathcal{E}_i^{KS}\psi_i^{KS}(\boldsymbol{r}) \\ \left[-\frac{1}{2}\nabla^2 + v_{eff}(\boldsymbol{r})\right]\psi_i^{KS}(\boldsymbol{r}) &= \mathcal{E}_i^{KS}\psi_i^{KS}(\boldsymbol{r}) \quad \forall \{\psi_i^{KS}(\boldsymbol{r})\} \end{aligned} \tag{2.82}$$

where $\psi_i^{KS}(\boldsymbol{r})$ represents the Kohn-Sham (KS) orbitals and $v_{eff}(\boldsymbol{r})$ is given by:

$$v_{eff}(\boldsymbol{r}) = v_r(\boldsymbol{r}) + \int d\boldsymbol{r}_2 \frac{\rho(\boldsymbol{r}_2)}{|\boldsymbol{r}_1 - \boldsymbol{r}_2|} + v_{xc}(\boldsymbol{r}) \tag{2.83}$$

Analogously to the HF method, the KS equations can be solved iteratively by starting from a trial set of initial KS orbitals, which can be used to calculate a trial electron density by means of Equation 2.69. This trial wavefunction can be then employed to calculate the effective potential using Equation 2.83 that is required to solve the eigenvalue problem in Equation 2.82, leading to a new set of KS orbitals.

As a final remark, it is important to emphasize the main difference between HF and DFT. While the former is an intentionally approximate theory, with a modest underlying motivation that is the ability to solve the relevant equations exactly; the latter is an exact theory, but the relevant equations must be solved approximately because the form of a crucial operator is yet to be discovered.[211]

Exchange-Correlation Functionals

The XC functional is a key ingredient of the DFT method since it allows computing the XC potential of the system. Even though the existence of an exact and universal functional has been proven, and its properties are well-defined, its shape is yet to be discovered. Some of these properties are:[213]

1. The self-iteration-free energy. For instance, the exchange and Coulomb energies for one-electron systems should exactly cancel each other and the correlation energy has to be zero.

2. Recovery of the uniform electron gas when the density becomes constant.

3. Linear scaling of the electronic coordinates should result in a similar linear scaling of the exchange energy $E_x[\rho(\boldsymbol{r})]$.

$$\begin{aligned} \rho_\lambda(x,y,z) &= \lambda^3 \rho_\lambda(\lambda x, \lambda y, \lambda z) \\ E_x[\rho_\lambda(\boldsymbol{r})] &= \lambda E_x[\rho(\boldsymbol{r})] \end{aligned} \tag{2.84}$$

4. No direct scaling for the correlation energy $E_c[\rho(\boldsymbol{r})]$. However, scaling the electronic coordinates by a factor larger than 1 should increase the magnitude of the correlation.

$$-E_c[\rho_\lambda(\boldsymbol{r})] > -\lambda E_c[\rho(\boldsymbol{r})] \qquad \text{when } \lambda > 1 \tag{2.85}$$

5. The correlation energy for a finite system approaches a negative constant when the scaling parameter goes to infinity.

The development of the XC functionals is based on fitting some parameters on mathematical expressions focusing on fulfilling the known properties that the exact functional should have, on the reproduction of some experimental values or on a combination of both.

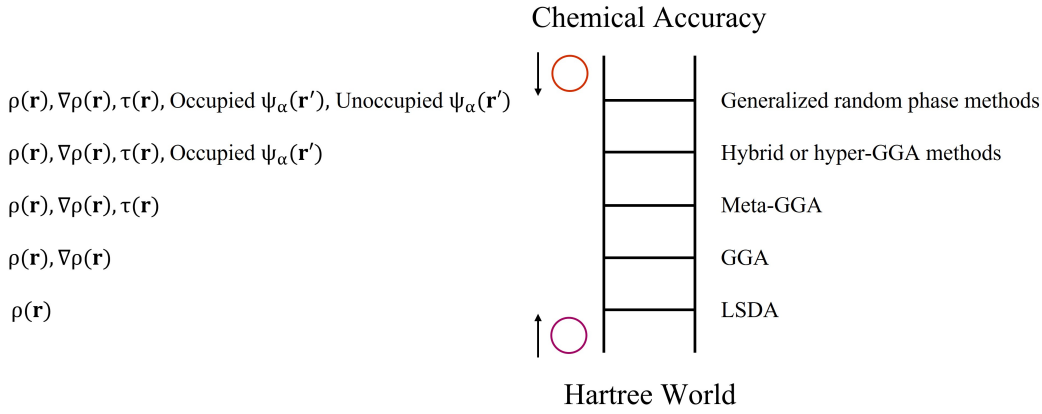


Figure 2.1: Jacob's ladder of density functional approximations.

Figure 2.1 represents the different families of XC functionals. The most simple of them are the LDA and its more general version, the Local Spin Density Approximation (LSDA). These XC functionals, which were first introduced by Walter Kohn and Lu Jeu Sham in 1965,[226] are based on the uniform electron gas model and only require the density for computing the XC energy. Another functional belonging to this rung is the VWN, described by Vosko, Wilk and

Nusair.[227] The Generalized Gradient Approximation (GGA) XC functionals additionally introduce the gradient of the density ($\nabla\rho(\boldsymbol{r})$), thus, allowing to account for slow variations on the density of the system and Meta-GGAs, apart from the previously mentioned properties, also consider the kinetic energy density given by Equation 2.86. LYP,[228, 229] constructed by Lee, Yang and Parr, is an example of this level of the ladder.

$$\tau(\boldsymbol{r}) = \sum_{i=1}^{N} |\nabla\psi_i(\boldsymbol{r})|^2 \tag{2.86}$$

Hybrid or hyper-GGA methods are found upper in Jacob's ladder. They are based on the Adiabatic Connection Formula[230] that arises from the exact connection between the XC energy and the corresponding hole potential connecting the non-interacting reference and the actual system. In this approach, the exchange energy is, exactly, that given by the HF theory for the occupied KS orbitals, while a good approximation of the correlation is obtained by means of mixing the HF exchange with the LDA. Some of the functionals used along this thesis, B3LYP[231, 232] and its long-ranged corrected version CAM-B3LYP,[233] belong to this level of the ladder and are expressed as a linear function of lower rungs in the ladder.

$$E_{xc}^{B3LYP} = aE_x^{HF} + (1-a)E_x^{LDA} + b\Delta E_x^{B88X} + (1-c)E_c^{GGA} \tag{2.87}$$

$$\begin{aligned} E_{xc}^{CAM-B3LYP} =& \alpha E_x^{HF} + (1-\alpha)E_x^{LDA} + c^{B88}\Delta E_x^{B88} + \\ & \beta(E_x^{LR,HF} - E_x^{LR,B88}) + cE_c^{LYP} + (1-c)E_c^{VWN} \end{aligned} \tag{2.88}$$

with $a = 0.2$, $b = 0.72$, $c = 0.81$, $\alpha = 0.19$, $\beta = 0.46$ and $c^{B88} = 0.81$.

Generalized random phase methods are found at the top of the Jacob's ladder. They require the knowledge of the full KS orbitals, not only the occupied but also the unoccupied ones, which significantly improves the description of dispersion interactions. The OEP2 functional[234–236] belongs to this rung, and it is computationally equivalent to an iterative Møller–Plesset second-order Perturbation theory (MP2) calculation.[237]

2.1.7 The Time Dependent Density Functional Theory

As stated by its first theorem, the original formulation of the DFT can only be applied to compute the ground state of the system and its properties. Nonetheless, in 1984 Erich Runge and Eberhard K. U. Gross were able to extend the Hohenberg-Kohn theorems of DFT to problems involving TD external potentials[238, 239] by proposing two new theorems: (i) the one-to-one correspondence and (ii) the variational principle, which provides a stationary (but not

a minimum) condition. The first one of them, which is only valid for external potentials, densities, and currents that can be expanded on a Taylor's series evaluated around the initial time, states that:

"There is a one-to-one correspondence via a TD current density $\boldsymbol{j}(\boldsymbol{r},t)$ between the v-representable TD density $\rho(\boldsymbol{r},t)$ and the TD external potential $v^{ext}(\boldsymbol{r},t)$ up to an additive time function $\phi(t)$."

$$\rho(\boldsymbol{r},t) \overset{\boldsymbol{j}(\boldsymbol{r},t)}{\Longleftrightarrow} v^{ext}(\boldsymbol{r},t) + \phi(t) \tag{2.89}$$

Hence, any operator is a functional of the TD density itself, since the expectation value of any operator can be evaluated by the TD density. Furthermore, the expectation values would, in principle, also depend on the initial state wavefunction $\Psi(\boldsymbol{r},t_0) = \Psi_0(\boldsymbol{r})$. However, as proven by the first Hohenberg–Kohn theorem, there is a bijective mapping between the initial density and the wavefunction. Hence, the operators can be expanded exclusively in terms of the TD density.[240]

The second of the aforementioned theorems was under debate for a long time since, even though it allowed the definition of a TD-DFT functional, it also introduced formal problems like the break of causality between the TD perturbation and the TD response. It states that:

"The action integral

$$\mathcal{A} = \int_{t_0}^{t_f} dt \int d\boldsymbol{r} \Psi^*(\boldsymbol{r},t) \left(i\frac{\partial}{\partial t} - \mathcal{H}(t) \right) \Psi(\boldsymbol{r},t) \tag{2.90}$$

can be represented as a functional of the density $\mathcal{A}[\rho(\boldsymbol{r},t)]$, which presents a stationary point at the exact density of the system that can be computed from the Euler equation:"

$$\frac{\partial \mathcal{A}}{\partial \rho(\boldsymbol{r},t)} = 0 \tag{2.91}$$

The total action can be written as:

$$\mathcal{A}[\rho(\boldsymbol{r},t)] = \mathcal{B}[\rho(\boldsymbol{r},t)] - \int_{t_0}^{t_f} dt \int d\boldsymbol{r} v^{ext}(\boldsymbol{r},t)\rho(\boldsymbol{r},t) \tag{2.92}$$

where $\mathcal{B}[\rho(\boldsymbol{r},t)]$ is a universal functional of the density, meaning that the same dependence on $\rho(\boldsymbol{r},t)$ holds for all external potentials $v^{ext}(\boldsymbol{r},t)$. Hence, in the case of the true density of the system, the (unconstrained) stationary condition

upon $\mathcal{A}$ implies that its functional derivative with respect to the density $\rho(\boldsymbol{r},t)$ vanishes, as given by Equation 2.91.

Similarly to the KS methodology for DFT discussed in Section 2.1.6, one can construct a TD non-interacting system with the same time evolution as the real interacting TD system, with a TD electron density given by:

$$\rho(\boldsymbol{r},t) = \sum_{p=1}^{K} f_p |\psi_p^{KS}(\boldsymbol{r},t)|^2 \tag{2.93}$$

in which f_p represents the orbital occupation and $\{\psi_p^{KS}(\boldsymbol{r},t)\}$, the set of K TD KS spin orbitals. Under this assumption, the universal functional $\mathcal{B}[\rho(\boldsymbol{r},t)]$ can be described by:

$$\begin{aligned}\mathcal{B}[\rho(\boldsymbol{r},t)] = &\sum_{i=1}^{N} f_p \int_{t_0}^{t_f} dt \int d\boldsymbol{r} \psi_i^*(\boldsymbol{r},t) \left(i\frac{\partial}{\partial t} - \frac{1}{2}\nabla_i^2 \right) \psi_i(\boldsymbol{r},t) \\ &- \frac{1}{2}\int_{t_0}^{t_f} dt \int\int d\boldsymbol{r}_1 d\boldsymbol{r}_2 \frac{\rho(\boldsymbol{r}_1,t)\rho(\boldsymbol{r}_2,t)}{|\boldsymbol{r}_1 - \boldsymbol{r}_2|} - \mathcal{A}_{xc}[\rho(\boldsymbol{r},t)]\end{aligned} \tag{2.94}$$

where the (unknown) XC functional $\mathcal{A}_{xc}[\rho(\boldsymbol{r},t)]$ has been defined. If $\mathcal{A}[\rho(\boldsymbol{r},t)]$ is minimized under the constraint in Equation 2.93, it is possible to demonstrate that the TD KS spin orbitals follow the Schrödinger-like set of one-electron equations commonly known as TD KS equations:

$$\hat{h}^{KS}(t)\psi_p^{KS}(\boldsymbol{r},t) = i\frac{\partial}{\partial t}\psi_p^{KS}(\boldsymbol{r},t) \quad \forall \{\psi_p^{KS}(\boldsymbol{r},t)\} \tag{2.95}$$

where the TD KS Hamiltonian $\hat{h}^{KS}(t)$, which is a functional of the TD density, is defined as:

$$\hat{h}^{KS}(\boldsymbol{r},t) = \hat{t}_r(\boldsymbol{r},t) + v^{ext}(\boldsymbol{r},t) + v_H[\rho(\boldsymbol{r}_1;t)] + v_{xc}[\rho(\boldsymbol{r},t)] \tag{2.96}$$

It is formed by addition of the non-interacting kinetic energy $\hat{t}_r(\boldsymbol{r},t)$, the TD external potential $v^{ext}(\boldsymbol{r},t)$, the TD Hartree potential $v_H[\rho(\boldsymbol{r}_1;t)]$, and the TD XC potential $v_{xc}[\rho(\boldsymbol{r},t)]$ with:

$$\hat{t}_r(\boldsymbol{r},t) = -\frac{1}{2}\nabla^2 \tag{2.97}$$

and:

$$v_H[\rho_t(\boldsymbol{r};t)] = \int d\boldsymbol{r}' \frac{\rho_t(\boldsymbol{r}';t)}{|\boldsymbol{r} - \boldsymbol{r}'|} \tag{2.98}$$

This TD Hartree potential exclusively depends on the instantaneous density $\rho_t(\boldsymbol{r};t)$ at time t. However, the TD XC potential depends on all the density history from t_0 to the current time. Equivalently to the KS DFT formulation in Equation 2.61, it can be described as the functional derivative of the XC action functional:

$$\frac{\delta \mathcal{A}_{xc}[\rho(\boldsymbol{r},t)]}{\delta \rho(\boldsymbol{r},t)} = v_{xc}[\rho(\boldsymbol{r},t)] \tag{2.99}$$

Accounting for the time-memory effect, or even trying to approximate it, presents a challenging task. Therefore, the most common and simplest description possible of the TD XC potential, called the adiabatic approximation, consists on neglecting the time-memory dependency. Within this approximation, the TD XC potential depends exclusively on the instantaneous density $\rho_t(\boldsymbol{r};t)$ and so the potential changes instantaneously following a change in the electron density. Hence, supposing that the XC functional of DFT at t_0 can properly describe the XC effects at all times, the TD XC potential (under the adiabatic approximation) can be defined as:

$$v_{xc}[\rho(\boldsymbol{r},t)] = \frac{\delta \mathcal{A}_{xc}[\rho(\boldsymbol{r},t)]}{\delta \rho(\boldsymbol{r},t)} \approx \frac{\delta E_{xc}[\rho_t(\boldsymbol{r};t)]}{\delta[\rho_t(\boldsymbol{r};t)]} = v_{xc}[\rho_t(\boldsymbol{r};t)] \tag{2.100}$$

This approximation is justified if the external TD is very slow and the system begins in a ground state, which is not usually the case. Nevertheless, and even though why this happens it still an open question, the adiabatic approximation works surprisingly good for most cases, and for those cases where it does not, such as the excited states of multiple-excitation character, the failure clearly arises from the lack of memory.[241]

Linear-Response Time Dependent Density Functional Theory

In cases for which the external perturbation is weak, the TD-DFT equations can benefit from the perturbative approach in the frequency domain. Conveniently, it is possible to express the Hamiltonian operator $\hat{\mathcal{H}}$ as the sum of an unperturbed Hamiltonian $\hat{\mathcal{H}}_0$ and a first-order TD perturbation term of the external potential $\delta\hat{v}^{ext}$:

$$\hat{\mathcal{H}} = \hat{\mathcal{H}}_0 + \delta\hat{v}^{ext} \tag{2.101}$$

where $\hat{\mathcal{H}}_0$ can be defined in terms of the kinetic energy operator $\hat{T}$, the electronic repulsion operator $\hat{V}_{ee}$ and the unperturbed external potential operator ${}^{(0)}\hat{v}^{ext}$:

$$\hat{\mathcal{H}}_0 = \hat{T} + \hat{V}_{ee} + {}^{(0)}\hat{v}^{ext} \tag{2.102}$$

Furthermore within the interaction picture, the wavefunction can be expressed as:

$$\Psi_I(\boldsymbol{r},t) = e^{i\hat{\mathcal{H}}_0 t}\Psi(\boldsymbol{r},t) \tag{2.103}$$

and any operator $\hat{\mathcal{O}}$, as:

$$\hat{\mathcal{O}}_I = e^{i\hat{\mathcal{H}}_0 t}\hat{\mathcal{O}}e^{-i\hat{\mathcal{H}}_0 t} \tag{2.104}$$

By substituting Equation 2.101 in the TDSE in Equation 2.1, in atomic units, it can be demonstrated that:

$$i\frac{\partial_I}{\partial t}\Psi_I(\boldsymbol{r},t) = \delta v_I^{ext}(\boldsymbol{r},t)\Psi_I(\boldsymbol{r},t) \tag{2.105}$$

Hence, the wavefunction in the interacting picture is given by:

$$\Psi_I(\boldsymbol{r},t) = \Psi_0(\boldsymbol{r}) - \int_0^t dt'\delta v_I^{ext}(\boldsymbol{r},t')\Psi_0(\boldsymbol{r}) + \hat{\mathcal{O}}(\delta v_I^{ext})^2 \tag{2.106}$$

where the first term correspond to the unperturbed (ground state) wavefunction; the second one, to the linear response of the wavefunction; and the third one, to higher-orders of the response. When any TD operator $\hat{\mathcal{O}}_I$ is applied to this $\Psi_I(\boldsymbol{r},t)$ wavefunction, the expectation value can be obtained by adding to the expectation value of the unperturbed system $<\hat{\mathcal{O}}>_0$ a first-order response term $\delta<\hat{\mathcal{O}}(t)>$ as defined in Equation 2.108, thus achieving time separation:

$$\begin{aligned} <\hat{\mathcal{O}}(t)> &= \int d\boldsymbol{r}\Psi_I^*(\boldsymbol{r},t)\hat{\mathcal{O}}_I\Psi_I(\boldsymbol{r},t) \\ &=<\hat{\mathcal{O}}>_0 -i\int_0^t dt'\int d\boldsymbol{r}\Psi_0^*(\boldsymbol{r})\left[\hat{\mathcal{O}}_I(t),\delta\hat{v}_I^{ext}(\boldsymbol{r},t')\right]\Psi_0(\boldsymbol{r}) \end{aligned} \tag{2.107}$$

$$\delta<\hat{\mathcal{O}}(t)>= -i\int_0^t dt'\int d\boldsymbol{r}\Psi_0^*(\boldsymbol{r})\left[\hat{\mathcal{O}}_I(t),\delta\hat{v}_I^{ext}(\boldsymbol{r},t')\right]\Psi_0(\boldsymbol{r}) \tag{2.108}$$

Hence, in the case of the density operator $\hat{\rho}(\boldsymbol{r})$, that is defined in the Schrödinger picture as:

$$\hat{\rho}(\boldsymbol{r}) = \sum_{i=1}^{N}\delta(\boldsymbol{r}-\boldsymbol{r}_i) \tag{2.109}$$

but can be easily transformed into the interaction picture via Equation 2.104, and considering that any multiplicative one-body operator such as $\delta\hat{v}_I^{ext}(\boldsymbol{r},t)$ can be written as:

$$\delta\hat{v}_I^{ext}(\boldsymbol{r},t) = \int d\boldsymbol{r}\delta v_I^{ext}(\boldsymbol{r},t)\hat{\rho}_I(\boldsymbol{r}) \tag{2.110}$$

then, introducing Equation 2.109 into Equation 2.108, the first-order response of the density is given by:

$$\delta\rho(\boldsymbol{r},t) = \int_0^\infty dt' \int d\boldsymbol{r}' \mathcal{X}(\boldsymbol{r},t,\boldsymbol{r}',t)\delta v^{ext}(\boldsymbol{r}',t') \tag{2.111}$$

with:

$$\begin{aligned}\mathcal{X}(\boldsymbol{r},t,\boldsymbol{r}',t) &= \frac{\delta\rho(\boldsymbol{r},t)}{\delta v^{ext}(\boldsymbol{r}',t')} \\ &= -i\Theta(t-t') \int d\boldsymbol{r}\Psi_0^*(\boldsymbol{r}) \left[\hat{\rho}_I(\boldsymbol{r},t), \hat{\rho}_I(\boldsymbol{r},t)\right] \Psi_0(\boldsymbol{r})\end{aligned} \tag{2.112}$$

where $\Theta(t-t')$ represents a step function that only activates when $t > t'$. This term appears because of the change of the integration limits when going from Equation 2.108 to Equation 2.111. $\mathcal{X}$ is known as the density-density response function or *susceptibility* and it is an essential object within the Linear Response (LR) approach since it allows describing the density in terms of the LR. The physical meaning behind this function, as represented schematically in Figure 2.2, is the description of the temporal behaviour of the density, how it responds, at a certain point of space $(\boldsymbol{r},t)$ upon poking at $(\boldsymbol{r}',t')$.

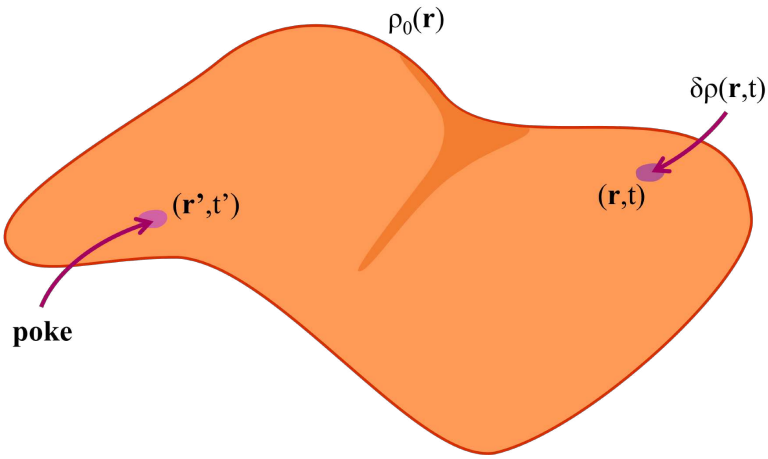


Figure 2.2: Graphical representation of the effect of the response function.

In order to obtain the energies (or the frequencies w), it is necessary to understand the dependencies between the two time variables t and t'. This can be achieved by means of the Fourier transformation:

$$t - t' \rightarrow w \tag{2.113}$$

which leads to:

$$\begin{aligned}\mathcal{X}(\boldsymbol{r},\boldsymbol{r}',w) &= \int d(t-t')\mathcal{X}(\boldsymbol{r},t,\boldsymbol{r}',t)e^{iw(t-t')} \\ &= \lim_{\eta\to 0^+}\sum_{J=1}^{\infty}\left(\frac{\int d\boldsymbol{r}\Psi_0^*(\boldsymbol{r})\hat{\rho}(\boldsymbol{r})\Psi_J(\boldsymbol{r})\int d\boldsymbol{r}\Psi_J^*(\boldsymbol{r})\hat{\rho}(\boldsymbol{r}')\Psi_0(\boldsymbol{r})}{w-\Xi_J+i\eta}\right. \\ &\quad \left.-\frac{\int d\boldsymbol{r}\Psi_0^*(\boldsymbol{r})\hat{\rho}(\boldsymbol{r})\Psi_J(\boldsymbol{r})\int d\boldsymbol{r}\Psi_J^*(\boldsymbol{r})\hat{\rho}(\boldsymbol{r}')\Psi_0(\boldsymbol{r})}{w-\Xi_J-i\eta}\right)\end{aligned} \tag{2.114}$$

where $\Psi_0(\boldsymbol{r})$ and $\Psi_J(\boldsymbol{r})$ represent the many body ground and the unperturbed excited state wavefunctions, respectively, and Ξ_J, the excitation energy of the state J given by the difference between the energies E of the states:

$$\Xi_J = E_J - E_0 \tag{2.115}$$

Following a similar procedure to the KS formalism, it is possible to describe a first-order response for the non-interacting reference state:

$$\delta\rho(\boldsymbol{r},t) = \int_0^{\infty} dt' \int d\boldsymbol{r}'\mathcal{X}_r(\boldsymbol{r},t,\boldsymbol{r}',t)\delta v_r^{ext}(\boldsymbol{r}',t') \tag{2.116}$$

with the response function:

$$\mathcal{X}_r(\boldsymbol{r},t,\boldsymbol{r}',t) = \frac{\delta\rho(\boldsymbol{r},t)}{\delta v_r^{ext}(\boldsymbol{r}',t')} \tag{2.117}$$

which, after performing a Fourier transformation, can be expressed in terms of the set of KS spin orbitals $\{\psi_p^{KS}(\boldsymbol{r},t)\}$ and their respective occupations f_p and KS eigenvalues ϵ_p:

$$\mathcal{X}_r(\boldsymbol{r},\boldsymbol{r}',w) = \lim_{\eta\to 0^+}\sum_{pq}^{K}(f_q-f_p)\frac{\psi_p(\boldsymbol{r})\psi_p^*(\boldsymbol{r})\psi_q(\boldsymbol{r})\psi_q^*(\boldsymbol{r})}{w-(\epsilon_q-\epsilon_p)+i\eta} \tag{2.118}$$

Hence, when the frequency is equal to the excitation energy ($w = \Xi_J$) or, in the case of the non-interacting state, when the frequency is equal to the difference of eigenvalues ($w = \epsilon_q - \epsilon_p$), the response function presents a pole ($\mathcal{X} = \pm\infty$).[240] Therefore, once the non-interacting response function and the exact TD XC kernel given by Equation 2.119 are known, it is possible to obtain the excitation energies of the interacting system.

$$f_{\hat{\mathcal{H}}_{xc}}(\boldsymbol{r},\boldsymbol{r}',w) = \frac{\delta v_H(\boldsymbol{r},w)}{\delta\rho(\boldsymbol{r},w)} + \frac{\delta v_{xc}(\boldsymbol{r},w)}{\delta\rho(\boldsymbol{r},w)} \tag{2.119}$$

It can be done by substituting the external potential of Equation 2.111 in Equation 2.116 and, after Fourier transforming, it is possible to achieve the Dyson-type equation for LR TD-DFT:

$$\left(\hat{\mathbb{1}} - \mathcal{X}_r(\boldsymbol{r},\boldsymbol{r}',w) f_{\hat{\mathcal{H}}_{xc}}(\boldsymbol{r},\boldsymbol{r}',w)\right) \mathcal{X}(\boldsymbol{r},\boldsymbol{r}',w) = w\mathcal{X}_r(\boldsymbol{r},\boldsymbol{r}',w) \tag{2.120}$$

In order to solve the indetermination that arises on this equation when $w = \Xi_J$ (the left-hand side tends to infinite while the right-hand side is finite) the first term in the left-hand side must be 0, that can be easily imposed by solving the eigenvalue problem:

$$\left(\hat{\mathbb{1}} - \mathcal{X}_r(\boldsymbol{r},\boldsymbol{r}',w) f_{\hat{\mathcal{H}}_{xc}}(\boldsymbol{r},\boldsymbol{r}',w)\right) X(\boldsymbol{r},\boldsymbol{r}',w) = wX(\boldsymbol{r},\boldsymbol{r}',w) \tag{2.121}$$

which, by expanding it in the KS spin orbital basis, leads to the usual Casida[242] form:

$$\begin{bmatrix} A(\boldsymbol{r},\boldsymbol{r}',w) & B(\boldsymbol{r},\boldsymbol{r}',w) \\ -B^*(\boldsymbol{r},\boldsymbol{r}',w) & -A(\boldsymbol{r},\boldsymbol{r}',w) \end{bmatrix} \begin{bmatrix} X(\boldsymbol{r},\boldsymbol{r}',w) \\ Y(\boldsymbol{r},\boldsymbol{r}',w) \end{bmatrix} = w \begin{bmatrix} X(\boldsymbol{r},\boldsymbol{r}',w) \\ Y(\boldsymbol{r},\boldsymbol{r}',w) \end{bmatrix} \tag{2.122}$$

with the matrices $A(\boldsymbol{r},\boldsymbol{r}',w)$ and $B(\boldsymbol{r},\boldsymbol{r}',w)$ given by:

$$A_{ia,jb}(\boldsymbol{r},\boldsymbol{r}',w) = \delta_{ij}\delta_{ab}(\epsilon_a - \epsilon_i) + (ia|\hat{f}_{\hat{\mathcal{H}}xc}(\boldsymbol{r},\boldsymbol{r}',w)|bj) \tag{2.123}$$

$$B_{ia,jb}(\boldsymbol{r},\boldsymbol{r}',w) = (ia|\hat{f}_{\hat{\mathcal{H}}xc}(\boldsymbol{r},\boldsymbol{r}',w)|jb) \tag{2.124}$$

where the two-electron integrals are expressed in the Mulliken notation:

$$\begin{aligned}(pq|\hat{f}_{\hat{\mathcal{H}}xc}(\boldsymbol{r},\boldsymbol{r}',w)|rs) \\ = \int d\boldsymbol{r} \int d\boldsymbol{r}' \psi_p^*(\boldsymbol{r})\psi_q(\boldsymbol{r}) \left(\frac{1}{|\boldsymbol{r}-\boldsymbol{r}'|} + f_{xc}(\boldsymbol{r},\boldsymbol{r}',w)\right) \psi_r^*(\boldsymbol{r}')\psi_s(\boldsymbol{r}')\end{aligned} \tag{2.125}$$

From the sole knowledge of the exact $\hat{v}_{xc}$ and an initial KS density, a true excitation energy of the system can thus be determined from this LR TD-DFT equation, which is formally exact to determine the first-order response of the TD density:

$$\delta\rho(\boldsymbol{r},\Xi_J) = \sum_{ia} (X_{ia}(\Xi_J)\phi_a^*(\boldsymbol{r})\phi_i(\boldsymbol{r}) + Y_{ia}(\Xi_J)\phi_a(\boldsymbol{r})\phi_i^*(\boldsymbol{r})) \tag{2.126}$$

Nonetheless, Equation 2.122 needs to be solved approximately. The most common approximation is the aforementioned adiabatic approximation which, within the LR framework, leads to a linear eigenvalue problem with a number of solutions equal to the dimensions of the $A(\boldsymbol{r},\boldsymbol{r}',w)$ and $B(\boldsymbol{r},\boldsymbol{r}',w)$ matrices

thus losing the extra solutions corresponding to double and higher electronic excitations since these matrices present the dimensions of one-electron excitations. Furthermore, the quality of the excitation energies is strongly influenced by the fact that the XC effects do not take into account the character of the excited state. Another approximation is the so-called Tamm–Dancoff Approximation (TDA),[243] which consist on the decoupling of the excitations from de-excitations by means of neglecting the terms coming from the B matrix in Equation 2.122 ($B(\boldsymbol{r}, \boldsymbol{r}', w) = 0$).[240] This approximation is less sensitive to instabilities of the ground state and only affects the excited state calculation, improving the description of the energies at the expense of other properties. It results in a solution of the form of a Hermitian eigenvalue problem:

$$A(\boldsymbol{r}, \boldsymbol{r}', w)X(\boldsymbol{r}, \boldsymbol{r}', w) = wX(\boldsymbol{r}, \boldsymbol{r}', w) \tag{2.127}$$

2.1.8 Typical Quantum Mechanical Calculations

Independently of the required computational QM modelling method, there exist several types of calculations that are common for most of them. Within this section, the difference between single-point calculations, geometry optimizations, normal mode frequencies and scanning of geometrical coordinates will be discussed, along with some important features of some of them. This discussion will be constrained within a BO static framework so that the time evolution of the nuclei, which is addressed in Section 2.2, will be neglected.

Single-Point Calculation

For a given arrangement of the position of the M nuclei, a Single Point (SP) calculation provides the electronic energy of the system under study without modifying its geometry. It is important to stress that SPs are the basis of many advanced QM molecular modelling techniques and they do not typically provide any information about the slope and curvature of the PES since they can be calculated for any nuclear configuration, regardless of whether it corresponds to a critical point or not. Nonetheless, it is also possible to study the curvature of the system following the procedures explained in the Geometry Optimization and Frequency Calculation sections. This curvature would inform on whether the geometry is a stable specie (a minimum in the PES), a TS (mathematically, a first-order saddle point in the PES), or a higher energy structure. Moreover, SP calculations allow the study of potential reaction pathways as will be discussed in the Scan Coordinate section and the relative stability between different molecular structures. Furthermore, a SP does not necessarily correspond to the calculation of the ground state but also refers to

the calculation of vertical excitations from a specific geometry. Therefore, SP calculations also allow for the calculation of absorption and emission spectra. Typically, this type of calculation is performed on top of an optimized geometry. However, it can be useful to calculate this property for a set of sampled geometries, as will be discussed in Section 2.2.7.

Geometry Optimization

Even thought SP calculations are the basis of the computational QM modelling methods, many chemical problems require the location of stationary points of the PES, points where the first derivative of the electronic energy with respect to the atomic coordinates is zero. In most cases, the desired stationary point is a minimum, which can be easily distinguished from a saddle point since all its second partial derivatives are positive. In general terms, all optimization algorithms follow the gradient $\boldsymbol{g}$, the first derivative of the electronic energy function with respect to all the coordinates thus pointing in the direction in which the function decreases the most, and, in some cases, they also take into account the matrix of second-order partial derivatives, the Hessian $\boldsymbol{H}$.

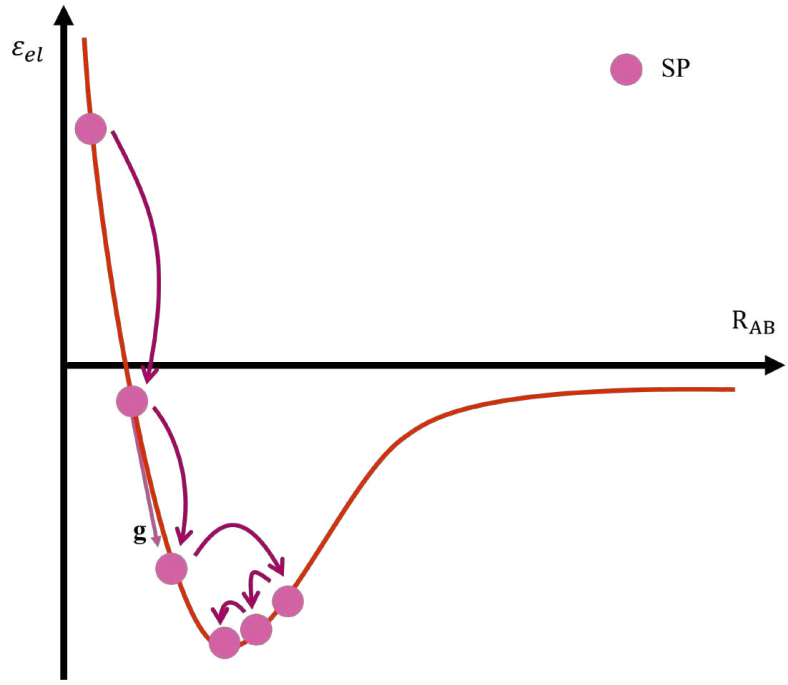


Figure 2.3: Graphical representation of a general optimization procedure.

As represented in Figure 2.3, for a given conformation of the system, both the SP, the gradient and, sometimes, the Hessian matrix of the PES are calculated. From this gradient (and Hessian), a new set of atomic coordinates is determined and a new SP and gradient are calculated. This process is repeated iteratively, until a saddle point is located.

The energy as a function of the atomic coordinates (and its derivatives) is calculated within a given precision during the optimization process. Therefore, the gradient can only be reduced to a certain value and, thus, any stationary point can not be located exactly. In practice, the optimization is considered to be converged if the gradient is reduced below a predefined threshold value and/or if the change of the function between two iterations becomes sufficiently small. These two criteria can be specially problematic in very flat regions, since convergence is very difficult to be reached as such stationary points are hard to locate.[213]

The most common optimization methods for finding minima can be divided into three types: (i) the Steepest Descent method, (ii) the Conjugate Gradient methods and (iii) Newton–Raphson methods.

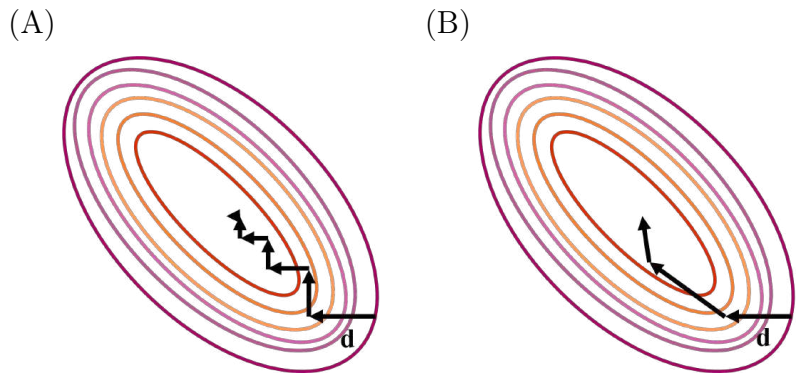


Figure 2.4: Graphical representation of (A) the Steepest Descent and (B) Conjugate Gradient methods.

The Steepest Descent method, firstly proposed by Augustin Cauchy in 1847,[244] consists of a series of function evaluations along a search direction $\boldsymbol{d}$ defined as the negative gradient:

$$\boldsymbol{d} = -\boldsymbol{g} \tag{2.128}$$

In this method, an approximate minimum may be determined by interpolation between the calculated points and, at this interpolated point, a new gradient is calculated for the next line search. This algorithm ensures that, given an

accurate enough line minimization, a lower function value will always be found. However, it presents two significant drawbacks. Firstly, two subsequent line searches are necessarily perpendicular to each other, as illustrated in Figure 2.4(A). Therefore, it neglects gradient components along the previous search direction, which is particularly problematic for surfaces having long narrow valleys. Secondly, the rate of convergence slows down as the minimum is approached. Hence, it is very difficult to reach the real minimum using this method, and it is mainly reserved for quick relaxations of poor starting points to obtain a better initial guess for a successive more advanced algorithm.

One of these more advanced algorithms is the Conjugate Gradient method, which improves the description and convergence of the search direction of each l step via "conjugation" to the previous one $(l-1)$, as graphically depicted in Figure 2.4(B):

$$\boldsymbol{d}_l = -\boldsymbol{g} + \gamma_l \boldsymbol{d}_{l-1} \tag{2.129}$$

where γ_l is just a function of the gradients at two consecutive search directions. There exist several formulations for its description, but the Polak–Ribiere[245] one in Equation 2.130 is usually preferred since it tends to restart the procedure, which is often required during the optimization.

$$\gamma_l = \frac{\boldsymbol{g}_l^T (\boldsymbol{g}_l - \boldsymbol{g}_{l-1})}{\boldsymbol{g}_{l-1}^T \boldsymbol{g}_{l-1}} \tag{2.130}$$

Finally, the Newton–Raphson method expands the true function to second-order around the current point $\boldsymbol{r}_0$:

$$f(\boldsymbol{r}) \approx f(\boldsymbol{r}_0) + \boldsymbol{g}^t(\boldsymbol{r} - \boldsymbol{r}_0) + \frac{1}{2}(\boldsymbol{r} - \boldsymbol{r}_0)^t \boldsymbol{H} (\boldsymbol{r} - \boldsymbol{r}_0) \tag{2.131}$$

Considering that, in a stationary point the gradient of the second-order approximation must be equal to 0, then the Newton–Raphson step is given by:

$$(\boldsymbol{r} - \boldsymbol{r}_0) = -\boldsymbol{H}^{-1} \boldsymbol{g} \tag{2.132}$$

which, in the coordinate system $(\boldsymbol{r}')$, where the Hessian is diagonal, becomes:

$$\Delta \boldsymbol{r}' = (\Delta \boldsymbol{r}'_1, \Delta \boldsymbol{r}'_2, ..., \Delta \boldsymbol{r}'_M) \tag{2.133}$$

with:

$$\Delta \boldsymbol{r}'_l = -\frac{f_l}{\lambda_l} \tag{2.134}$$

where f_l is the projection of the gradient along the Hessian eigenvector with eigenvalue λ_l. This method can be used iteratively for stepping towards a stationary point because the real function contains terms beyond second-order. By definition, in the vicinity of a minimum, all the Hessian eigenvalues are positive, and the step direction is opposite to that of the gradient. However, if one of the Hessian eigenvalues is negative, the step in that direction will be along the gradient component, and thus increase the function value. In this case, the optimization may end up at a stationary point with one negative Hessian eigenvalue, a first-order saddle point called TS in a chemical framework. In general, this method will converge on the "closest" stationary point, regardless of it curvature, whether it is a minimum, a saddle point or a maximum.

Newton–Raphson methods can be combined with extrapolation procedures such as the Geometry Direct Inversion in the Iterative Subspace (GDIIS)[246] in which, the Newton–Raphson step is taken from an interpolated point with a corresponding interpolated gradient based on the previously calculated points on the surface instead of from the last geometry. This approach attempts to find a low-gradient point within the already searched subspace. However, this might be counterproductive since this procedure tends to pull the structure back to the flat energy region whenever the optimization moves across a flat part of the surface before entering the region of the local minimum. An advanced version of this algorithm, the Berny algorithm using the Energy-Represented GEDIIS[247] based on an earlier program written by H. Bernhard Schlegel[248] in 1982, is precisely the one used in the software Gaussian, the QM software used along this thesis.

Frequency Calculation

As mentioned before, not all the geometry optimization algorithms ensure that the final geometry is a minimum of the PES, but a stationary point. Thus, a frequency calculation is a crucial final step of any geometry optimization procedure, since it informs of the curvature of the system for a given nuclear arrangement.

In the vicinity of any minima, the PES can be approximated by means of the Harmonic Oscillator (HO) model. For simplicity, let's consider a A–B diatomic molecule. As illustrated in Figure 2.5, the PES would present a minimum at the equilibrium distance R_e, while at very small or large values of the interatomic distance ($r_{AB} = |\boldsymbol{r}_A - \boldsymbol{r}_B|$) the surface would present a vertical asymptote to $+\infty$ or a horizontal asymptote to 0, respectively.

Nonetheless, and independently of the shape of this PES, it is possible to express its shape as a Taylor expansion of the interatomic distance function

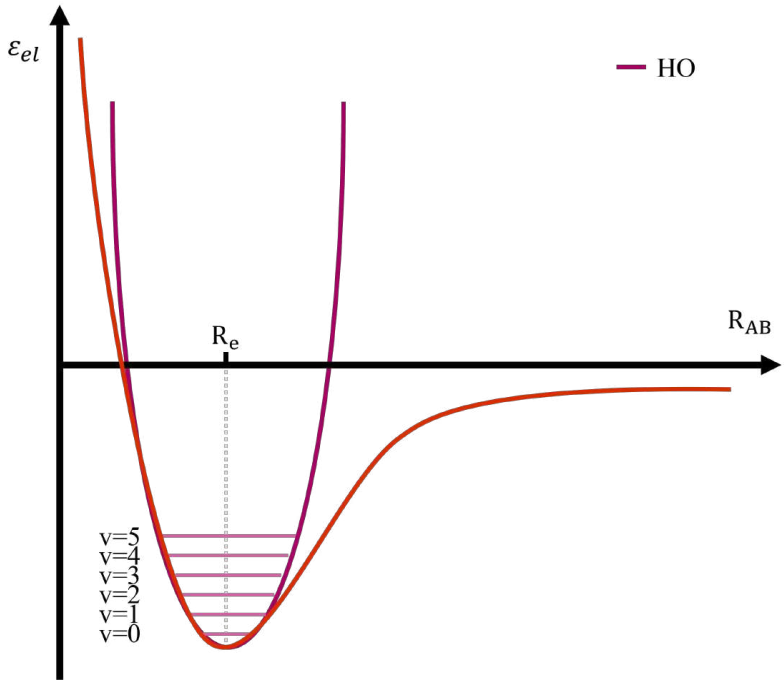


Figure 2.5: Graphical representation of the HO model.

around R_e:

$$E(r_{AB}) = E(R_e) + \left(\frac{dE(r_{AB})}{dr_{AB}}\right)_{R_e}(r_{AB} - R_e) + \frac{1}{2!}\left(\frac{d^2E(r_{AB})}{dr_{AB}^2}\right)_{R_e}(r_{AB} - R_e)^2 + \frac{1}{3!}\left(\frac{d^3E(r_{AB})}{dr_{AB}^3}\right)_{R_e}(r_{AB} - R_e)^3 + \ \ldots \tag{2.135}$$

Taking into account that the PES presents a minimum at R_e, then the first derivative of the second term of Equation 2.135 must be 0. Moreover, it is possible to truncate this equation at second-order, since higher-order derivatives become ever smaller. Hence:

$$E(r_{AB}) = E(R_e) + \frac{1}{2!}\left(\frac{d^2E(r_{AB})}{dr_{AB}^2}\right)_{R_e}(r_{AB} - R_e)^2 \tag{2.136}$$

where $E(R_e)$ is just a system dependent constant and the second derivative of the second term corresponds to the definition of k_e, the force constant of

the oscillator, that is related to the vibrational frequency ν_{AB} between the two atoms:

$$\nu_{AB} = \frac{1}{2\pi}\sqrt{\frac{k_e}{\mu}} \tag{2.137}$$

with the reduced mass μ given by:

$$\mu = \frac{m_a m_B}{m_A + m_B} \tag{2.138}$$

In the case of a polyatomic molecule, the k_{nm} force constants of each of the nm number of vibrational normal modes ($3M-5$ or $3M-6$ for linear and non-linear molecules, respectively) correspond to the eigenvalues of the Hessian matrix. Furthermore, the eigenvectors describe the relative movement of the M nuclei for each normal mode. Hence, according to Equation 2.137, if all the vibrational frequencies are real, then the geometry of the molecule would correspond to a minimum of the PES. On the flip side, whenever a vibrational frequency is imaginary, the PES would present a maxima in as many directions as imaginary frequencies there are.

In addition to the curvature information of the PES for a given nuclear arrangement, frequency calculations also allow accessing several thermodynamic properties via the molecular partition function:[209]

$$q = q^E q^V q^R q^T \tag{2.139}$$

where the electronic (q^E), vibrational (q^V), rotational (q^R) and translational (q^T) partition function contributions can be expressed in terms of the Boltzmann constant k_B; the temperature T; the Plank constant h; the speed of light in vacuum c; the degeneracy of the electronic ground state g_0; the number of indistinguishable orientations of the molecule σ, called the symmetry number; the rotational constants A, B and C; the volume V and the mass of the molecule m:

$$q^E = g_o \tag{2.140}$$

$$q^V = \frac{1}{1 - e^{-\frac{\Theta_V}{T}}} \quad \text{with} \quad \Theta_V = \frac{h\nu}{k_B} \tag{2.141}$$

$$\begin{aligned} \text{Linear molecules} \qquad & q^R = \frac{T}{\Theta_R} \quad \text{with} \quad \Theta_R = \frac{hcB}{k_B} \\ \text{Non-linear molecules} \qquad & q^R = \frac{1}{\sigma}\left(\frac{k_B T}{hc}\right)^{\frac{3}{2}}\left(\frac{\pi}{ABC}\right)^{\frac{1}{2}} \end{aligned} \tag{2.142}$$

$$q^T = \frac{V}{\Lambda^3} \quad \text{with} \quad \Lambda = \frac{h}{\sqrt{2\pi m k_B T}} \tag{2.143}$$

In the case of independent molecules, the molecular partition function can be used to calculate the canonical partition function:

$$\begin{aligned} &\text{Indistinguishable molecules} \qquad Q = \frac{q^M}{M!} \\ &\text{Distinguishable molecules} \qquad Q = q^M \end{aligned} \tag{2.144}$$

which, at the same time, is related to the internal energy ($U - U(T = 0)$) relative to its value at $T = 0$ ($U(T = 0)$), to the entropy S and to the pressure P:

$$U - U(T = 0) = -\left(\frac{\partial lnQ}{\partial \beta}\right)_V \tag{2.145}$$

$$S = \frac{U - U(T = 0)}{T} + k_B lnQ \tag{2.146}$$

$$P = k_B T \left(\frac{\partial lnQ}{\partial V}\right)_T \tag{2.147}$$

where:

$$\beta = \frac{1}{k_B T} \tag{2.148}$$

These equations can be used to derive those of the Helmholtz energy ($A - A(T = 0)$), the entropy ($H - H(T = 0)$) and the Gibbs Free energy ($G - G(T = 0)$) by means of the standard thermodynamics relationships:

$$A = U - TS \quad \longrightarrow \quad A - A(T = 0) = -k_B T lnQ \tag{2.149}$$

$$H = U + pV \quad \longrightarrow \quad H - H(T = 0) = -\left(\frac{\partial lnQ}{\partial \beta}\right)_V + k_B TV \left(\frac{\partial lnQ}{\partial V}\right)_T \tag{2.150}$$

$$G = A + pV \quad \longrightarrow \quad G - G(T = 0) = -k_B T lnQ + k_B TV \left(\frac{\partial lnQ}{\partial V}\right)_T \tag{2.151}$$

Scan Coordinate

Sometimes, it is interesting to investigate how certain properties of the system are affected by specific changes in the position of the nuclei. For example, what happens when the distance between two atoms is gradually increased, leading to a dissociation process or when a torsion around a single bond takes place. In these cases, it is possible to perform a so-called rigid scan, used in Chapters III, IV and VI of this thesis. It consists on the generation of a series of distorted geometries along a specific reaction coordinate in order to perform SP calculations on top of each of them to obtain the desired property.

Nonetheless, it is also possible to use these geometries as starting points of geometry optimizations to obtain a relaxed scan. In this case, it is required to apply some constrains during the optimization procedure which can be done by either using (i) a penalty function, in which the reaction coordinate is constrained to be near a target value by applying an external potential; (ii) Lagrange undetermined multipliers, so the optimization is then performed on a Lagrange function by requiring that the gradient components with respect to the constrained variables of the system are equal to zero, or (iii) a projection method, in which a general non-linear constraint condition can be considered to be fulfilled by projecting out the first-order linear Taylor approximation to the function.[213]

Charge Transfer Analysis

Additional to these types of calculation, it is also possible to study specific electronic properties of the system, like the CT number. This property allows for the quantification of the number of electrons that are transferred from one fragment A to a different fragment B of the system under study upon an electronic transition.[249] Upon fragment identification, it is possible to construct different excited states by creating an electron–hole pair whose effective wavefunction can be extracted from the one-electron transition density matrix between the ground and excited states:

$$T^{OI}(\boldsymbol{r}_h, \boldsymbol{r}_e) = N \int \cdots \int \Psi_0(\boldsymbol{r}_h, \boldsymbol{r}_2, \cdots, \boldsymbol{r}_N)\Psi_I(\boldsymbol{r}_e, \boldsymbol{r}_2, \cdots, \boldsymbol{r}_N) d\boldsymbol{r}_2 \cdots d\boldsymbol{r}_N \tag{2.152}$$

where Ψ_0 and Ψ_I are, respectively, the ground and excited state wavefunctions and $\boldsymbol{r}_h$ and $\boldsymbol{r}_e$ coordinates represent the hole and electron that originate from the two electronic states. From this one-electron transition density matrix it is possible to decompose the excitation into different local and CT contributions by computing the CT numbers:

$$CT_{AB} = \int_A \int_B |T^{OI}(\boldsymbol{r}_h, \boldsymbol{r}_e)|^2 d\boldsymbol{r}_e d\boldsymbol{r}_h \tag{2.153}$$

where the hole and electron are restricted to fragments A and B, respectively. Practically speaking, this integral can be evaluated using a population analysis scheme like, for instance, the Löwdin orthogonalization.[250, 251]

$$CT_{AB} = \sum_{\mu \in A} \sum_{\nu \in B} \left(S^{1/2} \bar{T}^{OI} S^{1/2} \right)^2_{\mu\nu} \tag{2.154}$$

where S is the again the overlap matrix and $\bar{T}$ represents the one-electron transition density matrix in the basis of the atomic orbitals:

$$\bar{T}^{OI} = CT^{OI}C^T \tag{2.155}$$

2.2 Nuclear Motion Simulation

Some chemical reactions are driven by the energetic differences between the chemical species along the reaction coordinate. Nonetheless, the reactivity can not be explained exclusively by a thermodynamic control. In many cases, it is crucial to understand the kinetics of the reaction and therefore the nuclear motion of a system over time. A typical example of this, is the degradation of diamonds to graphite that, although it is thermodynamically favourable under standard conditions, it can take from millions to billions of years to happen.

In order to study these TD processes, the most accurate description of the nuclear motion is the TDSE in Equation 2.1, since it contains the first derivative of the wavefunction with respect to time and allows the calculation of the future wavefunction at any time, given an initial wavefunction at time t_0.[209] Even thought the electronic and nuclear contributions to the wavefunction are separable in the BO framework, as shown in Equation 2.17, and therefore the electronic part for a given nuclear configuration can be used to determine the future nuclear position, this approach is very computationally demanding even for systems of few atoms. In practice, it is possible to use deterministic simulations, given by the Molecular Dynamics (MD) method, to obtain a time evolution of the system by integrating Hamilton's equations of motion.[252] In these framework the BO approximation is implicit included and so the motion of the nuclei is inherently decoupled from that of the electrons.

2.2.1 The Hamilton's Equations of Motion

Hamilton's Equations of motion, as expressed in Equation 2.156, are the pair of equations that allow for the propagation of nuclear coordinates and momenta of the system to calculate its time evolution:

$$\frac{d\boldsymbol{p}_k}{dt} = -\frac{dU}{d\boldsymbol{R}_k} \qquad \frac{dT_k}{d\boldsymbol{p}_k} = \frac{d\boldsymbol{R}_k}{dt} \tag{2.156}$$

They can be easily derived from Newton's laws and the definition of the kinetic energy. Hence, starting from Newton's Second law:

$$\boldsymbol{F}_k = \frac{d\boldsymbol{p}_k}{dt} \tag{2.157}$$

where $\boldsymbol{F}_k$ and $\boldsymbol{p}_k$ represent the forces and the momentum for a given atom; and knowing that, for conservative systems, the relationship between the forces acting on each atom and the potential U is:

$$\boldsymbol{F}_k = -\frac{dU}{d\boldsymbol{R}_k} \tag{2.158}$$

then, the first Hamilton equation of Equation 2.156 is simply obtained by equalling Equations 2.157 and 2.158. In the case of the second Hamilton equation, it is obtained by deriving the expression of the kinetic energy with respect to the momentum:

$$\frac{dT_k}{d\boldsymbol{p}_k} = \frac{d}{d\boldsymbol{p}_k}\left(\frac{\boldsymbol{p}_k^2}{2m_k}\right) = \frac{2\boldsymbol{p}_k}{2m_k} = \frac{m_k\boldsymbol{v}_k}{m_k} = \boldsymbol{v}_k = \frac{d\boldsymbol{R}_k}{dt} \tag{2.159}$$

In order to solve these equations, three different ingredients are required: (i) an integrator to propagate particle positions and velocities in time, (ii) an interaction model to describe the potential energy of the system and (iii) a selected thermodynamic ensemble to determine the thermodynamic quantities that need to be controlled during the simulation.[252] Next three sections are dedicated to thoroughly describe each of these elements. Furthermore, more specific details of some of them like, for instance, the specific algorithms employed along this thesis will also be discussed.

2.2.2 The Integrator

The integrator is a key ingredient for the implementation of MD, since it allows propagating the positions and velocities of all the atoms of the system in time by numerically solving the Hamilton equations of motion in Equation 2.156. It must meet some requirements: (i) accuracy so that the resulting trajectory corresponds to a good approximation of the real trajectory; (ii) stability in the sense that total energy is conserved and that small perturbations of the system do not lead to instabilities and (iii) robustness, that allows the use of large time steps during the propagation.

There exist several types of integrators: (i) the expansion based methods, constructed by expanding the positions and velocities in a Taylor series; (ii) the operator splitting methods, based on the phase space description of a classical system and that additionally allow splitting the propagator into several time scales, and (iii) the multiple time step methods, that allow to decompose the time scales of the system depending on the speed of the motion of the different degrees of freedom (DOF).[252] All along this thesis, the MD simulations exclusively use the Velocity Verlet algorithm,[253] an expansion based integrator that is explained thereupon.

The Velocity Verlet Algorithm

The velocity Verlet algorithm allows obtaining the atomic coordinates $\boldsymbol{R}_k$ and the velocities $\boldsymbol{v}_k$ at a certain time from the positions and velocities of the previous step and from the acceleration of previous and current time steps:

$$\boldsymbol{R}_k(t+\Delta t) = \boldsymbol{R}_k(t) + \boldsymbol{v}_k(t)\Delta t + \frac{1}{2}\boldsymbol{a}_k(t)\Delta t^2 \tag{2.160}$$

$$\boldsymbol{v}_k(t+\Delta t) = \boldsymbol{v}_k(t) + \frac{1}{2}\left(\boldsymbol{a}_k(t) + \boldsymbol{a}_k(t+\Delta t)\right)\Delta t \tag{2.161}$$

where Δt represents the time step, the time between two consecutive steps of the trajectory. As commented before, the velocity Verlet algorithm is an expansion based method. Therefore, its equations are directly obtained by performing a Taylor expansion around $\boldsymbol{R}_k(t)$ and $\boldsymbol{v}_k(t)$ for coordinates and velocities, respectively.

In the case of the coordinates:

$$\boldsymbol{R}_k(t+\Delta t) = \boldsymbol{R}_k(t) + \Delta t\frac{d\boldsymbol{R}_k(t)}{dt} + \frac{\Delta t^2}{2!}\frac{d^2\boldsymbol{R}_k(t)}{dt^2} + \frac{\Delta t^3}{3!}\frac{d^3\boldsymbol{R}_k(t)}{dt^3} + ... \tag{2.162}$$

Hence, truncating up to second-order and identifying the derivatives of the first and second-order terms in Equation 2.162 with the velocity and the acceleration, respectively, the Equation 2.160 for the propagation of the coordinates is directly obtained. Doing the same for the velocities:

$$\boldsymbol{v}_k(t+\Delta t) = \boldsymbol{v}_k(t) + \Delta t\frac{d\boldsymbol{v}_k(t)}{dt} + \frac{\Delta t^2}{2!}\frac{d^2\boldsymbol{v}_k(t)}{dt^2} + \frac{\Delta t^3}{3!}\frac{d^3\boldsymbol{v}_k(t)}{dt^3} + ... \tag{2.163}$$

However, in this case, it is not possible to identify the second-order term of the expansion with a property of the system so an expression for it needs to be obtained, which can be achieved by doing another Taylor expansion around $\frac{d\boldsymbol{v}_k(t)}{dt}$:

$$\frac{d\boldsymbol{v}_k(t+\Delta t)}{dt} = \frac{d\boldsymbol{v}_k(t)}{dt} + \Delta t\frac{d^2\boldsymbol{v}_k(t)}{dt^2} + \frac{\Delta t^2}{2!}\frac{d^3\boldsymbol{v}_k(t)}{dt^3} + ... \tag{2.164}$$

Truncating Equations 2.163 and 2.164, solving the latter for the first-order term and substituting it in the former; then, it is obtained that:

$$\begin{aligned}
\boldsymbol{v}_k(t+\Delta t) &= \boldsymbol{v}_k(t) + \Delta t\frac{d\boldsymbol{v}_k(t)}{dt} + \frac{\Delta t}{2}\frac{d\boldsymbol{v}_k(t+\Delta t)}{dt} - \frac{\Delta t}{2}\frac{d\boldsymbol{v}_k(t)}{dt} \\
&= \boldsymbol{v}_k(t) + \frac{\Delta t}{2}\frac{d\boldsymbol{v}_k(t+\Delta t)}{dt} + \frac{\Delta t}{2}\frac{d\boldsymbol{v}_k(t)}{dt} \\
&= \boldsymbol{v}_k(t) + \frac{\Delta t}{2}\left(\frac{d\boldsymbol{v}_k(t+\Delta t)}{dt} + \frac{d\boldsymbol{v}_k(t)}{dt}\right)
\end{aligned} \tag{2.165}$$

In this equation, the terms in the parenthesis can be easily identified as the acceleration of the future and current steps, respectively; thus obtaining the equation for the propagation of the velocities of Equation 2.161.

As it is necessary the information of the acceleration $\boldsymbol{a}_k$ of the future step, the calculation of the velocity is performed in two steps. Initially, the first two terms of Equation 2.161 are calculated since they only require the already known information of the previous time step and then, after the new coordinates are obtained and the acceleration recomputed, the third term is added. Accelerations are calculated by means of Newton's second law:

$$\boldsymbol{a}_k = \frac{\boldsymbol{F}_k}{m_k} \tag{2.166}$$

where the forces acting on each atom $\boldsymbol{F}_k$ are determined from the potential energy of the system, which assuming it is conservative, only depends on the atomic positions of the atoms:

$$\boldsymbol{F}_k\left(\boldsymbol{R}_1, ..., \boldsymbol{R}_M\right) = -\frac{\partial U\left(\boldsymbol{R}_1, ..., \boldsymbol{R}_M\right)}{\partial \boldsymbol{R}_k} \tag{2.167}$$

At the beginning of the simulation, the coordinates are taken either from experimental results, as in the case of proteins, or previous theoretical calculations and the velocities, from a Maxwell-Boltzmann distribution given by Equation 2.168 at a certain temperature, supposing that the system is in thermal equilibrium.

$$P(\boldsymbol{v}_k) = 4\pi \left(\frac{m_k\beta}{2\pi}\right)^{\frac{3}{2}} \boldsymbol{v}_k^2 e^{-\frac{m_k \boldsymbol{v}_k^2 \beta}{2}} \tag{2.168}$$

2.2.3 Interaction Model

Apart from the integrator, Hamilton's equations of motion in Equation 2.156 require the description of the potential $U(\boldsymbol{R})$. This description can be given at different levels of theory, leading to two flavours of MD: (i) *Ab Initio* Molecular Dynamics (AIMD), in which the potential is obtained from a QM description of the electrons, and (ii) Force Field (FF) based MD, in which the potential is calculated from a sum of analytical and empirical functions.

Ab Initio Molecular Dynamics

The most accurate way for describing the potential on which the nuclei move during a MD simulation $U(\boldsymbol{R}) = (\boldsymbol{R}_1, \boldsymbol{R}_2, ..., \boldsymbol{R}_M)$, as it appears in Equation 2.156, is to solve the TISE in Equation 2.18 at a suitable QM level.[254] In

this strategy, that is referred to as AIMD, the Hamilton's equations of motion are solved for a given set of coordinates at a specific time $\boldsymbol{R}(t')$ and then, they are propagated numerically using, for instance, the Velocity Verlet algorithm presented in Section 2.2.2. At every time step t', the coordinates $\boldsymbol{R}(t')$ are used to compute the energy $U(\boldsymbol{R}(t'))$ and the gradient $\nabla U(\boldsymbol{R}(t'))$ at a QM level, which in turn is used to determine the position of the atoms at time $t' + \Delta t$ by introducing it in Equation 2.156.

Some typical examples of applications include the exploration of reaction pathways and the research of transition state structures,[255, 256] and the computation of thermodynamic properties at a high level of theory.[257, 258]

Force Fields

In the FF based MD methods, the potential energy of the system is approximated by a sum of analytical and empirical functions. The collection of these functions is called FF. Usually, the functions are divided in bonding and non-bonding terms[259] and each of them tries to describe a certain kind of atomic movement or pairwise interactions as represented in Figure 2.6.

$$U^{total} = U^{bonding} + U^{non-bonding} \tag{2.169}$$

Within bonding interactions, there are three main types: (i) bond stretching, (ii) bond angle bending and (iii) dihedral angle interactions. The first one, which represents the elongation and contraction of the interatomic distance

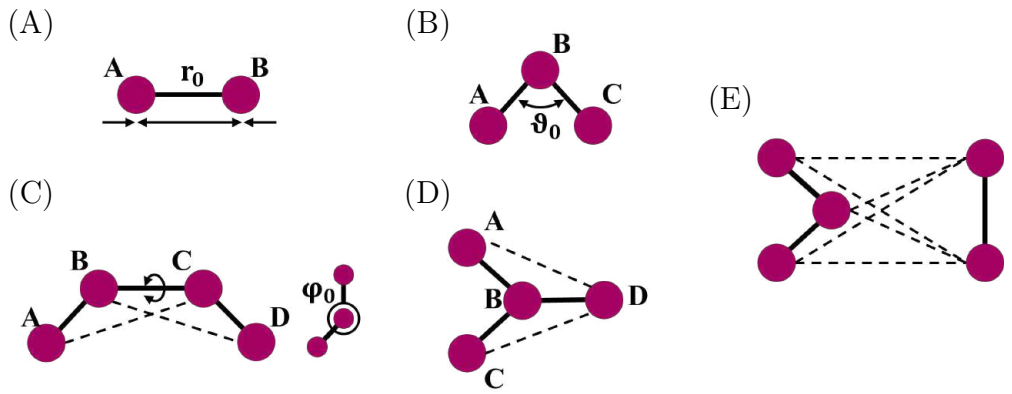


Figure 2.6: Graphic representation of the main kinds of interactions employed in FFs. (A - D) Represent the bonding terms (bond stretching, angle bending, proper dihedral and improper dihedral, respectively) and (E), the non-bonding terms.

between two covalently bonded atoms, is usually given by either a harmonic (in non-reactive FFs) or a Morse potential (in reactive FFs):

$$\begin{aligned} \text{Harmonic} \qquad & U_{AB}(R_{AB}) = \frac{1}{2}k_{AB}(R_{AB} - R_0)^2 \\ \text{Morse} \qquad & U_{AB}(R_{AB}) = \frac{1}{2}k_{AB}(1 - e^{a(R_{AB}-R_0)})^2 \end{aligned} \tag{2.170}$$

The second one describes the angle variations between three consecutively bonded atoms:

$$U_{ABC}(\vartheta_{ABC}) = \frac{1}{2}k_{\vartheta}(\vartheta_{ABC} - \vartheta_0)^2 \tag{2.171}$$

These two terms are, respectively, 1-2 and 1-3 interactions. The last one, the 1-4 interactions given by Equation 2.172, can be subdivided in proper dihedrals, which describe the rotation around the BC bond (angle formed by the planes ABC and BCD, see Figure 2.6(C)), and improper dihedrals that describe the variation of the angle formed by the planes ABC and ACD in Figure 2.6(D). The latter are used to ensure planarity and, sometimes chirality of some groups.

$$\begin{aligned} \text{Proper dihedrals} \qquad & U_{A-D}(\varphi_{A-D}) = \frac{1}{2}k_{\varphi}(1 + cos(n\varphi_{A-D} - \varphi_0)) \\ \text{Improper dihedrals} \qquad & U_{A-D}(\xi_{AB}) = \frac{1}{2}k_{\xi}^{qs}(\xi_{A-D} - \xi_0)^2 \end{aligned} \tag{2.172}$$

The most common function to describe the bond stretching, the bond angle bending and the improper dihedral is an harmonic potential, a quadratic function that reproduces oscillations around an equilibrium value (r_0, ϑ_0 or ξ_0) with a certain frequency, associated to the force constant (k_{AB}, k_{ϑ} or k_{ξ}). However, this formulation is only able to describe the behavior of the system close to equilibrium and it is not capable of describing bond breaking and formation. Hence, other functions like the Morse potentials can be used to obtain a better description at regions of the PES far from equilibrium. In the case of the proper dihedral, the description uses a periodic function where k_{φ} is related to the height of the energy barrier; n to the number of minima in the energy function and φ_0 is the phase factor associated to the position of the minima. This formulation can also be generalized to the description of all the dihedral angles, like it is done in the Amber20 software.[260] One of the problems of splitting the bonding terms in different contributions is that, in real systems, they are usually coupled. To avoid this problem, it is possible to describe crossed terms like in the Urey-Bradley potential, where the variation in energy upon angle bending depends on the distance between atoms in positions 1–3.

Non-bonding terms represent the interactions between atoms separated by more than three bonds or not bonded at all. They are divided in electrostatic and van der Waals interactions. The former comes from the interaction between two-point charges (q), expressed as a Coulombic potential:

$$U_{AB}^{Coulomb}(R_{AB}) = \frac{q_A q_B}{R_{AB}} \tag{2.173}$$

The latter, typically described by a Lennard-Jones 12-6 potential, comprises a repulsive term based on Pauli's principle and an attractive term that arises from dispersion forces:

$$U_{AB}^{6-12} = 4\epsilon_{AB}\left(\left(\frac{\sigma_{AB}}{R_{AB}}\right)^{12} - \left(\frac{\sigma_{AB}}{R_{AB}}\right)^{6}\right) \tag{2.174}$$

This expression includes the minimum of the potential well (ϵ) and interatomic distance at which the value of the potential is 0 (σ).

To apply a FF, it is important to determine the force and equilibrium constants that characterize the interactions. In practice, they are assumed to be transferable, which means that the functions developed on a small set of molecules can be applied to a larger group of them of similar characteristics, similar functional groups. Therefore, each atom of the system can be classified into a specific atom type depending on its surrounding and, for each atom type, these constants are parameterized. Nonetheless, although FFs are suitable to explore the conformational space of a system consisting of hundreds of thousands of atoms, as will be discussed in Section 2.2.7, the form of the potential in Equation 2.169 is not suitable to study phenomena which involve, for example, bond breaking and formation.

Along this thesis, different sets of FFs have been used to describe different biological systems. Specifically, the General Amber Force Field (GAFF)[261] has been used to describe small organic molecules, the ff19SB FF[262] to describe the luciferase protein and the TIP3P[263] to describe water molecules. Furthermore, the OxyLUciferin Force Field (OLUFF)[264] has been developed, as will be discussed in Chapter IV, in order to describe the behaviour of the OLU chromophore in both the ground and first electronic excited states.

The Seminario method - FFs are usually developed for the S_0. However, during this thesis not only the PES of the ground state but also that of the S_1, was aimed to be reproduced. Hence, in order to sample this two electronic surfaces, it is possible to parameterize the equilibrium and force constants values of a molecule by means of QM calculations. For instance, it is possible

to use the Seminario method[265], that relies on the harmonic approximation previously described in Section 2.1.8, to obtain them.

The equilibrium bond distances and angles can be directly taken from an optimized geometry computed with the preferred software, while the force constant values can be obtained from the Hessian tensor $[k]$ of the system:

$$[k] = k_{ij} = \frac{\partial^2 E}{\partial R_i \partial R_j} \tag{2.175}$$

The Hessian is exactly related (up to second-order on a Taylor series expansion) to the $3M$ component reaction force ΔF generated as a consequence of a small displacement ΔR of the M atoms in a molecular system, as expressed Equation 2.176. Moreover, it represents the intramolecular FF of the system. Thus, its eigenvalues λ_i are the $3M$ force constants (3 translational, 3 rotational and $3M-6$ vibrational modes) and its eigenvectors $\boldsymbol{\nu}_i$, the directions of the displacements of the corresponding normal modes.

$$\Delta F = -[k]\Delta R \tag{2.176}$$

In the case of the bond force constants, the previous equation can be simplified to a pairwise interaction in which the reaction force on an atom A (ΔF_A) is expressed as a function of the displacement of an atom B (ΔR_B) (see Equation 2.177 where the minus sign is comprised in the tensor). Therefore, the tensor is a 3x3 matrix that has 3 eigenvalues and eigenvectors.

$$\Delta F_A = [k_{AB}]\, \Delta R_B \tag{2.177}$$

If one or more eigenvalues or two of the eigenvectors are complex, then A and B are not bonded since the pair wise interaction is unstable. The force constant can have contributions from more than one eigenvalue λ of $[k_{AB}]$ when any of the eigenvectors coincide with the direction from A to B and so, it can be expressed as a projection of each of the eigenvectors $\boldsymbol{\nu}^{AB}$ in the unitary vector from A to B ($\boldsymbol{u}^{AB}$):

$$k_{AB} = \sum_{i=1}^{3} \lambda_i^{AB} |\boldsymbol{u}^{AB} \cdot \boldsymbol{\nu}_i^{AB}| \tag{2.178}$$

In the case of bond angles, the force constant is defined as:

$$\frac{1}{k_\vartheta} = \frac{1}{r_{AB}^2 \sum_{i=1}^{3} \lambda_i^{AB} |\boldsymbol{u}^{PA} \cdot \boldsymbol{\nu}_i^{AB}|} + \frac{1}{r_{CB}^2 \sum_{i=1}^{3} \lambda_i^{CB} |\boldsymbol{u}^{PC} \cdot \boldsymbol{\nu}_i^{CB}|} \tag{2.179}$$

where r_{AB} and r_{CB} represent the bond lengths and $\boldsymbol{u}^{PA}$ and $\boldsymbol{u}^{PC}$, the projection of the unitary vectors in the direction of the bond onto a unitary vector $\boldsymbol{u}_N$ perpendicular to the bonds AB and CB:

$$\boldsymbol{u}^{PA} = \boldsymbol{u}_N \times \boldsymbol{u}^{AB} \qquad \boldsymbol{u}^{PC} = \boldsymbol{u}_N \times \boldsymbol{u}^{CB} \tag{2.180}$$

$$\boldsymbol{u}_N = \frac{\boldsymbol{u}^{CB} \times \boldsymbol{u}^{AB}}{|\boldsymbol{u}^{CB} \times \boldsymbol{u}^{AB}|} \tag{2.181}$$

For both cases the order in which the atoms are considered does not affect the computation of the force constant since, taking into account the third law of Newton, a pair of forces acting on the two interacting objects are the same in modulus and direction but opposite senses. A similar formulation for the reparametrization of the dihedral angles can also be described, taking into account the projection of the bond unitary vectors to the unitary vectors perpendicular to the ABC and BCD planes. However, for small molecules dihedral angles are not usually parameterized, but directly taken from the GAFF, since bond and angle potentials contribute more to the potential energy and are, in principle, more relevant. Moreover, software's such as AMBER[260] only allow the description of the dihedral angles as a periodic function and not as a harmonic potential, as in the Seminario method. Hence, a parameterization of the proper dihedral angle would require a different approach, as the one followed in Chapter IV.

2.2.4 Thermodynamic Ensembles

A thermodynamic ensemble is the assembly of all possible microstates, all the possible states consistent with the thermodynamic constraints of the system.[266] *A priori*, the selection of one ensemble or another should provide the same results. However, different stages of the MD simulation might benefit from one specific ensemble, as they provide a control mechanism to ensure a proper description of the system. For example, during a simulation in which the temperature of the system is increased, it is beneficial to use an ensemble that maintains constant the volume to avoid an abnormal expansion. On the other hand, using an ensemble that fixes temperature and pressure allows a control of the density along the simulation since the volume is allowed to change.

There exist several types of control mechanisms that can be applied to the measurable properties of the system: (i) differential control, in which the thermodynamic quantity is fixed and no fluctuations are allowed; (ii) proportional control, in which some variables are coupled to a thermodynamic property and thus, it is corrected along the simulation by means of a coupling constant that determines the strength of the fluctuations of the property; (iii) integral control, in which an extended Hamiltonian, where variables that represent the

effect of an external system, are introduced to fix the state to the desired ensemble; and (iv) stochastic control, in which modified equations of motion are used to propagate the values of the variables coupled to the thermodynamic property.[252] Typically, the measurable properties that are controlled are temperature and pressure and the algorithms applying these constrains are denominated thermostats and barostats. However, it is also possible to introduce a constraint to induce energy conservation, as in the microcanonical ensemble (NVE). Thereupon, the thermostat and barostats used along this thesis will be explained, although there exist many others.

Langevin Thermostat - The Langevin thermostat[267, 268] is an stochastic thermostat that is used to control that the simulation runs within a canonical ensemble, an ensemble in which the number of particles, the volume and the temperature are constant (NVT). This model is based on collisions between virtual particles and the DOFs of the system. To do so, Gaussian friction random forces $\boldsymbol{F}_{rand}(t)$ with zero mean, taken from Stokes' law, are applied on each atom and, therefore, Equation 2.167 is modified to:

$$\boldsymbol{F}_i\left(\boldsymbol{R}_1, ..., \boldsymbol{R}_M\right) = -\frac{\partial U\left(\boldsymbol{R}_1, ..., \boldsymbol{R}_M\right)}{\partial \boldsymbol{R}_i} - \gamma m_i \boldsymbol{v}_i(t) + \boldsymbol{F}_{rand}(t) \tag{2.182}$$

where γ represents a friction constant parameter and the amplitude of the random force is determined by the second fluctuation dissipation theorem:

$$\langle \boldsymbol{F}_{rand,i}(t) \boldsymbol{F}_{rand,j}(t') \rangle = 2 m_i \gamma k_B T_0 \delta_{ij} \delta(t - t') \tag{2.183}$$

where k_B represents Boltzmann's constant and T_0, the objective temperature. The collision frequency, related to γ, should not be extremely high to avoid a big disturbance of the phase space trajectory and a strong loss of the memory of the particle and thus, a fast decay of dynamic correlation functions.

The instantaneous temperature (T) of the system can be calculated along the simulation through the definition of the averaged translational kinetic energy obtained from a Maxwell-Boltzmann distribution of molecular speeds.[269] This distribution, that agrees with the equipartition result, can be used to determine the average kinetic energy of a particle in a gas is given by Equation 2.168. Therefore, the average kinetic energy of a particle can be expressed as:

$$\langle T_{ki} \rangle = \langle \frac{1}{2} m_i \boldsymbol{v}_i^2 \rangle = \int_0^\infty \frac{1}{2} m_i \boldsymbol{v}_i^2 P(\boldsymbol{v}_i) d\boldsymbol{v}_i \tag{2.184}$$

Substituting Equation 2.168 into Equation 2.184:

$$\langle T_{ki} \rangle = \frac{1}{2} m_i 4\pi \left(\frac{m_i \beta}{2\pi}\right)^{\frac{3}{2}} \int_0^\infty \boldsymbol{v}_i^4 e^{-\frac{m_i \boldsymbol{v}_i^2 \beta}{2}} d\boldsymbol{v_i} \tag{2.185}$$

and knowing that the result of the integral is given by:

$$\int_0^\infty \boldsymbol{v_i}^4 e^{-\frac{m_i \boldsymbol{v}_i^2 \beta}{2}} d\boldsymbol{v}_i = \frac{3}{2}\left(\frac{k_B T}{m_i}\right)^2 \left(\frac{2\pi k_B T}{m_i}\right)^{\frac{1}{2}} \tag{2.186}$$

Then:

$$\langle T_{ki} \rangle = \frac{1}{2} m_i 4\pi \left(\frac{m_i \beta}{2\pi}\right)^{\frac{3}{2}} \frac{3}{2}\left(\frac{k_B T}{m_i}\right)^2 \left(\frac{2\pi k_B T}{m_i}\right)^{\frac{1}{2}} = \frac{3}{2} k_B T \tag{2.187}$$

On the other hand, the kinetic energy is usually expressed as:

$$\langle T_k \rangle = \sum_{i=1}^{M} \frac{1}{2} m_i \cdot \boldsymbol{v}_i^2 \tag{2.188}$$

Hence, combining Equation 2.187 for all the atoms of the system and Equation 2.188:

$$\langle T_{ki} \rangle = \sum_{i=1}^{M} \frac{3}{2} k_B T = \sum_{i=1}^{M} \frac{1}{2} m_i \cdot \boldsymbol{v}_i^2 \tag{2.189}$$

and solving for T:

$$T = \frac{\sum_{i=1}^{M} m_i \cdot \boldsymbol{v}_i^2}{3 M k_B} = \frac{\sum_{i=1}^{M} m_i \cdot \boldsymbol{v}_i^2}{k_B N_{DOF}} \tag{2.190}$$

being M the number of atoms in the system; k_B, the Boltzmann's constant and N_{DOF}, the number of DOF.

Monte Carlo Barostat - In the case of requiring an additional control of the pressure (P), like in the isothermal-isobaric ensemble (NTP), the use of a barostat is mandatory. One of the barostats employed along this thesis is the Monte Carlo (MC)[270] barostat, in which the control of the pressure is obtained stochastically by means of a trial volume change:

$$\Delta V = n_{rand} \Delta V_{max} \tag{2.191}$$

where n_{rand} represents a random number between -1 and 1 and ΔV_{max} is the volume variation limit chosen so that a typical MC acceptance ratio of about 40–50% is achieved. Then, the box lengths (l_i) and centre of mass coordinates ($\boldsymbol{r}_i$) of each molecule are rescaled according to:

$$l_i \longrightarrow l_i \sqrt[3]{\frac{V'}{V}} \qquad\qquad \boldsymbol{R_i} \longrightarrow (\boldsymbol{R_i} - \boldsymbol{c_i}) \sqrt[3]{\frac{V'}{V}} + \boldsymbol{c_i} \tag{2.192}$$

being $V' = V + \Delta V$ and $\boldsymbol{c}_i$ the coordinates of the centre of the periodic box. This new configuration is accepted with the standard MC probability of $e^{-\beta \Delta W}$ if the Metropolis weight function, given by Equation 2.193 where P_0 represents the target pressure, is larger than 0, and it is always rejected otherwise.

$$\Delta W = U' - U + P_0 \Delta V - N k_B T ln \frac{V'}{V} \tag{2.193}$$

Berendsen barostat - The other barostat used along this thesis is the Berendsen barostat,[271] in which a proportional control of the pressure is obtained by means of a weak coupling to an external constant pressure bath. In this case, and extra term is added to the definition of the velocities of the system in Equation 2.161 and so:

$$\boldsymbol{v}_i \longrightarrow \boldsymbol{v}_i - \frac{\xi \left(P_0 - P\right)}{3\tau_P} \boldsymbol{R}_i \tag{2.194}$$

where ξ represents the isothermal compressibility; P_0 and P, the objective and the instantaneous pressures, respectively, and τ_P is the time constant of the coupling, a characteristic timescale in which the pressure will approach to P_0. P can then be calculated from the kinetic energy (T_k, defined as in Equation 2.188), the volume (V) and the FF:

$$P = \frac{2}{3V} \left(T_k + \frac{1}{2} \sum_{i=1}^{N} \boldsymbol{R_i} \cdot \boldsymbol{F_i} \right) \tag{2.195}$$

This formulation represents a scaling of the coordinates $\boldsymbol{R}$ and box length L [252] $\boldsymbol{R} \rightarrow s\boldsymbol{R}$ and $L \rightarrow sL$ being s given by:

$$s = 1 - \frac{\xi \Delta t \left(P_0 - P\right)}{3\tau_P} \tag{2.196}$$

2.2.5 Periodic Boundary Conditions

MD simulations are preformed for a relatively small number of atoms, when compared to the usual amount of them in real systems, which modifies the ratio of atoms on the surface of the system and, therefore, in contact with vacuum. To overcome this problem, Periodic Boundary Conditions (PBC) are applied to the simulation. Within this model, the computed system[213] is replicated in space along the three dimensions (see a two-dimensional scheme in Figure 2.7), which allows accounting for the long-range non-bonding interactions while the equations of motion are solved only for the atoms in the primary cell.[213]

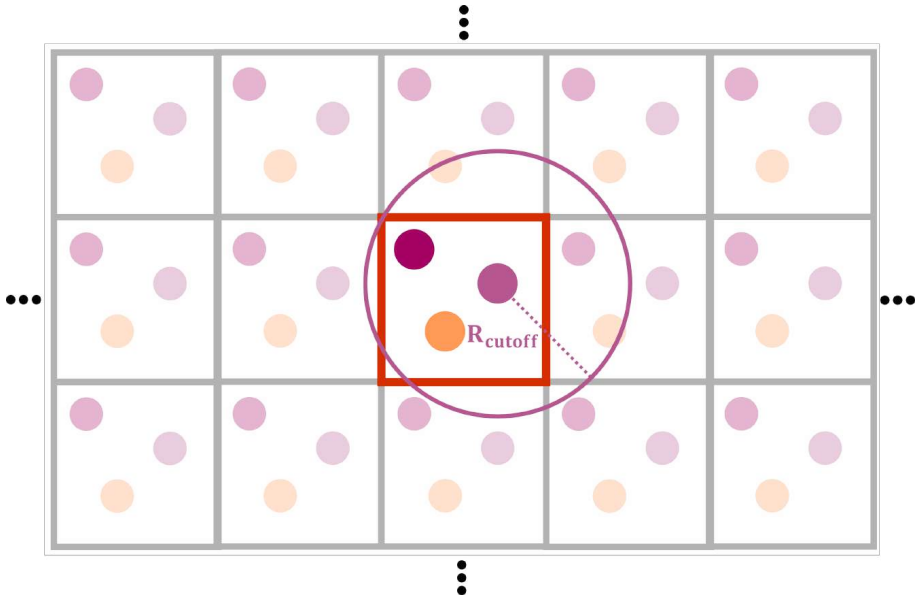


Figure 2.7: Schematic 2D representation of the replication of the primary cell (dark orange) in virtual cells (black) performed when periodic boundary conditions are applied to the MD simulation.

Each replica, called virtual cell, contains exactly the same amount of atoms as the primary cell and are obtained by adding to the coordinates of the system a set of vectors $\{\boldsymbol{n}\}$ expressed in terms of the axis direction ($\boldsymbol{L}_x$, $\boldsymbol{L}_y$, $\boldsymbol{L}_z$), whose length is the size of the box (L). Moreover, for each of the virtual atoms, momenta are the same as in the primary cell and therefore behave in the same way:

$$\boldsymbol{R}_i \longrightarrow \boldsymbol{R}_i + \boldsymbol{n} \qquad\qquad \boldsymbol{p}_i \longrightarrow \boldsymbol{p}_i \tag{2.197}$$

The set of $\{\boldsymbol{n}\}$ vectors is given by Equation 2.198 where a, b and c can have the values of -1, 0 and 1 and n represents the number of times in which the primary cell is replicated in a specific direction.

$$\boldsymbol{n} \longrightarrow n(a\boldsymbol{L}_x + b\boldsymbol{L}_y + c\boldsymbol{L}_z) \tag{2.198}$$

Additionally, if one atom leaves the primary box on one side during the simulation, another one with the same momenta enters it on the opposite side to ensure that the number of atoms is conserved. This description of the virtual cells saves a lot of memory in the calculation since they can be very easily computed from the coordinates and momenta of the primary cell.

In order to compute the non-bonding interactions between the atoms, a cutoff (R_{cutoff}), inside which these interactions are calculated for each atom, is

used to save computational time. However, this approach presents the problem of the energy continuity with the distance. In the case of van der Waals

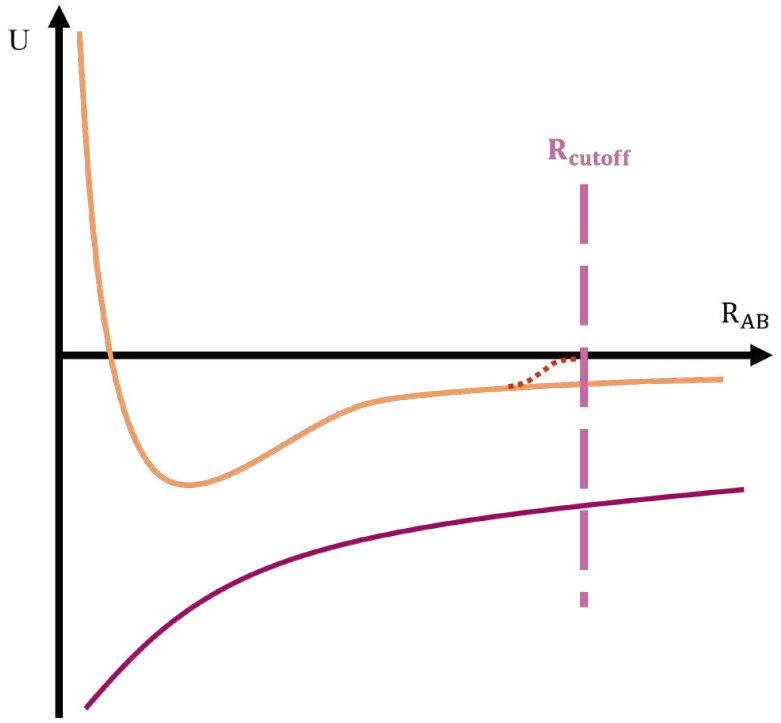


Figure 2.8: Schematic representation of the van der Waals (dark orange) and coulomb (dark pink) pair potentials as a function of the interatomic distance.

interactions, since they do not present a strong long-range character, and so they decay to 0 very fast as shown in Figure 2.8.The problem can be solved by multiplying the potential by a switching function $S(R)$ (represented in orange dashed line) that gradually reduces the van der Waals contributions to zero over a range of a few angstroms (R):

$$S(R) = \begin{cases} 1 & R_{AB} \leq R_0 \\ \frac{\left(R_{cutoff}^2 - R_{AB}^2\right)^2 \left(R_{cutoff}^2 + 2R_{AB}^2 - 3R_0^2\right)}{\left(R_{cutoff}^2 - R_0^2\right)^3} & R_0 \leq R_{AB} \leq R_{cutoff} \\ 0 & R_{AB} \geq R_{cutoff} \end{cases} \tag{2.199}$$

where:

$$R_0 = R_{cutoff} - R \qquad \text{and} \qquad R_{AB} = |\boldsymbol{R_A} - \boldsymbol{R_B}| \tag{2.200}$$

In the case of electrostatic interactions, this approach would lead to a completely wrong description of this contribution since the interaction at the cutoff distance in not negligible and so, introducing a switching function would lead to a lot of error. In addition, it can happen that, along the simulation time, some atoms that are around R_{cutoff} distance are not taken into account at certain the snapshots which would introduce some discontinuities in the PES. A way to solve this problem is by means of the Ewald summation method.[272]

The Ewald Summation

The Ewald summation method[272] is a widespread technique that accelerates the calculation of the M-body force summation of periodic or pseudo-periodic systems in which long-range contributions, such as electrostatic terms, are relevant.

This method consists in the application of a Gaussian charge distribution of opposite sign on top of each of the point charges of the system, as shown in Figure 2.9, to screen them. This transforms the long-range interactions into short-range interactions that are easier to compute. Then, the added Gaussian distributions are removed in the reciprocal space to compute the real long-range interactions.[252, 273] The electrostatic potential can be then computed

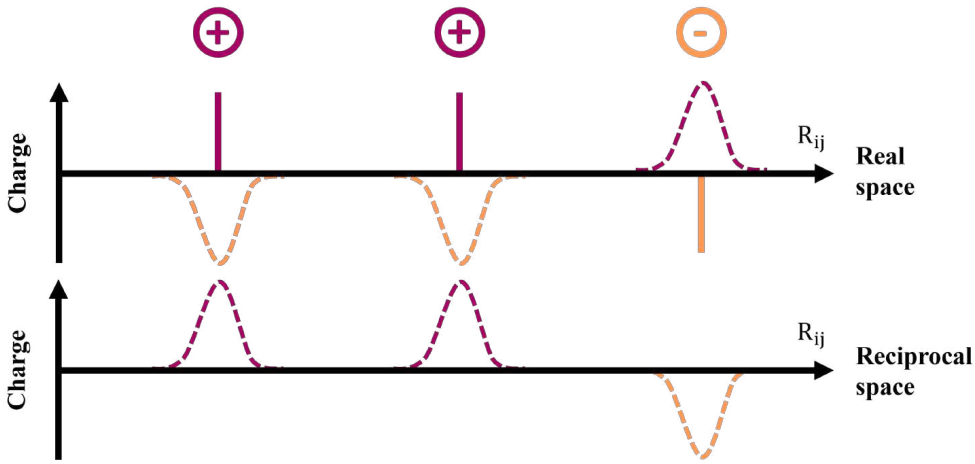


Figure 2.9: Schematic representation of the Ewald summation.

as a sum of the short-range screened potential V_{direct} that now decays very rapidly to 0 when $\boldsymbol{n} \longrightarrow \infty$ due to the error function $ercf$, the potential in the reciprocal space $V_{reciprocal}$ and a correction term $V_{correction}$:

$$V = V_{direct} + V_{reciprocal} + V_{correction} \tag{2.201}$$

that are given by:

$$V_{direct} = \frac{1}{2}\sum_{\boldsymbol{n}}^{'}\sum_{i,j=1}^{M}\frac{q_i q_j}{|\boldsymbol{R_i} - \boldsymbol{R_j} + \boldsymbol{n}|}ercf\left(\alpha|\boldsymbol{R_i} - \boldsymbol{R_j} + \boldsymbol{n}|\right) \tag{2.202}$$

$$V_{reciprocal} = \frac{1}{2\pi V}\sum_{i,j=1}^{M} q_i q_j \sum_{\boldsymbol{m}\neq 0}\frac{exp\left(-\left(\frac{\pi \boldsymbol{m}}{\alpha}\right)^2\right) + 2\pi i \boldsymbol{m}|\boldsymbol{R_i} - \boldsymbol{R_j}|}{\boldsymbol{m}^2} \tag{2.203}$$

$$V_{correction} = -\frac{\alpha}{\sqrt{\pi}}\sum_{i=1}^{M} q_i^2 \tag{2.204}$$

where $'$ represents that each i atom is able to interact with its replicas but not with itself; α is the width of the Gaussian and $\boldsymbol{m}$, a reciprocal-space vector expressed as:

$$\boldsymbol{m} = m_1\boldsymbol{B_x} + m_2\boldsymbol{B_y} + m_3\boldsymbol{B_z} \tag{2.205}$$

with:

$$\begin{aligned}
\boldsymbol{B}_x &= 2\pi\frac{\boldsymbol{L}_y \times \boldsymbol{L}_z}{\boldsymbol{L}_x\left(\boldsymbol{L}_y \times \boldsymbol{L}_z\right)} \\
\boldsymbol{B}_y &= 2\pi\frac{\boldsymbol{L}_z \times \boldsymbol{L}_x}{\boldsymbol{L}_y\left(\boldsymbol{L}_z \times \boldsymbol{L}_x\right)} \\
\boldsymbol{B}_z &= 2\pi\frac{\boldsymbol{L}_x \times \boldsymbol{L}_y}{\boldsymbol{L}_z\left(\boldsymbol{L}_x \times \boldsymbol{L}_y\right)}
\end{aligned} \tag{2.206}$$

2.2.6 Constrained Dynamics

The restriction of the motion in MD simulations presents several advantages: (i) saving computational time by neglecting several DOF, (ii) increasing of the time step and therefore the simulation time, (iii) enhancing the sampling by applying external forces to keep the system in a certain state and (iv) preventing wrong descriptions due to models that poorly describe certain DOF, like in the case of the O−H bonds in the TIP3P model.[263] However, these constrains need to be applied carefully because they can lead to poor results.[252]

The fixing of the bond length can be achieved by introducing Lagrangian multipliers, as shown in Equation 2.207, to the first expression at Equation 2.156:

$$\frac{d\boldsymbol{p}_k}{dt} = -\frac{dV}{d\boldsymbol{R}_k} + g'(\boldsymbol{R}_k)\lambda \tag{2.207}$$

with a bond distance constrain given by:

$$g'(\boldsymbol{R}_k) = R_{kB}^2 - d_{kB}^2 \tag{2.208}$$

where d_{AB}^2 represents the target distance between the atoms k and B. Additionally, it is also possible to impose the constraints in an iterative way. First, an unconstrained motion of the system is obtained, and then the positions are corrected:

$$\boldsymbol{R}_k(t + \Delta t) \longrightarrow \boldsymbol{R}_k(t + \Delta t) + \frac{\Delta t^2}{2m_k} \sum_{\gamma} F_{\gamma}^c \tag{2.209}$$

where γ runs over all the constraints and the constrained forces F_{γ}^c are expressed as:

$$F_{\gamma}^c = \frac{\mu}{2\Delta t^2} \left(\frac{|\boldsymbol{d}_0 - \boldsymbol{d}|}{|\boldsymbol{d}_0 \boldsymbol{d}|} \right) \tag{2.210}$$

being μ the reduced mass; $\boldsymbol{d}_0$, the constrained bond vector and $\boldsymbol{d}$, the unconstrained bond vector. This numerical method is used in the algorithms SHAKE,[274] developed by Jean-Paul Ryckaert et al. in 1977, and in its extension RATTLE,[275] developed by Hans C. Andersen in 1983, in which also momenta are constrained. Furthermore, a similar approach is followed in the SETTLE algorithm,[276] an analytical version of both SHAKE and RATTLE algorithms employed for rigid water models such as the TIP3P FF that has been used along this thesis.

2.2.7 The Sampling

Biological systems, such as the OLU/luciferase complex, present many possible stable geometrical configurations due to the big amount of DOF. Hence, to properly describe the properties of the system, it is important to account for the whole configurational space, within a specific ensemble. Since, as previously discussed, a full QM description of the motion would be unfeasible, to obtain a proper sampling it is possible to use deterministic simulations, given by the MD method presented in Section 2.2.[252] This sampling is based on the ergodic hypothesis that states that the probability of finding the system in some region of the configurational space that present the same energy over a long period of time is proportional to the volume of this region and therefore that all accessible configurations are equiprobable over a long time period.

Within the MD framework, and considering that the time step for integrating Hamilton's equations of motion is between 1 and 2 fs, the maximum affordable simulation time is short, up to microseconds (although some groups are already starting to reach the millisecond timescale). Thus, considering that

these simulations are usually performed at normal conditions (293.15 K and 1 atm), only the local area around the starting point is sampled since only relatively small barriers (a few tens of a kJ/mol) can be overcome. Nonetheless, this is not a real problem because MD methods are typically used for sampling the conformational space when the starting geometry is derived from experimental information, such as an X-ray or Nuclear Magnetic Resonance (NMR) structures. Hence, they can be used to search out the low-energy deformation modes and for sampling the conformational space for large confined systems.[213]

Apart from deterministic MD simulations, it is also possible to sample the conformational space using trial geometries that are accepted or not my means of a Metropolis algorithm, without any knowledge of the time propagation. This methodology is covered by the MC methods[277], which will not be further discussed as they have not be employed at any point of this thesis.

2.2.8 Typical Analysis for Molecular Dynamics Simulations

Once the time propagation of the system is simulated, the sampled geometries can be used to perform a wide variety of analysis of the dynamical properties of the system under study. All of them employ the relative position of the atoms of the system and in some cases, they also require information about the velocities and atomic charges.

The most straight forward analysis consists of the measurement of distances R_{AB} (Equation 2.211) and angles α (Equation 2.212) between specific atoms or residues along the simulation time, which can be then averaged to obtain a general picture of the system.

$$R_{AB} = |\boldsymbol{R}_A - \boldsymbol{R}_B| \tag{2.211}$$

$$\alpha = arcos\left(\frac{(\boldsymbol{R}_A - \boldsymbol{R}_B)(\boldsymbol{R}_C - \boldsymbol{R}_B)}{|\boldsymbol{R}_A - \boldsymbol{R}_B||\boldsymbol{R}_C - \boldsymbol{R}_B|}\right) \tag{2.212}$$

Using these two criteria, it is also possible to study not only the number of contacts between atoms or residues as a qualitative way to investigate the non-bonding interactions between two regions of interest of the system, but also the hydrogen bonding pattern among them. The former only takes into account a distance threshold to determine whether there exists or not a contact between the regions, while the latter also considers an angle threshold and the atom type, as hydrogen bonding only appears when there is a H atom bridging two electronegative ones, usually O atoms.

Moreover, it is also possible to measure the average distance between the atoms of a superimposed ensemble to quantify how much the system diverges from a reference structure over time. This measurement, which typically employs the first structure of the time evolution as a reference, is obtained from the Root Mean Square Deviation (RMSD) of the atomic positions:

$$RMSD = \sqrt{\frac{1}{M}\sum_{k=1}^{M}\left|\boldsymbol{R}_A - \boldsymbol{R}_A^{ref}\right|^2} \tag{2.213}$$

Furthermore, the flexibility of the system can be measured by means of the Root Mean Square Fluctuation (RMSF) (Equation 2.214). This quantity; which, in contrast to the RMSD, is calculated using an averaged structure; reveals which regions of the system are more mobile. Hence, a region with a larger RMSF value presents larger differences from the reference structure, thus indicating a larger mobility. If this analysis is restricted to the C_α carbons of a protein, it can help to visualize conformational changes along the dynamics.

$$RMSF_A = \sqrt{< \left|\boldsymbol{R_A} - \boldsymbol{R_A^{ref}}\right|^2 >} \tag{2.214}$$

Additionally, it is also possible to characterize the solvation sphere of a region of interest. This is achieved by means of the Radial Distribution Functions (RDF) (Equation 2.215), which effectively counts the average number of b neighbours in the shell at distance R around an a particle and represents its density. This value is normalized so it tends to 1 for large separations in a homogeneous system.

$$g_{ab} = \frac{1}{M_A}\frac{1}{M_B}V\sum_{A=1}^{M_A}\sum_{B=1}^{M_B} < \delta(|\boldsymbol{R}_A - \boldsymbol{R}_B| - R) > \tag{2.215}$$

Infrared Spectrum

Other property that can be calculated from a classical MD trajectory is the Infrared (IR) spectrum of the system. In this case, both velocities, derived during the simulations, and charges from the FF are required for the calculation of the dipole derivative autocorrelation function:

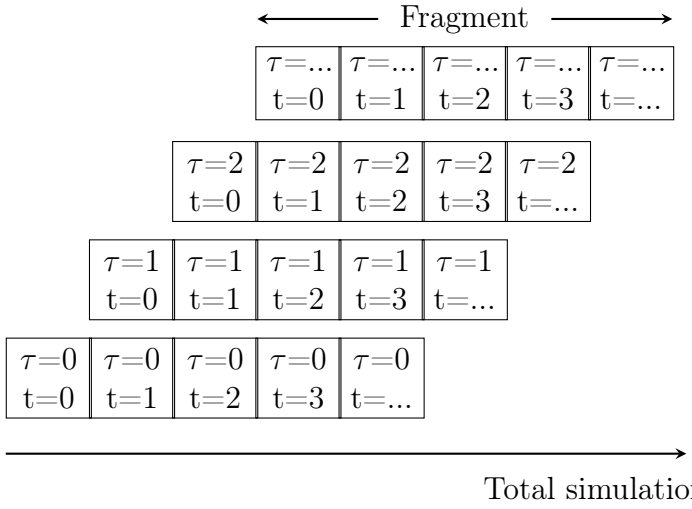
$$C_{\dot{\boldsymbol{\mu}}\dot{\boldsymbol{\mu}}}(t) =< \dot{\boldsymbol{\mu}}(t+\tau)\dot{\boldsymbol{\mu}}(\tau) > \tag{2.216}$$

where the dipole derivative $\dot{\boldsymbol{\mu}}$ is given by:

$$\dot{\boldsymbol{\mu}}(\boldsymbol{t}) = \sum_{A=0}^{M} q_A \boldsymbol{v}_A(t) \tag{2.217}$$

Moreover, it is possible to account for different time origins τ following Scheme 2.1 and to compute the time-averaged–trajectory-averaged dipole derivative autocorrelation function:

$$< C_{\dot{\boldsymbol{\mu}}\dot{\boldsymbol{\mu}}}(t) >= \frac{< \dot{\boldsymbol{\mu}}(t+\tau)\dot{\boldsymbol{\mu}}(\tau) >}{< \dot{\boldsymbol{\mu}}(\tau=0)\dot{\boldsymbol{\mu}}(\tau=0) >} \tag{2.218}$$



Scheme 2.1: Calculation of the dipole derivative autocorrelation function. t represents the internal time of each of the trajectory fragments while τ is the time displacement of each of the trajectory fragments with respect to the beginning of the simulation.

This autocorrelation function is then related to the intensity of the infrared spectrum *I(ω)* by its Fourier transform:[278, 279]

$$I(\omega)\ \alpha\ \int_0^\infty < C_{\dot{\boldsymbol{\mu}}\dot{\boldsymbol{\mu}}}(t) > e^{-i\omega t}dt \tag{2.219}$$

Principal Components Analysis

Analogously to the QM picture represented in Section 2.1.8, classical MD simulations also allow the characterization of the relative movement of the M nuclei.[280] Each of these independent modes of movement, usually denominated Principal Components (PC), can be calculated from the Mass-Weighted (MW) covariance matrix C^{covar} of the trajectory, whose elements are given by:

$$C_{AB}^{covar} =< \sqrt{M_{AA}}(R_A - R_A^{ref})\sqrt{M_{BB}}(R_B - R_B^{ref}) > \tag{2.220}$$

where <> denotes the time average, M corresponds to the diagonal matrix of atomic masses and the reference structure usually refers to the average structure along the dynamics. Upon the diagonalization of C^{covar}, the PC described by the eigenvectors of the transformation matrix while the frequency of the motion, in s^{-1}, can be calculated from the λ_i eigenvalues of the diagonal matrix:[281]

$$\omega_1 = \frac{1}{2\pi}\sqrt{\frac{k_B T}{\lambda_i}} \tag{2.221}$$

An important consideration to take into account during this analysis is that the system needs to be aligned to the reference structure in order to eliminate the rotational and translational degrees of freedom. This allows for a more direct comparison to the normal modes of the system computed from a QM frequency calculation.

2.3 Hybrid QM/MM Approaches

Classical MD methods allow for a fast evaluation of very large systems while accounting for environmental effects. However, they suppose a loss in the accuracy of the results. To overcome this problem, the theory behind hybrid Quantum-Mechanics/Molecular-Mechanics (QM/MM) approaches have been deduced.[282–285] This approach began its development in the 1970s,[283, 284] but it was not until the late 1990s[286] that it finally emerged as a relevant method for the study of large biological macromolecules, culminating in the Nobel Prize in Chemistry 2013 award *for the development of multiscale models for complex chemical systems* to its developers: Martin Karplus, Michael Levitt and Arieh Warshel. In this methodology, the system is usually divided in two regions, each treated at a different level of theory. The inner (I) region, treated with a QM method, is composed by the part of the system that is of interest. For example, a chromophore, as in this thesis, or the reacting site. The rest of the system, the solvent and some other environmental molecules like proteins, are described at MM level, given by the previously described treatment in Section 2.2.

There exist two approaches to describe the QM/MM energy of the system, that are graphically represented in Figure 2.10: (i) the substractive and (ii) the additive schemes. In the former, the energy E^{subs} is computed as:

$$E^{subs} = E^{QM}(I) + E^{MM}(I+O) - E^{MM}(I) \tag{2.222}$$

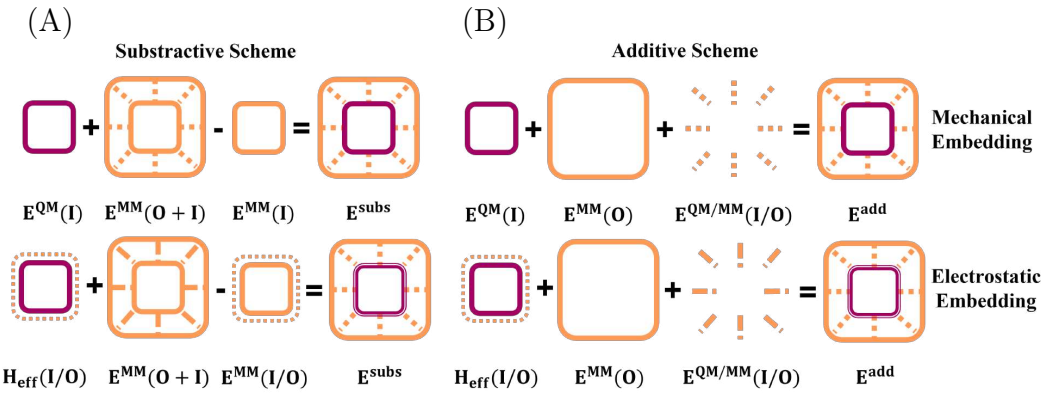


Figure 2.10: Graphical representation of the QM/MM schemes and embeddings. Light orange and dark pink represent, respectively, the classic and QM treatments of the inner (I) and outer (O) parts. Dashed lines represent the point charges of the outer region.

where $E^{QM}(I)$ represents the energy of the inner region computed at QM level and $E^{MM}(I+O)$ and $E^{MM}(I)$, the classical energies of the whole system and the inner region, respectively. As shown Figure 2.10(A), this approach accounts for the interaction energy between the two regions during the calculation of the whole system. Moreover, it computes twice the energy of the inner region (classically and QM) and therefore, it requires a substrative term to avoid the repetition.

In the additive scheme, represented in Figure 2.10(B), the energy of the inner and outer parts, as well as the interactions between them ($E^{QM/MM}(I/O)$), are computed separately:

$$E^{add} = E^{QM}(I) + E^{MM}(O) + E^{QM/MM}(I/O) \tag{2.223}$$

where $E^{QM/MM}(I/O)$ is given by the sum of the classically described bonding energy $E^{MM}_{bonded}(I/O)$, the van der Waals interactions E^{MM}_{vdW} and the electrostatic interactions E_{elec}:

$$E^{QM/MM}(I/O) = E^{MM}_{bonded}(I/O) + E^{MM}_{vdW} + E_{elec} \tag{2.224}$$

Regardless of the QM/MM scheme, the embedding can be described as: (i) mechanical, (ii) electrostatic or (iii) polarizable. In the mechanical embedding, $E^{mechanical}_{elec}$ is computed by means of the coulomb expression:

$$E^{mechanical}_{elec} = \sum_{I}^{M_{QM}} \sum_{O}^{M_{MM}} \frac{q^{QM}_I q^{MM}_O}{|\boldsymbol{R}_I - \boldsymbol{R}_O|} \tag{2.225}$$

where M_{QM} and M_{MM} represent the number of QM and classical nuclei, respectively; q_I^{QM} can be calculated from the Mulliken population analysis as in Equation 2.51 and q_O^{MM} is directly taken from the FF. This model does not take into account the polarization of the inner layer by the charges of the outer one because the interaction is computed classically. Furthermore, it has limitations when describing the interface between the inner and outer layers, leading to potential artifacts in the system's properties. In the case of the electrostatic embedding, the inner region is polarized by the charges of the outer region by means of an effective Hamiltonian that accounts for both the energy of the QM inner region and its QM interaction with the outer part:

$$\hat{\mathcal{H}}^{eff} = \hat{\mathcal{H}}^0(I) + \hat{\mathcal{H}}(I/O) \tag{2.226}$$

where the interaction Hamiltonian $\hat{\mathcal{H}}^0(I/O)$ is given by:

$$\hat{\mathcal{H}}(I/O) = \sum_{O}^{M_{MM}} q_O^{MM}\left(\boldsymbol{R_O}\right) \int d\boldsymbol{r} \frac{\phi_\mu(\boldsymbol{r})\phi_\nu^*(\boldsymbol{r})}{|\boldsymbol{r} - \boldsymbol{R}_O|} \tag{2.227}$$

Finally, in the polarizable embedding, both the inner and outer parts are polarized in self-consistent way by means of an additional polarization term $\hat{\mathcal{H}}_{pol}$ in the interaction Hamiltonian:

$$\begin{aligned} \hat{\mathcal{H}}(I/O) &= \hat{\mathcal{H}}_{elec}(I/O) + \hat{\mathcal{H}}_{pol}(I/O) \\ &= \sum_{O}^{M_{MM}} q_O^{MM}\left(\boldsymbol{R_O}\right) \int d\boldsymbol{r} \frac{\phi_\mu(\boldsymbol{r})\phi_\nu^*(\boldsymbol{r})}{|\boldsymbol{r} - \boldsymbol{R}_O|} - \sum_a \mu_a^O\left(\boldsymbol{R_a}\right) E^I\left(\boldsymbol{R_a}\right) \end{aligned} \tag{2.228}$$

where $\mu_a^0\left(\boldsymbol{R_a}\right)$ represents the dipoles of the outer atoms located at $\boldsymbol{R_a}$ and $E^I\left(\boldsymbol{r_a}\right)$, the electric field created by inner atoms at $\boldsymbol{R_a}$.

As mentioned before, a typical problem of multiscale systems appears for the description of the interface between the inner and outer layers.[211] This effect is specially stressed when both regions are covalently bonded since the interface between the layers results in unsaturated atomic valences in the inner region, which means that the interface artificially generates a radical at the boundary atom of the inner region with the outer region, resulting in a unrealistic electronic structure of the system. A general approach to overcome this problem is the link-atom approximation, in which a hydrogen atom is added to the boundary atom of the inner region to re-saturate the unpaired electron that remains unpaired after the breaking of the bond with the boundary atom of the outer region. However, the addition of this hydrogen atom in the inner region leads to extremely high and unrealistic interactions with the

point charge representing the boundary atom of the outer region due to its proximity to this link atom. Hence, to avoid this situation, the point charge representing the boundary atom of the outer layer is often removed, and its charge is redistributed among the nearby point charges in the MM region.

Finally, it is important to notice that there exist two different types of QM/MM calculations. The first of them is static, in the sense that the MD simulation is used for the sampling of the conformational space of the system and, QM/MM SPs are calculated on top of the classically sampled geometries to obtain specific QM properties. Nonetheless, it is also possible to calculate the QM/MM potential gradient at every step of the trajectory in order to propagate the system along the QM/MM PES. This dynamic approach is much more computationally demanding since QM energies and gradients need to be computed for every step. However, the accuracy of the description is higher, and it also allows for the study of a wider variety of processes such as excited state deactivation and chemical reactivity.

2.4 Interaction Energy

An important part of the work presented in this thesis is dedicated to the study of the interactions between the OLU emitting species and the luciferase protein, the biological environment in which the chromophores are embedded in. This interaction energy can be described at both classical and QM/MM levels by means of the MM-Generalized Born (GB) Surface Area (SA) method and QM/MM-Energy Decomposition Analysis (EDA), respectively, from representative snapshots of an ensemble of conformations. This section is dedicated to introduce the theoretical framework behind these two different approaches for the description of the interaction energy.

2.4.1 MM-Generalized Born Surface Area

The MM-GBSA method[287–289] is used to classically calculate the free energy change between two states, typically a bound and free state of a protein receptor (P) and a ligand (L) isolated, and the complex (C) when both P and L are interacting. This method, which achieves a good balance between computational efficiency and accuracy, has been widely used to evaluate docking poses, determine structural stability, and predict binding affinities and hotspots. Furthermore, it allows the analysis of the contributions from individual residues or energy terms by free energy decomposition analysis, resulting in detailed residue-specific energetic contributions to the system binding, iden-

tifies dominant interactions in the binding process, and therefore facilitates individualized drug design.[290]

In this method, the binding free energy in solution ΔG^{Sol}_{Bind}, represented in the upper part of Figure 2.11 and defined from the Gibbs free energies of the individual solvated systems (Equation 2.229), is expressed in terms of the thermodynamic cycle represented in Figure 2.11 (Equation 2.230).

$$\Delta G^{Sol}_{Bind} = G^{C,Sol} - G^{L,Sol} - G^{P,Sol} \tag{2.229}$$

$$\Delta G^{Sol}_{Bind} = \Delta G^{Vac}_{Bind} + \Delta G^{C}_{Solv} - T\Delta S^{Sol}_{Bind} \tag{2.230}$$

where ΔG^{Vac}_{Bind} represents the binding free energy in vacuum; ΔG^{C}_{Solv}, the free energy of solvation of the complex (Solv) and ΔS^{Sol}_{Bind}, the difference in conformational entropy upon binding. Typically, the ΔG^{Vac}_{Bind} is approximated to the gas-phase MM energy ΔU^{MM}, which can be calculated from the FF using the expressions in Section 2.2.3:

$$\Delta G^{Vac}_{Bind} = \Delta U^{MM} = \Delta U^{bonding} + \Delta U^{Coulomb}_{AB} + \Delta U^{6-12}_{AB} \tag{2.231}$$

where $\Delta U^{bonding}$ represents the changes in the internal energies; $\Delta U^{Coulomb}_{AB}$, the changes in the electrostatic energies and ΔU^{6-12}_{AB}, the changes in the van der Waals energies.

The free energy of solvation is the sum of a polar contribution coming from the electrostatic solvation energy ΔG^{GB} and a non-polar contribution ΔG^{SA}

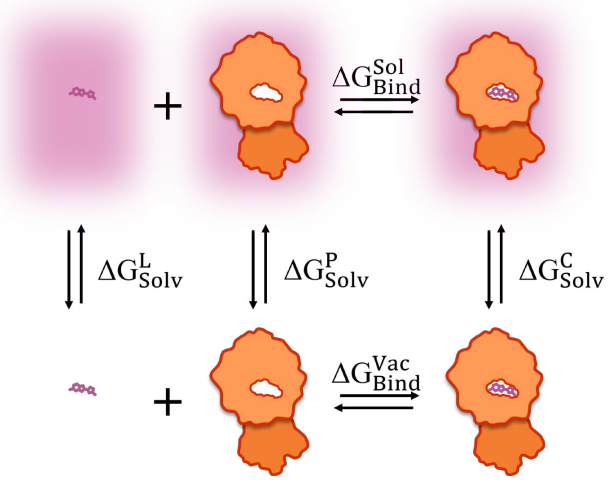


Figure 2.11: Schematic representation of the thermodynamic cycle used for the MM-GBSA calculation. The ligand is represented in dark pink; the receptor, in orange and the continuum solvent, in light pink.

between the solute and a continuum solvent:

$$\Delta G^{C}_{Solv} = \Delta G^{GB} + \Delta G^{SA} \tag{2.232}$$

The polar contribution is calculated using, for instance, the GB model, which presents an analytical expression for the polar solvation energy that allows for a faster performance than other methods. In this model,[291, 292] atoms are described as charged spheres of radius a_A whose internal dielectric constant ε_{in} is lower than that of the continuous environment ε_{Sol} so that the screening that each atom experiences is determined by their local environment. Hence, a larger presence of other atoms surrounding a specific charged sphere would result in a lower screening of its electrostatics since it is surrounded by a lower-dielectric material. This property, called descreening of one atom by another, is used to calculate the effective Born radius of each atom B^r_A, that is related to its the degree of descreening. On the one hand, a large Born radius would represent a small screening (and so a strong electric field), as if the atom was in vacuum. On the other hand, a small Born radius is related to a large screening (weak electric field), as if the atom was in bulk solvent. The canonical GB equation[293] with the absence of salt can be written as:[294]

$$\Delta G^{GB} = -\left(\frac{1}{\varepsilon_{in}} - \frac{1}{\varepsilon_{Sol}}\right)\sum_{AB}^{M}\frac{q_A q_B}{f^{GB}} \tag{2.233}$$

However, it is possible to include the electrostatic screening effects of (monovalent) salts via the Debye–Huckel screening parameter κ, which depends on the salt concentration:[295]

$$\Delta G^{GB} = -\left(\frac{1}{\varepsilon_{in}} - \frac{e^{-\kappa f^{GB}_{AB}}}{\varepsilon_{Sol}}\right)\sum_{AB}^{M}\frac{q_A q_B}{f^{GB}_{AB}} \tag{2.234}$$

These two equations present a smooth function f^{GB}_{AB} of the $A-B$ interatomic distance R_{AB} and their effective Born radii given by:

$$f^{GB}_{AB} = \sqrt{R^2_{AB} + B^r_A B^r_B \cdot exp\left(\frac{R^2_{AB}}{4B^r_A B^r_B}\right)} \tag{2.235}$$

The effective Born radii can be described in many ways. However, the Amber[260] software uses the following expression:[296]

$$(B^r_A)^{-1} = \bar{a_A}^{-1} - a_A^{-1} tanh(\alpha\varsigma_A - \beta\varsigma_A^2 + \gamma\varsigma_A^3) \tag{2.236}$$

with $\alpha = 1.0$, $\beta = 0.8$ and $\gamma = 4.85$ and:

$$\varsigma_A = \frac{\bar{a_A}^{-1}}{4\pi} \int_{vdW} \Theta(|\boldsymbol{R}| - \bar{a_A}) r^{-4} d\boldsymbol{R} \tag{2.237}$$

where the integral is taken over the van der Waals volume of the atomic spheres and $\Theta(|\boldsymbol{R}| - \bar{a_A})$ represents a step function.

In the case of the non-polar energy ΔG^{SA}, it is usually estimated using the Solvent-Accessible Surface Area (SASA) approximation,[297] which accounts for the unfavourable formation of a solute cavity within a continuum solvent and the favourable van der Waals interactions between the solute and the solvent around the cavity. In this method, the solvation free energy is given by an empirical formula in which the solvent accessible surface area of the solute $SASA$ is proportional to this magnitude:

$$\Delta G^{SA} = \gamma \cdot SASA + b \tag{2.238}$$

where γ is the surface tension constant and b is a correction constant. In this sense, the ΔG^{SA} term can be understood as the free energy of solvation process of a molecule from which all charges have been removed.

Counter-intuitively, the polar term is not favourable to the binding energy. When the ligand and the receptor form a complex, their combined SA that is exposed to the solvent is reduced when compared to their separate states. Hence, fewer solvent molecules can interact with the complex with the unbounded states, resulting in a less favourable (more positive) solvation free energy for the complex. On the flip side, the non-polar terms results in a favourable interaction, as the combined SA exposed to the solvent is reduced when the ligand and receptor form a complex.

Finally, the change in conformational entropy $-T\Delta S^{Sol}_{Bind}$ can be calculated by normal-mode analysis[287] on a set of conformational snapshots taken from a MD simulations. However, due to the large computational cost, changes in the conformational entropy are usually neglected when only the relative binding free energies of similar ligands are needed.[290]

The most accurate protocol for the calculation of the binding free energies by means of the MM-GBSA approach is the so-called three-average (3A) MM-GBSA, in which three independent MD simulations for the isolated ligand, the protein receptor, and the bound complex are propagated for the generation of a set of conformations for each of the independent system. Then, the sampled geometries are used to calculate the averages of each of the free energies and the total binding energy:

$$\Delta G^{Sol}_{Bind} = \langle G^{C,Sol} \rangle^C - \langle G^{L,Sol} \rangle^L - \langle G^{P,Sol} \rangle^P \tag{2.239}$$

In practice, it is preferable to propagate just one MD trajectory for the bound complex to obtain the structures of all three components. This one-average (1A) approach, which is based on the assumption that the structures sampled for the complex in solution are sufficiently similar to those sampled for the isolated protein and ligand, gives more accurate results with lower standard errors because the internal energy errors of the systems can be cancelled due to the use of exactly identical configurations for the calculation of the energy difference between the complex and its individual components:

$$\Delta G_{Bind}^{Sol} = \langle G^{C,Sol} - G^{L,Sol} - G^{P,Sol} \rangle^{C} \tag{2.240}$$

Moreover, it is important for reliable prediction of the binding free energies to calculate the MD trajectory in an explicit water solvent system since water molecules play an important role in protein–ligand interactions during the propagation. For example, water molecules participate in the hydrogen bond network of the system, and therefore they can displace the ligand inside the active site of the enzyme by either bridging it to the protein or filling a region of the pocket.

2.4.2 Energy Decomposition Analysis

Another approach for the calculation of interaction energies is the EDA, a QM-based method that has been described in literature with either variational or perturbative approaches.[298–301] The EDA scheme employed in Chapter VII of this thesis corresponds to the perturbative approach devised by Marcos Mandado and collaborators in 2011,[302, 303] which is based on the electronic deformation densities that describe the first-order perturbation of the electron density.

Given a QM system consisting of two interacting monomers A and B, the interaction energy between them is defined, equivalently to that of the MM-GBSA approach in Equation 2.229, as the difference between the energy of the complex AB (the entire system) minus the energies of the monomers A and B in the same geometry:

$$E_{int} = E_{AB}^{AB} - E_{A}^{AB} - E_{B}^{AB} \tag{2.241}$$

where the superscript AB indicates that the basis set of the complex is being used in order to correct for Basis Set Superposition Error (BSSE)[304, 305] with the Counterpoise (CP) correction.[306]

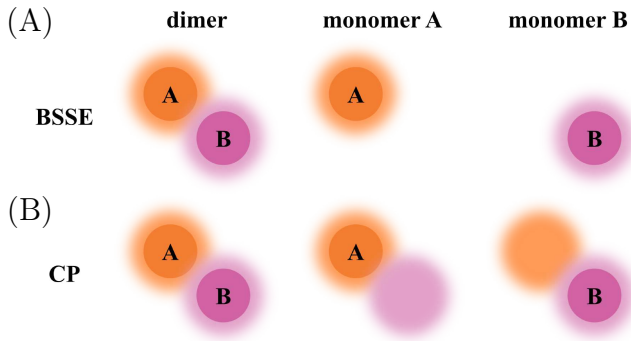


Figure 2.12: Graphical representation of the (A) BSSE and (B) CP correction; where solid circles represent the presence of monomers A (orange) and B (pink) and blurred circles, their corresponding basis set.

The BSSE, graphically represented in Figure 2.12, consists on the energy compensation for the basis set incompleteness of each of the independent monomers that arises in dimers from the presence of basis set of the other monomer. The dimer would therefore be artificially lowered in energy when exclusively compared to the individual monomers. In the limit of a complete basis set, the BSSE would be zero, and so adding more basis functions would not result in any improvement. Conceptually, the simplest approach for eliminating the BSSE is, therefore, to add more basis functions until the interaction energy no longer changes. Unfortunately, and considering that some non-bonded interactions are weak, this requires very large basis sets since inclusion of electron correlation is mandatory for calculating the dispersion interaction between molecules.[213] To overcome this computational conundrum, this artifact is typically corrected by means of the CP correction[307, 308] by introducing the basis set of the other monomer at the corresponding nuclear positions, but without its nuclei and electrons present. Therefore:

$$E_{int}^{CP} > E_{int}^{BSSE} \tag{2.242}$$

In an ideal situation, the energy of the complex E_{AB} and each of the monomers E_A and E_B would be calculated on the optimized structure of each of the systems and then, the CP correction ΔCP would be calculated, on the complex conformation, as:

$$\Delta CP = \left(E_A^{AB} - E_A^A\right) + \left(E_B^{AB} - E_B^B\right) \tag{2.243}$$

However, in a dynamic framework in which sampling is required, it is computationally challenging to obtain the optimized structures for the monomers

and the complex. Hence, in these cases, Equation 2.241 is directly applied to the sampled conformation and the interaction energy averaged in time.

As for the EDA framework, let's consider the one-particle density matrix $\rho(\boldsymbol{r}_1, \boldsymbol{r}'_1)$, defined in terms of the N-electron wavefunction $\Psi(\boldsymbol{x}_1, ..., \boldsymbol{x}_N)$:

$$\rho(\boldsymbol{r}_1, \boldsymbol{r}'_1) = \int\int d\sigma_1 d\boldsymbol{x}_2 ... d\boldsymbol{x}_N \Psi^*(\boldsymbol{r}_1, \omega_1, ..., \boldsymbol{x}_N)\Psi(\boldsymbol{r}'_1, \omega_1, ..., \boldsymbol{x}_N) \tag{2.244}$$

where $\boldsymbol{x}$ represents a spin coordinate collective variable as in Equation 2.20 and $\boldsymbol{r}_1$ and $\boldsymbol{r}'_1$ are two different coordinate variables for a particular electron. This one-particle density matrix is useful to represent non-multiplicative operators in terms of electron densities, such as, for instance, the kinetic energy operator. On the flip side, the two-particle density $\rho(\boldsymbol{r}_1, \boldsymbol{r}_2)$, defined as:

$$\rho(\boldsymbol{r}_1, \boldsymbol{r}_2) = N(N-1)\int\int d\sigma_1 d\sigma_2 d\boldsymbol{x}_3 ... d\boldsymbol{x}_N |\Psi^*(\boldsymbol{x}_1, ..., \boldsymbol{x}_N)| \tag{2.245}$$

implicitly accounts for the electron correlation between pairs of electrons since it can be understood as the probability density of finding an electron at $\boldsymbol{r}_1$ and another electron at $\boldsymbol{r}_2$ simultaneously. From these two density matrices, it is possible to define the total energy of a molecular system as:

$$\begin{aligned} E = \int d\boldsymbol{r}_1 h(\boldsymbol{r}_1)\rho(\boldsymbol{r}_1, \boldsymbol{r}'_1)|_{\boldsymbol{r}'_1=\boldsymbol{r}_1} + \int\int d\boldsymbol{r}_1 d\boldsymbol{r}_2 \frac{\rho(\boldsymbol{r}_1, \boldsymbol{r}_2)}{|\boldsymbol{r}_1 - \boldsymbol{r}_2|} \\ + \sum_{A=1}^{M-1}\sum_{B>A}^{M} \frac{Z_A Z_B}{|\boldsymbol{R}_A - \boldsymbol{R}_B|} \end{aligned} \tag{2.246}$$

where last term in represents the Coulomb repulsion between two nuclei and $h(\boldsymbol{r}_1)$ is the one-electron Hamiltonian, which represents the kinetic energy and the coulomb attraction between electrons and nuclei $\hat{\mathcal{V}}_M$ as given by Equation 2.26 from which the spin coordinate has been integrated out. In addition, if $\rho(\boldsymbol{r}_1, \boldsymbol{r}_2)$ is represented as a sum of an uncorrelated term, given by the product of two ordinary electron densities $\rho(\boldsymbol{r}_1)$ and $\rho(\boldsymbol{r}_2)$ at $\boldsymbol{r}_1$ and $\boldsymbol{r}_2$ as defined in Equation 2.58, plus an XC term:

$$\rho(\boldsymbol{r}_1, \boldsymbol{r}_2) = \rho(\boldsymbol{r}_1)\rho(\boldsymbol{r}_2) + \rho_{xc}(\boldsymbol{r}_1, \boldsymbol{r}_2) \tag{2.247}$$

then, Equation 2.246 can be re-written as:

$$\begin{aligned} E = &-\frac{1}{2}\int d\boldsymbol{r}_1 \nabla^2 \rho(\boldsymbol{r}_1, \boldsymbol{r}'_1)|_{\boldsymbol{r}'_1=\boldsymbol{r}_1} + \int d\boldsymbol{r}_1 \hat{\mathcal{V}}_M \rho(\boldsymbol{r}_1) \\ &+ \frac{1}{2}\int\int d\boldsymbol{r}_1 d\boldsymbol{r}_2 \frac{\rho(\boldsymbol{r}_1)\rho(\boldsymbol{r}_2)}{|\boldsymbol{r}_1 - \boldsymbol{r}_2|} + \frac{1}{2}\int\int d\boldsymbol{r}_1 d\boldsymbol{r}_2 \frac{\rho_{xc}(\boldsymbol{r}_1, \boldsymbol{r}_2)}{|\boldsymbol{r}_1 - \boldsymbol{r}_2|} \\ &+ \sum_{A=1}^{M-1}\sum_{B>A}^{M} \frac{Z_A Z_B}{|\boldsymbol{R}_A - \boldsymbol{R}_B|} \end{aligned} \tag{2.248}$$

where the first term can be interpreted as the kinetic energy operator acting upon $\rho(\boldsymbol{r}_1, \boldsymbol{r}'_1)$, after which the coordinate $\boldsymbol{r}'_1$ is set to be equal to $\boldsymbol{r}_1$. The rest of the terms in Equation 2.248 can be represented as a sum of contributions from a fictitious system formed by two non-interacting monomers A and B, whose interactions are accounted for by means of additional perturbative terms. Moreover, the last term of Equation 2.248, which represents the total internuclear repulsion $V_{M,M}$, can be divided into a nuclear repulsive potential involving exclusively nuclei of each monomer (V_{M_A,M_A} and V_{M_B,M_B}) and the nuclear repulsive potential involving nuclei of both monomers V_{M_A,M_B}:

$$V_{M,M} = V_{M_A,M_A} + V_{M_B,M_B} + V_{M_A,M_B} \tag{2.249}$$

which is given by:

$$\begin{aligned}\sum_{A=1}^{M-1}\sum_{B>A}^{M}\frac{Z_A Z_B}{|\boldsymbol{R}_A - \boldsymbol{R}_B|} = &\sum_{A=1}^{M_A-1}\sum_{B>A}^{M_A}\frac{Z_A Z_B}{|\boldsymbol{R}_A - \boldsymbol{R}_B|} \\ &+ \sum_{A=M_A+1}^{M-1}\sum_{B>A}^{M}\frac{Z_A Z_B}{|\boldsymbol{R}_A - \boldsymbol{R}_B|} + \sum_{A=1}^{M_A}\sum_{B=M_A+1}^{M}\frac{Z_A Z_B}{|\boldsymbol{R}_A - \boldsymbol{R}_B|}\end{aligned} \tag{2.250}$$

where M_A and M_B are, respectively, the total number of nuclei in monomers A and B. Furthermore, the nuclear electrostatic potential operator $\hat{\mathcal{V}}_M$, that appears in the second term of 2.248, can be split into two terms, one for each monomer:

$$\hat{\mathcal{V}}_M = \hat{\mathcal{V}}_{M_A} + \hat{\mathcal{V}}_{M_B} \tag{2.251}$$

and the electron density $\rho(\boldsymbol{r}_1)$ and the two-particle XC density into:

$$\rho(\boldsymbol{r}_1) = \rho_A(\boldsymbol{r}_1) + \rho_B(\boldsymbol{r}_1) + \Delta\rho_{Pau}(\boldsymbol{r}_1) + \Delta\rho_{pol}(\boldsymbol{r}_1) \tag{2.252}$$

and:

$$\begin{aligned}\rho_{xc}(\boldsymbol{r}_1, \boldsymbol{r}_2) = &\rho_{xc,A}(\boldsymbol{r}_1, \boldsymbol{r}_2) + \rho_{xc,B}(\boldsymbol{r}_1, \boldsymbol{r}_2) \\ &+ \Delta\rho_{x,AB}(\boldsymbol{r}_1, \boldsymbol{r}_2) + \Delta\rho_{xc}(\boldsymbol{r}_1, \boldsymbol{r}_2)\end{aligned} \tag{2.253}$$

where the first two terms of Equations 2.252 and 2.253 represent the unperturbed one- and two-particle electron densities for each of the monomers; $\Delta\rho_{Pau}(\boldsymbol{r}_1)$ and $\Delta\rho_{pol}(\boldsymbol{r}_1)$ are, respectively, the Pauli and polarization deformation densities; and the last two terms of Equation 2.253, the interfragment exchange (only) and XC densities. Considering these definitions, the integrands in the second and third terms of Equation 2.248, which are associated

to the electrostatic nucleus-electron attraction and the electron-electron repulsion, can be embedded into two monomer-wise unperturbed potentials:

$$\begin{aligned}\hat{\mathcal{V}}_A(\boldsymbol{r}_1) &= \hat{\mathcal{V}}_{M_A}(\boldsymbol{r}_1) + \int d\boldsymbol{r}_2 \frac{\rho_A(\boldsymbol{r}_2)}{|\boldsymbol{r}_2 - \boldsymbol{r}_1|} \\ \hat{\mathcal{V}}_B(\boldsymbol{r}_1) &= \hat{\mathcal{V}} u_{M_B}(\boldsymbol{r}_1) + \int d\boldsymbol{r}_2 \frac{\rho_B(\boldsymbol{r}_2)}{|\boldsymbol{r}_2 - \boldsymbol{r}_1|}\end{aligned} \tag{2.254}$$

where $\rho_A(\boldsymbol{r}_2)$ and $\rho_B(\boldsymbol{r}_2)$ represent the unperturbed electron densities of monomers A and B.

Considering all the aforementioned definitions, it is possible to calculate the energy terms in Equation 2.241 from the expression of the energy in Equation 2.248 leading to the partitioning of the interaction energy:[302]

$$E_{int} = E_{elec} + E_x + E_{rep} + E_{pol} \tag{2.255}$$

whose terms are defined as:

$$\begin{aligned}E_{elec} = &\int d\boldsymbol{r}_1 \hat{\mathcal{V}}_{M_A} \rho_B(\boldsymbol{r}_1) + \int d\boldsymbol{r}_1 \hat{\mathcal{V}}_{M_B} \rho_A(\boldsymbol{r}_1) \\ &+ \int\int d\boldsymbol{r}_1 d\boldsymbol{r}_2 \frac{\rho_A(\boldsymbol{r}_1)\rho_B(\boldsymbol{r}_1)}{|\boldsymbol{r}_2 - \boldsymbol{r}_1|} + \sum_{A=1}^{M_A} \sum_{B=M_A+1}^{M} \frac{Z_A Z_B}{|\boldsymbol{R}_A - \boldsymbol{R}_B|}\end{aligned} \tag{2.256}$$

$$E_x = \frac{1}{2} \int\int d\boldsymbol{r}_1 d\boldsymbol{r}_2 \frac{\rho_{x,AB}(\boldsymbol{r}_1, \boldsymbol{r}_2)}{|\boldsymbol{r}_2 - \boldsymbol{r}_1|} \tag{2.257}$$

$$\begin{aligned}E_{rep} = &-\frac{1}{2} \int d\boldsymbol{r}_1 \nabla^2 \Delta\rho_{Pau}(\boldsymbol{r}_1, \boldsymbol{r}_1')|_{\boldsymbol{r}_1' = \boldsymbol{r}_1} \\ &+ \int d\boldsymbol{r}_1 \hat{\mathcal{V}}_A \Delta\rho_{Pau}(\boldsymbol{r}_1) + \int d\boldsymbol{r}_1 \hat{\mathcal{V}}_B \Delta\rho_{Pau}(\boldsymbol{r}_1) \\ &+ \frac{1}{2} \int\int d\boldsymbol{r}_1 d\boldsymbol{r}_2 \frac{\Delta\rho_{Pau}(\boldsymbol{r}_1)\Delta\rho_{Pau}(\boldsymbol{r}_2)}{|\boldsymbol{r}_2 - \boldsymbol{r}_1|}\end{aligned} \tag{2.258}$$

$$\begin{aligned}E_{pol} = &-\frac{1}{2} \int d\boldsymbol{r}_1 \nabla^2 \Delta\rho_{pol}(\boldsymbol{r}_1, \boldsymbol{r}_1')|_{\boldsymbol{r}_1' = \boldsymbol{r}_1} + \frac{1}{2} \int\int d\boldsymbol{r}_1 d\boldsymbol{r}_2 \frac{\Delta\rho_{xc}(\boldsymbol{r}_1, \boldsymbol{r}_2)}{|\boldsymbol{r}_2 - \boldsymbol{r}_1|} \\ &+ \int d\boldsymbol{r}_1 \hat{\mathcal{V}}_A \Delta\rho_{pol}(\boldsymbol{r}_1) + \int d\boldsymbol{r}_1 \hat{\mathcal{V}}_B \Delta\rho_{pol}(\boldsymbol{r}_1) \\ &+ \frac{1}{2} \int\int d\boldsymbol{r}_1 d\boldsymbol{r}_2 \frac{\Delta\rho_{pol}(\boldsymbol{r}_1)\Delta\rho_{pol}(\boldsymbol{r}_2)}{|\boldsymbol{r}_2 - \boldsymbol{r}_1|} \\ &+ \int\int d\boldsymbol{r}_1 d\boldsymbol{r}_2 \frac{\Delta\rho_{Pau}(\boldsymbol{r}_1)\Delta\rho_{pol}(\boldsymbol{r}_2)}{|\boldsymbol{r}_2 - \boldsymbol{r}_1|}\end{aligned} \tag{2.259}$$

Both the exchange and repulsion terms of Equations 2.257 and 2.258 appear as a consequence of the Pauli exclusion principle,[216] so that they are can be into the single term E_{Pau}:

$$E_{Pau} = E_x + E_{rep} \tag{2.260}$$

Moreover, the polarization energy not only includes the classical induction energy but also, the inherent QM dispersion energy. Within the second-order perturbation theory framework and considering the molecular wavefunction $\Psi(\boldsymbol{x}_1, ..., \boldsymbol{x}_N)$, [309] these two contributions can be exactly separated. In the case of the induction energy E_{ind}, it can be demonstrated that it is given by :

$$E_{ind} = \sum_{m\neq 0} \frac{\left[\int d\boldsymbol{r}_1 \hat{\mathcal{V}}_A \rho_B^{m0}(\boldsymbol{r}_1)\right]^2}{E_B^m - E_B^0} + \sum_{n\neq 0} \frac{\left[\int d\boldsymbol{r}_1 \hat{\mathcal{V}}_B \rho_A^{n0}(\boldsymbol{r}_1)\right]^2}{E_A^n - E_A^0} \tag{2.261}$$

where $\rho_A^{m0}(\boldsymbol{r}_1)$ and $\rho_B^{m0}(\boldsymbol{r}_1)$ are the induced transition one-electron densities within each monomer from the ground state configuration 0 to the excited state m or n and the denominators contain the corresponding energy differences.

The first-order correction to the electron density $\Delta\rho(\boldsymbol{r}_1)$, which corresponds to the deformation density and is required to obtain the second-order correction for the energy, is given, at the second-order perturbation theory, by the polarization densities of monomers A and B:[309]

$$\begin{aligned}\Delta\rho(\boldsymbol{r}_1) =& \Delta\rho_A(\boldsymbol{r}_1) + \Delta\rho_B(\boldsymbol{r}_1) \\ =& 2\sum_{n\neq 0} \frac{\int d\boldsymbol{r}_1 \hat{\mathcal{V}}_B \rho_A^{n0}(\boldsymbol{r}_1)}{E_A^n - E_A^0} \rho_A^{n0} + 2\sum_{m\neq 0} \frac{\int d\boldsymbol{r}_1 \hat{\mathcal{V}}_A \rho_B^{m0}(\boldsymbol{r}_1)}{E_B^m - E_B^0} \rho_B^{m0}\end{aligned} \tag{2.262}$$

Starting from this expression, the polarization contribution to the interaction energy in Equation 2.259 and taking into account the framework presented along this section; it is possible to obtain an expression for the induction energy E_{ind}. To do so, it is necessary to define the charge induction energy E_{ch-ind} as the sum of the third and fourth terms of Equation 2.259:

$$E_{ch-ind} = \int d\boldsymbol{r}_1 \hat{\mathcal{V}}_A \Delta\rho_{pol}(\boldsymbol{r}_1) + \int d\boldsymbol{r}_1 \hat{\mathcal{V}}_B \Delta\rho_{pol}(\boldsymbol{r}_1) \tag{2.263}$$

which, upon the introduction of Equation 2.262, becomes:

$$\begin{aligned} E_{ch-ind} = 2\sum_{m\neq 0}\frac{\left[\int d\boldsymbol{r}_1\hat{\mathcal{V}}_A\rho_B^{m0}(\boldsymbol{r}_1)\right]^2}{E_B^m - E_B^0} + 2\sum_{n\neq 0}\frac{\left[\int d\boldsymbol{r}_1\hat{\mathcal{V}}_B\rho_A^{n0}(\boldsymbol{r}_1)\right]^2}{E_A^n - E_A^0} \\ + 2\sum_{m\neq 0}\frac{\int d\boldsymbol{r}_1\hat{\mathcal{V}}_A\rho_B^{m0}(\boldsymbol{r}_1)\int d\boldsymbol{r}_1\hat{\mathcal{V}}_B\rho_B^{m0}(\boldsymbol{r}_1)}{E_B^n - E_B^0} \\ + 2\sum_{n\neq 0}\frac{\int d\boldsymbol{r}_1\hat{\mathcal{V}}_B\rho_A^{n0}(\boldsymbol{r}_1)\int d\boldsymbol{r}_1\hat{\mathcal{V}}_A\rho_A^{n0}(\boldsymbol{r}_1)}{E_A^m - E_A^0} \end{aligned} \tag{2.264}$$

From this equation, the induction energy coming from the second-order perturbation theory on the molecular wavefunction given by Equation 2.261 can be identified as half of the first two terms. Hence, introducing Equation 2.261 into this expression and solving for E_{ind}:

$$\begin{aligned} E_{ind} = \frac{1}{2}\Bigg[E_{ch-ind} - 2\sum_{m\neq 0}\frac{\int d\boldsymbol{r}_1\hat{\mathcal{V}}_A\rho_B^{m0}(\boldsymbol{r}_1)\int d\boldsymbol{r}_1\hat{\mathcal{V}}_B\rho_B^{m0}(\boldsymbol{r}_1)}{E_B^n - E_B^0} \\ - 2\sum_{n\neq 0}\frac{\int d\boldsymbol{r}_1\hat{\mathcal{V}}_B\rho_A^{n0}(\boldsymbol{r}_1)\int d\boldsymbol{r}_1\hat{\mathcal{V}}_A\rho_A^{n0}(\boldsymbol{r}_1)}{E_A^m - E_A^0}\Bigg] \end{aligned} \tag{2.265}$$

which becomes:

$$E_{ind} = \frac{1}{2}\left[E_{ch-ind} - \int d\boldsymbol{r}_1\hat{\mathcal{V}}_A\Delta\rho_A(\boldsymbol{r}_1) - \int d\boldsymbol{r}_1\hat{\mathcal{V}}_B\Delta\rho_B(\boldsymbol{r}_1)\right] \tag{2.266}$$

upon the introduction of the expression for the first-order correction to the electron density in Equation 2.262 in the last two terms of Equation 2.265. Finally, applying the definition of the charge induction energy in Equation 2.264, the induction energy can be expressed as:

$$E_{ind} = \frac{1}{2}\left[\int d\boldsymbol{r}_1\hat{\mathcal{V}}_A\Delta\rho_B(\boldsymbol{r}_1) + \int d\boldsymbol{r}_1\hat{\mathcal{V}}_B\Delta\rho_A(\boldsymbol{r}_1)\right] \tag{2.267}$$

which is, in fact, consistent with the classical result for the induction energy.[303] Considering all the aforementioned, the interaction energy can be thus decomposed as a sum of electrostatic E_{elec}, Pauli E_{Pau}, inductive E_{ind} and dispersive $E_{disp+respol}$ terms:

$$E_{int} = E_{elec} + E_{Pau} + E_{ind} + E_{disp+respol} \tag{2.268}$$

where the second-order dispersion energy represents the highest contribution to the $E_{disp+respol}$.

The EDA scheme presented above can be extended to the QM/MM framework[310] where monomers A and B consist of a substrate and (part of) the environment which, along this thesis, will be the OLU chromophore and the luciferase protein, respectively. In this case, the environment would present atoms represented either at QM level or as MM fixed point charges, and the interactions between the MM region and monomer A are introduced in the calculation of the different energy terms in Equations 2.256-2.259 by adding the $\mathcal{V}_{FF}$ potential generated by these point charges to the potential generated by the nuclei of monomer B:

$$\hat{\mathcal{V}}_{M_B}^{QM/MM}(\boldsymbol{r}) = \hat{\mathcal{V}}_{M_B}(\boldsymbol{r}) + \hat{\mathcal{V}}_{FF} \tag{2.269}$$

Furthermore, this $\hat{\mathcal{V}}_{FF}$ potential is taken into account for the calculation of the internuclear interaction of the electrostatic energy, which corresponds to the last term of Equation 2.256.

Chapter III

Electronic Properties of the Phenolate-Keto Oxyluciferin inside the Wild Type Luciferase

The electronic characterization of the luciferin/luciferase complex is fundamental to tune its photophysical properties and develop more efficient devices based on this luminescent system. In this chapter, MD simulations, hybrid QM/MM calculations and transition density analysis are employed to compute the absorption and emission spectra of the phenolate-keto OLU/luciferase complex. These results are analysed to unravel the nature of the relevant electronic state, the S_1, and its behaviour with the intramolecular and intermolecular DOF. It is found that the torsional motion of the chromophore is hampered by the presence of the enzyme, reducing the intramolecular CT nature of the absorbing and emitting state. In addition, such a reduced CT character does not correlate in a strong way neither with the intramolecular motion of the chromophore nor with the chromophore/amino-acid distances. However, the presence of a polar environment around the oxygen atom of the tz ring of the OLU, coming from both the protein and the solvent, enhances the CT character of the emitting state.

* This chapter is based on the released article *Phys. Chem. Chem. Phys.* **2023**, 25, 16507-16519. DOI: 10.1039/D3CP01387A with contributions of Henar Mateo-delaFuente,[a] Davide Avagliano,[b] Marco Garavelli,[b] and Juan J. Nogueira.[a,c]

[a] Chemistry Department, Universidad Autónoma de Madrid, 28049 Madrid, Spain.

[b] Dipartimento di Chimica industriale Toso Montanari, Università di Bologna, 40136 Bologna, Italy.

[c] Institute for Advanced Research in Chemical Sciences (IAdChem), Universidad Autónoma de Madrid, 28049 Madrid, Spain.

Preprint: 10.26434/chemrxiv-2023-88slr

3.1 Introduction

As commented before in Section 1.2.3 and despite the extensive research on the topic,[2, 31, 119, 131, 167] many of the factors that affect the emission mechanism and the nature of the emitting electronic state of the OLU chromophore are still unclear. Recent computational work has characterized the emitting state as a partial intramolecular CT character state.[311] In addition, previous studies[124, 312] have pointed out the relevance of the torsion of the molecule around the SCCS dihedral angle (see Scheme 1.9) on the electronic properties of the chromophore and the possible generation of a twisted intramolecular CT state with a red-shifted emission. Although, computational research discarded this TICT state mechanism,[154, 155] these works were performed without taking into account neither the environment nor the dynamic effects. Therefore, and even though the consensus is that the molecule emits in a planar conformation,[119] it is fundamental to characterise the electronic properties with respect to the torsion in the biological media in order to understand the emission photophysics of the chromophore. In the present chapter, the absorption and emission spectrum of the phenolate-keto OLU/luciferase complex are computationally investigated by means of MD simulations and hybrid QM/MM calculations. The CT character of the absorbing and emitting excited state is quantitatively investigated by analysing the transition density, which is not basis set dependent and, thus, more robust than previous analysis based on, for example, Mulliken population analysis.[155] In addition, the effect of the intramolecular torsion and possible micro-environmental effect due to the presence of the protein and the solvent on the CT state is discussed. It is found that the WT enzyme restricts the intramolecular motion of the chromophore, avoiding the formation of an electronic state with strong CT character.

3.2 Computational Details

3.2.1 Static Potential Energy Rigid Scan

The OLU species emit light from an electronic state which has been classified by computational methods as a state with partial intramolecular CT character.[311] The torsional motion around the SCCS dihedral angle in Scheme 1.9, which leads to a breakdown of the planarity of the molecule, could be relevant in the photophysics since the distortion of the geometry could significantly affect the electronic delocalization of the chromophore. Therefore, the effect of this intramolecular degree of freedom on the electronic structure of the

phenolate-keto OLU is analysed here in detail. As a first approach, a rigid scan around the SCCS dihedral angle was carried out for OLU *in vacuum*, along which the energies of the S_0 and S_1 states were computed at the TD-DFT level using different density functionals and the 6-311G(2d,p)[313–315] basis set. Specifically, the B3LYP,[231, 232, 316] CAM-B3LYP,[233] M062X,[317] PBE0[318] and ωB97XD[319] functionals were employed to evaluate their performance. For these calculations the Gaussian 16[320] software was used. In addition, the transition density matrix was analysed by the TheoDORE software[249] to quantify the CT character of the S_1 state along the torsion. The use of this methodology complements the analysis performed on previous works,[155] where the CT character was discussed based on Mulliken population analysis. These calculations for a simplified model were then employed to rationalise the results obtained for the more complex situation, where the chromophore is embedded in the protein environment and conformational sampling is introduced by MD simulations, as explained below.

3.2.2 Molecular Dynamics Simulations

For this chapter, MD simulations of the phenolate-keto OLU/luciferase complex were carried out, as graphically represented in Figure 3.1, considering two different approaches to compute the potential energy along the dynamics on the S_0 and S_1 states. In the simplest one, the interatomic interactions were computed by means of a FF while, in the more accurate approach some of the conformations sampled from the classical trajectory were selected as initial conditions to evolve hybrid QM/MM MD simulations to obtain a better description of the chromophore structure, which is the only part of the system included in the QM region. For each simulation, a few snapshots were selected and SP QM/MM calculations were performed on top of them to compute the absorption and emission spectra and characterize the electronic states involved. On the flip side, and despite on the ongoing debate about the emitting species of the system,[119] it is in general a good approximation to use the OLU mono-anion in its keto form as chromophore in the model since it is the direct product of the BL reaction, after the ESPT takes place. Furthermore, not only this chromophore has been found to be one of the emitting species but also it is considered to be the red-emitting specie.[321–323]

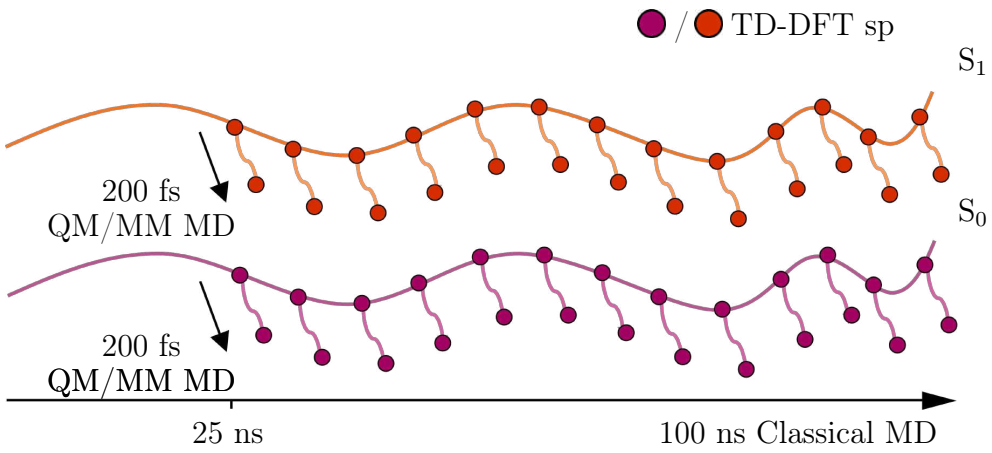


Figure 3.1: Scheme of the computational protocol followed in Chapter III.

System Setup

The OLU/luciferase complex was built with PyMol[324] from the X-ray diffraction structure file of the Japanese firefly, *Luciola cruciata*, luciferase published in the Protein Data Bank (PDB) (PDB ID: 2D1R[124]). The AMP residue was removed and the system was solvated using the *tleap* module of AmberTools20[260] with TIP3P[263] water molecules within a truncated octahedral box, ensuring a solvent shell of at least 12 Å from any solute molecule. This solvation process led to the inclusion of three and one initial water molecules inside the cavity of the protein for the S_0 and S_1 classical trajectories, although during the dynamics more water molecules diffused inside it. Moreover, no extra ions were required since the charge of the protein compensates for the negative charge of the phenolate-keto OLU. The enzyme was described using the ff19SB[262] FF, whereas the GAFF[261] was used to describe the dihedral and improper torsions and Lennard-Jones parameters of the chromophore selected by means of the Antechamber *parmchk2* module of the AmberTools20. The rest of the bonded parameters for OLU, that is, the bond and bond angle parameters for both the S_0 and S_1 electronic states, were obtained from QM calculations. Equilibrium distances and angles were obtained from a geometry optimization whereas force constants were calculated by means of the Seminario method,[265] described in Section 2.2.3. Specifically, the optimized geometries for S_0 and S_1 and the Hessian matrices were obtained using the Gaussian16[320] software by means of DFT and TD-DFT *in vacuo* calculations, respectively, where the B3LYP[231, 232, 316]/6-311G(2d,p)[313–315] level of theory was employed for both the S_0 and S_1 minimum-energy geome-

tries. In the case of the S_1 optimization, the TD-DFT calculation took into account 10 roots. Moreover, the electrostatic charges were computed by using the Merz–Singh–Kollman (MSK) scheme[325] at the B3LYP/6-311G(2d,p) level of theory, according to previous studies.[311, 326, 327] While it is true that some parametrization effort has been previously done for an OLU analogue,[328] it was only carried out for the electronic ground state. However, the parameters used in this chapter, which are listed in Tables B.1-B.8, not only aim at describing the electronic ground state but also the first excited state. Additionally, another system was built from the same PDB file in which the AMP monoanion was kept in other to study if its presence affects the emission spectrum of the OLU chromophore. In this case, the previously discussed S_1 parameters for the chromophore were used, while the AMP was described by the GAFF together with the MSK electrostatic charges computed at the B3LYP/6-31G*[231, 232, 313–316] level of theory.

Classical Molecular Dynamics

After setting up the solvated OLU/luciferase complex, classical MD simulations were performed for both the S_0 and S_1 states of the chromophore. The same computational protocol was used for the two simulations. First, the system was minimized for 5000 steps by means of the steepest descent algorithm,[244] and for additional 5000 steps using the conjugate gradient algorithm.[245] Then, it was progressively heated for 500 ps from 0 to 300 K in the NVT ensemble using a 2 fs timestep. An additional simulation was evolved at 300 K for another 500 ps in the same ensemble to reach an stable state before the production phase. The temperature was controlled by means of the Langevin thermostat[267, 268] with a 2.0 ps^{-1} of collision frequency. Afterwards, a production of 100 ns with a 2 fs time step and within the NPT ensemble was performed. The temperature was also controlled by means of the Langevin thermostat with a 1.0 ps^{-1} of collision frequency, while pressure was set to be 1.0 bar by means of the Monte Carlo barostat.[270, 277] The particle-mesh Ewald[272] method was used with a grid spacing of 1.0 Å, and the non-bonded interactions were computed with a cutoff of 9 Å and a switching distance of 7 Å. The bonds involving hydrogen atoms were constrained using the SHAKE[274] algorithm. Moreover, this simulation was extended to 500 ns in order to see whether the chromophore is able to rotate inside the cavity of the protein or not and therefore, evaluate the convergence of the probability distributions of the SCCS dihedral angle during the classical MD simulations. In the case of the S_1 system with AMP inside the cavity, after the minimization and thermalization processes, 200 ns of production were run.

These simulations were carried out using the previously described computational protocol.

From each of the 100 ns simulations in the S_0 and S_1 electronic states, 200 time-equidistant snapshots were extracted from the last 75 ns in order to prepare the Gaussian09 QM/MM input files for the calculation of the absorption and emission spectra. The MoBioTools software[329] was used to automatize the preparation of the QM/MM input files, which considered 10 excited states (for the absorption spectrum) at the TD-B3LYP/6-311G(2d,p) level of theory. In the case of the emission spectrum, only the lowest-energy electronic transition was considered since the BL occurs from the S_1 state. The vertical energies were convoluted with Gaussian functions of full width at half maximum of 0.1 eV to obtain the absorption and emission spectra. For the emission spectrum computed from the 200 ns S_1 trajectory of the system with AMP inside the cavity, other 200 time-equidistant snapshots were extracted from the last 100 ns in order to prepare the Gaussian09 input files following the same procedure described before.

Hybrid QM/MM Molecular Dynamics

From the classical S_0 and S_1 trajectories 100 time-equidistant geometries, together with their respective velocities, were extracted to be used as initial conditions for 100 QM/MM MD simulations in each electronic state. Ideally, one long QM/MM MD simulation should be performed for each of the electronic states, from which the snapshots could be selected to compute the spectra. However, running such long simulations at QM/MM level is not computationally feasible. Instead, an accurate conformational sampling of the chromophore is obtained by propagating short QM/MM trajectories, starting from geometries extracted from the classical MD, well separated in time (every ns), in order to also guarantee a statically-accurate thermal distribution of the solvent molecules interacting with the chromophore. Thus, these QM/MM simulations can be compared with the FF based simulations to evaluate the performance of the GAFF torsion parameters against the QM/MM potential. Each of these snapshots were processed with the recently developed tools included in the COBRAMM 2 software[330, 331] in order to define three different regions for the simulation: (i) the high layer, composed by the chromophore and treated with the CAM-B3LYP[231–233, 316]/6-31+G(d)[313–315, 332, 333] level of theory (with 3 roots in the case of the S_1 simulation); (ii) the medium layer, formed by the protein and the water molecules within 4 Å from the chromophore treated classically; and (iii) the low layer, including the rest of the water molecules and frozen during the simulation. All the

QM/MM dynamics were run with COBRAMM 2,[330, 331] using its interfaces with Gaussian16 and AMBER20 suites to calculate QM and MM energies and gradients and the internal routines to build QM/MM potentials and gradients within the electrostatic embedding scheme, and to propagate the adiabatic dynamics on S_0 and S_1 with the velocity-Verlet algorithm.[253]

The protein and the water molecules were described, respectively, by the previously used ff19SB and TIP3P FFs. The 200 trajectories (100 in the S_0 and 100 in the S_1 states) were run for 200 fs in the microcanonical ensemble (NVE) without periodic boundary conditions, which are not available in COBRAMM 2.0. For this reason, the atoms of the low layer were frozen during the QM/MM MD simulations. The time step was set to 1 fs, the cutoff for the non-bonded interactions was 10 Å, and positions and velocities of the atoms involved in X–H bonds in the medium layer were constrained with the RATTLE algorithm[275] to get a good conservation of the total energy.[334] The last step of these trajectories in the S_0 and S_1 states was used to compute the absorption and emission spectra, respectively. The preparation of the Gaussian09 input files was again performed by means of the MoBioTools software.[329]

3.2.3 Analysis of the Electronic Properties

The electronically excited states, computed by QM/MM based on the snapshots of the S_0 and S_1 states taken from the classical and QM/MM MD trajectories, were analysed by means of the TheoDORE software.[249] Specifically, the intramolecular CT number, computed from the one-electron transition density matrix between the ground state and the S_1 state as explained in Section 2.1.8, was calculated by dividing the chromophore into two fragments: the tz and btz moieties, represented in Scheme 1.9. In addition, the absorption and emission spectra were computed by considering snapshots for which the SCCS dihedral angle lie within a particular range. Moreover, the variation of the CT number with the dihedral angle was analysed and compared with the results obtained from the static scan of OLU *in vacuum* to assess the effect of the protein environment.

The QM/MM trajectory on the S_0 surface was also employed to compute the binding free energy of OLU to the enzyme by means of the 1A-MM GBSA approach[287, 288] by using the *MMPBSA.py* tool.[335] In addition, the binding free energy was decomposed into amino acid contributions by selecting only the residues located within a sphere of 5.0 Å radius from the OLU chromophore for at least 1% of the simulation time, following a similar procedure employed in a previous work.[336] The residue ARG339 was also included in the analysis since it presents a positive charge and, therefore, can interact

strongly with OLU. Moreover, a visual inspection of the dynamics showed that ARG339 was close to the chromophore in some of the snapshots with large CT values. A total of 16 residues were considered in the per-residue decomposition of the free binding energy. In addition, this analysis was also performed for the S_1 classical trajectories without and with the AMP contained in the cavity.

Finally, the location of the OLU inside the cavity for the QM/MM trajectory on the S_0 surface was analysed by a RMSD and divided accordingly in 3 clusters using the K-means clustering algorithm[337] randomizing the initial set of points implemented in the *cpptraj* module of AmberTools20.[260]

3.3 Results and Discussion

This chapter focuses on the influence of intramolecular and intermolecular DOFs on the electronic properties of the chromophore in order to investigate both the TICT and the micro-environment mechanisms for OLU colour tuning by means of computational methods. Specifically, the variation of the excitation energies, absorption and emission intensities, and CT character with the SCCS dihedral angle are analysed. This is done by comparing the results obtained from the classical and QM/MM MD trajectories with the static potential energy scan. Regarding the impact of the intermolecular DOFs, the protein residues that are more relevant in the phenolate-keto OLU/enzyme interaction are identified and, then, a correlation between the distance between those residues and the chromophore with the CT character of the emitting state is evaluated. In the following sections, first the results from the static scan are presented, followed by the discussion based on the analyses of the MD simulations.

3.3.1 Static Scan around the C–C' Bond

The energy change of the phenolate-keto OLU anion with the SCCS torsion have been computed by means of TD-DFT calculations using five different density functionals, namely, B3LYP, CAM-B3LYP, M06-2X, PBE0 and ωB97XD, and the 6-311G(2d,p) basis set. The energy scans shown in Figure 3.2 show that the S_0 surface presents two minima in the two planar situations (0 and 180º respectively) and two maxima when the btz and tz fragments are perpendicular to one another, at 90 and 270º respectively. On the flip side, the S_1 surface presents two additional minima, corresponding to the maxima of the ground state, and four maxima in total at 40, 130, 230 and 320º. Even though all the density functionals present a similar behaviour, B3LYP and PBE0 show

an abnormally pronounced energy decrease in the S_1 minima corresponding to the perpendicular fragments situations. This could be due to the fact that these functionals do not include long-range corrections, thus leading to a wrong description of CT states.[338] The other three density functionals, which do

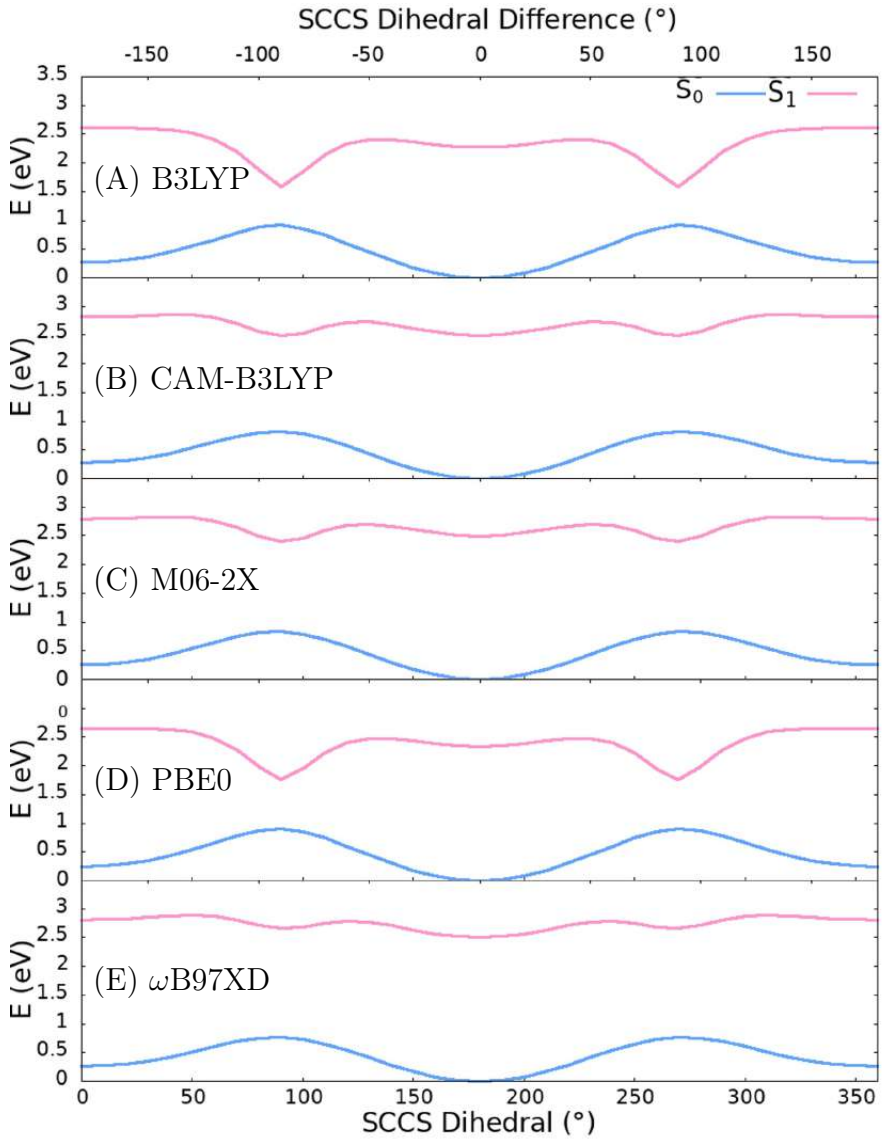


Figure 3.2: S_0 (blue) and the S_1 (pink) rigid potential energy curves calculated with the density functionals: (A) B3LYP, (B) CAM-B3LYP, (C) M06-2X, (D) PBE0 and (E) ωB97XD. The secondary x-axis shows the angle difference with respect to the planar energy minimum found at a SCCS angle of 180º.

include long-range corrections, present a smoother shape of the S_1 potential curve around the energy minima. Previous high-level wavefunction-based calculations also showed such a smooth shape of the potential.[155] Therefore, one can conclude that a long-range corrected functional is needed to properly describe the shape of the potential energy curve. Thus, CAM-B3LYP was the functional selected to carry out the QM/MM MD trajectories and it is the scan that will be discussed below.

In addition to the energies of the S_0 and S_1 states along the scan, the oscillator strength and the CT character of the S_0 - S_1 transition have been computed. These results are shown in Figure 3.3(A) at the CAM-B3LYP/6-311G(2d,p) level although, for the sake of completeness, an equivalent representation of Figure 3.2 but for the variation of the oscillator strength and CT number can be found in Figure B.2. The oscillator strength (f) presents two maxima and two minima corresponding, respectively, to the planar and perpendicular geometrical situations. This means that a break of the planarity of the chromophore reduces the probability of the transition to take place. This correlates well with the CT number, which increases its value when the planarity of the molecule is broken, reaching its maximum value at 90 and 270^{o}. Both results can be rationalized by looking at the molecular orbitals involved in the electronic transition in the planar and perpendicular cases (see Figure 3.3(B)). When both fragments are in the same plane (SCCS angle of 180^{o}), the Highest Occupied Molecular Orbital (HOMO) and Lowest Unoccupied Molecular Orbital (LUMO) orbitals are delocalized on the two fragments, which favours the overlap of the initial and final electronic densities and, thus, increases the transition dipole moment and the oscillator strength. Moreover, since both orbitals are delocalized over the whole molecule, the CT from one fragment to the other is not important and the CT number is small. On the flip side, when the fragments are perpendicular to each other, the HOMO is completely localized in the btz fragment, while the LUMO is localized in the tz one. Therefore, the overlap between the initial and final electronic densities vanishes and the oscillator strength becomes very small. Furthermore, the CT increases since the initial and final positions of the electron are localised in different fragments.

In summary, the potential energy scan suggests a red shift of the absorption and emission energies when the chromophore leaves the planarity during the dynamics since the energy of the ground state increases, while the energy of the first excited state is nearly constant around both the ground and excited-state minima. In addition to the red shift, the torsional motion of OLU around the C–C' single bond should also induce a decrease of the intensity of the spectra and an increase of the CT character of the S_1 state. In the following, the MD

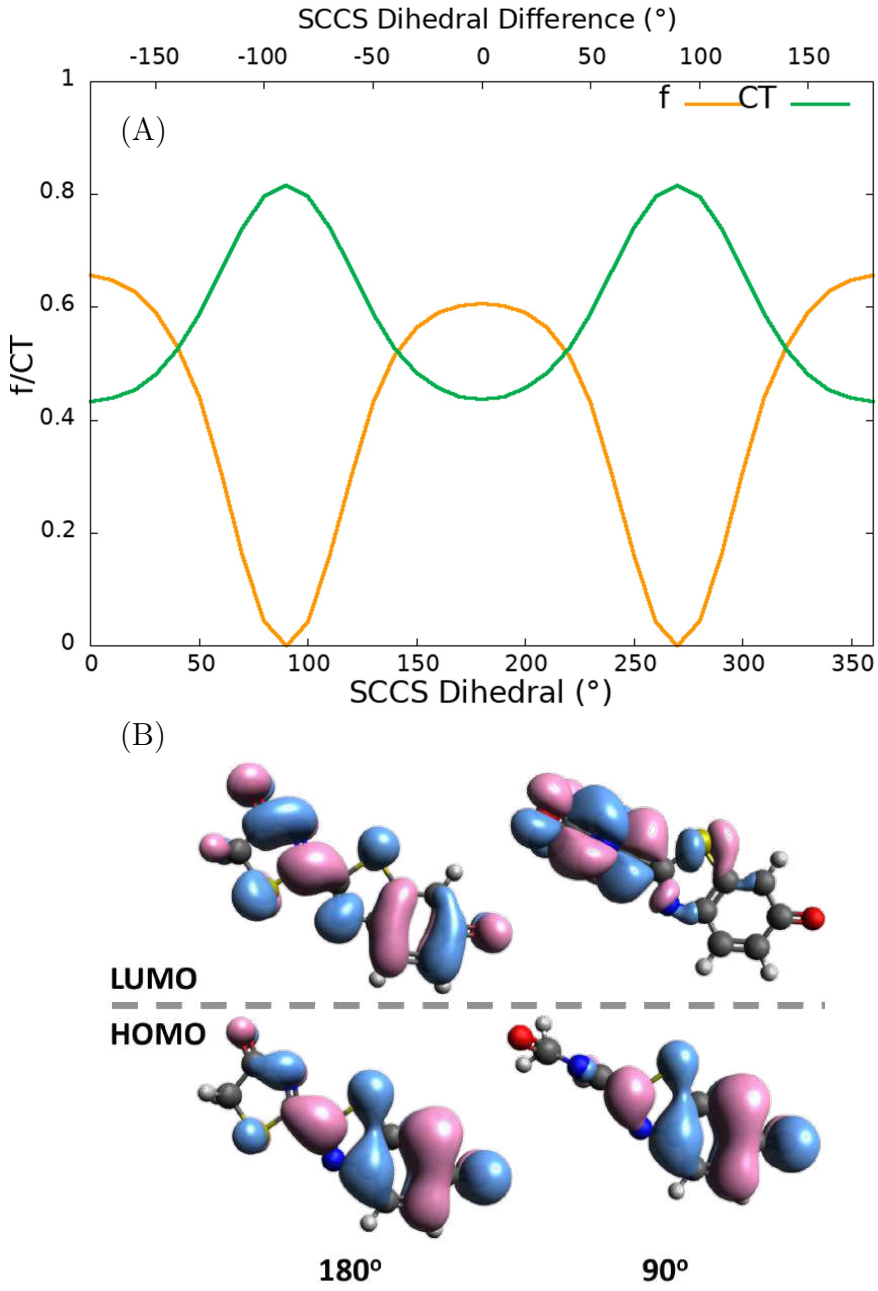


Figure 3.3: (A) Variation of the oscillator strength (orange) and CT number (green) for the S_0 - S_1 electronic transition along the dihedral angle calculated with the CAM-B3LYP density functional. The secondary x-axis shows the angle difference with respect to the energy minimum. (B) HOMO and LUMO orbitals for the planar and perpendicular geometrical situations.

simulations will be analysed to see whether the expected trend predicted by the static simple model is satisfied or not.

3.3.2 Absorption and Emission Spectra from the Classical Trajectories

The QM/MM vertical excitation energies of the 200 equidistant snapshots taken from the S_0 and S_1 classical trajectories were convoluted to obtain the absorption and emission spectra, represented in Figure 3.4. In the case of the absorption band, two peaks can be observed at 2.34 eV (530 nm) and 3.08 eV (403 nm), which correspond with the S_0 - S_1 and S_0 - S_3 transitions, respectively. The emission peak was found at 2.16 eV (574 nm), in very good agreement with the experimental value (560 nm).[124] Therefore, the methodology consisting of classical MD followed by QM/MM vertical energy calculations provides accurate results in terms of energy. Moreover, as it can be observed in Figure 3.4(B), the inclusion of AMP in the model induces a slight blue shift of the emission spectrum of 0.11 eV, which is within the DFT error. This proves that the inclusion of the AMP inside the protein does not affect much the electronic properties of the system.

The S_0 - S_1 contribution of the absorption spectra and emission spectra were decomposed into bands where the chromophore presents different dihedral angles. In particular, the bands have been analysed by classifying the 200 sampled snapshots in 10 groups of 20 geometries according to the dihedral angle difference with respect to planarity (180^{o}). Each group of snapshots contains 20 geometries to ensure that all the divisions are equally statistically meaningful, since the population of planar geometries is larger than the ones with distorted angles. However, since the 140 geometries corresponding to the dihedral angles difference between 1.6 and 21.3^{o}, in the case of the S_0 trajectory, and between 1.9 and 15.7^{o}, in the case of the S_1 trajectory, resulted in very similar bands, they have been represented together taking into account the number of geometries so that the relative intensity of each group is comparable. The same figure without combining geometries is presented in Figure B.3.

As it can be observed in Figure 3.4, there is a very small red shift for the absorption and emission bands when the chromophore goes from the planar to the twisted configuration. In the case of the absorption spectrum, the planar contribution of the band peaks at 2.31 eV, while the two non-planar contributions peak at 2.24 and 2.23 eV, respectively. In the case of emission, the planar contribution of the band peaks at 2.17 eV, while the two contributions from the most rotated dihedral angle peak at 2.14 and 2.16 eV. According to the static scan for the chromophore *in vacuum* (Figure 3.3), the red shift is

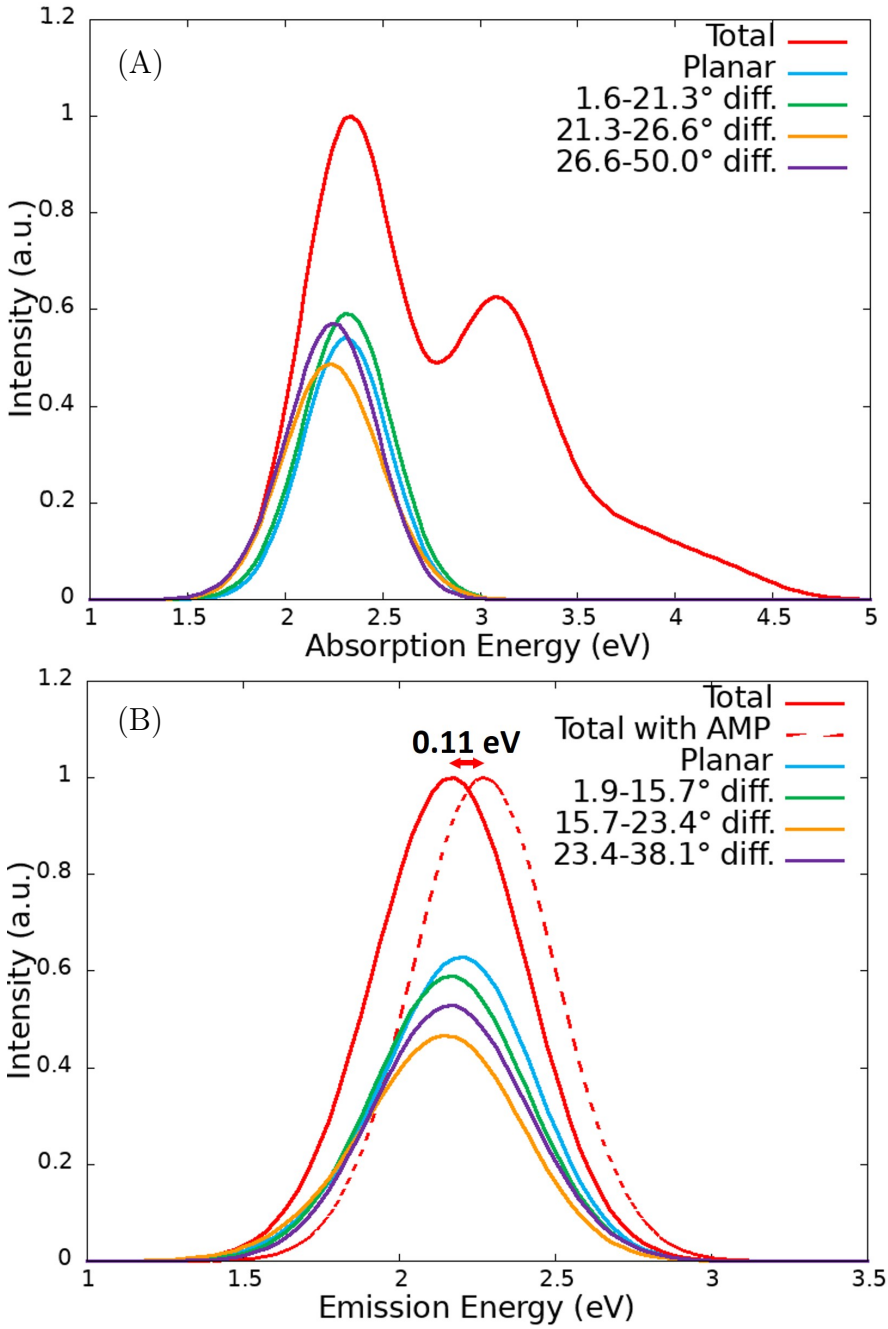


Figure 3.4: (A) Absorption and (B) emission spectra computed from the classical trajectories, together with their respective contributions from the S_0 - S_1 band depending on the SCCS dihedral angle. Additionally, the emission spectra obtained from the classical MD with the AMP inside the cavity is shown with a dashed line.

important only when the deviation from planarity is large (more than 50^{o}), reaching values for the SCCS angle smaller than 130^{o} and larger than 230^{o}. Nonetheless, such a large deviations are not reached during the classical MD trajectories; as it can be observed in Figure 3.5(A), which shows the probability distribution of the dihedral angle for the classical trajectories.

Along these trajectories, the population of dihedral angle deviations from planarity larger than 50^{o}, for both the S_0 and S_1 trajectories, is insignificant. Moreover, this small deviations from planarity not only happen for the 100 ns classical simulations but also for the 500 ns trajectories as it can be observed in Figure B.4. This is the reason behind the very small red shift observed in the spectra. Moreover, the distribution of the classical trajectory on the ground state (with a Full Width at Half Maxima (FWHM) of 36^{o}) is wider than that of the excited state (with a FWHM of 29^{o}). This difference between both states arises from the stretching and bending FF parameters and the atomic charges since the torsional and non-bonding parameters are the same for both simulations, as they were directly taken from the GAFF. These values of FWHM do not coincide with the expected behaviour attending to the static scan represented in Figure 3.2 since the energy barrier of the S_1 is lower than that of the S_0. This could indicate that the FF employed to describe the motion of the chromophore is not adequate and a more accurate description could be needed. For this reason, the quality of the analysis has been improved by performing QM/MM simulations in the S_0 and S_1 states of OLU.

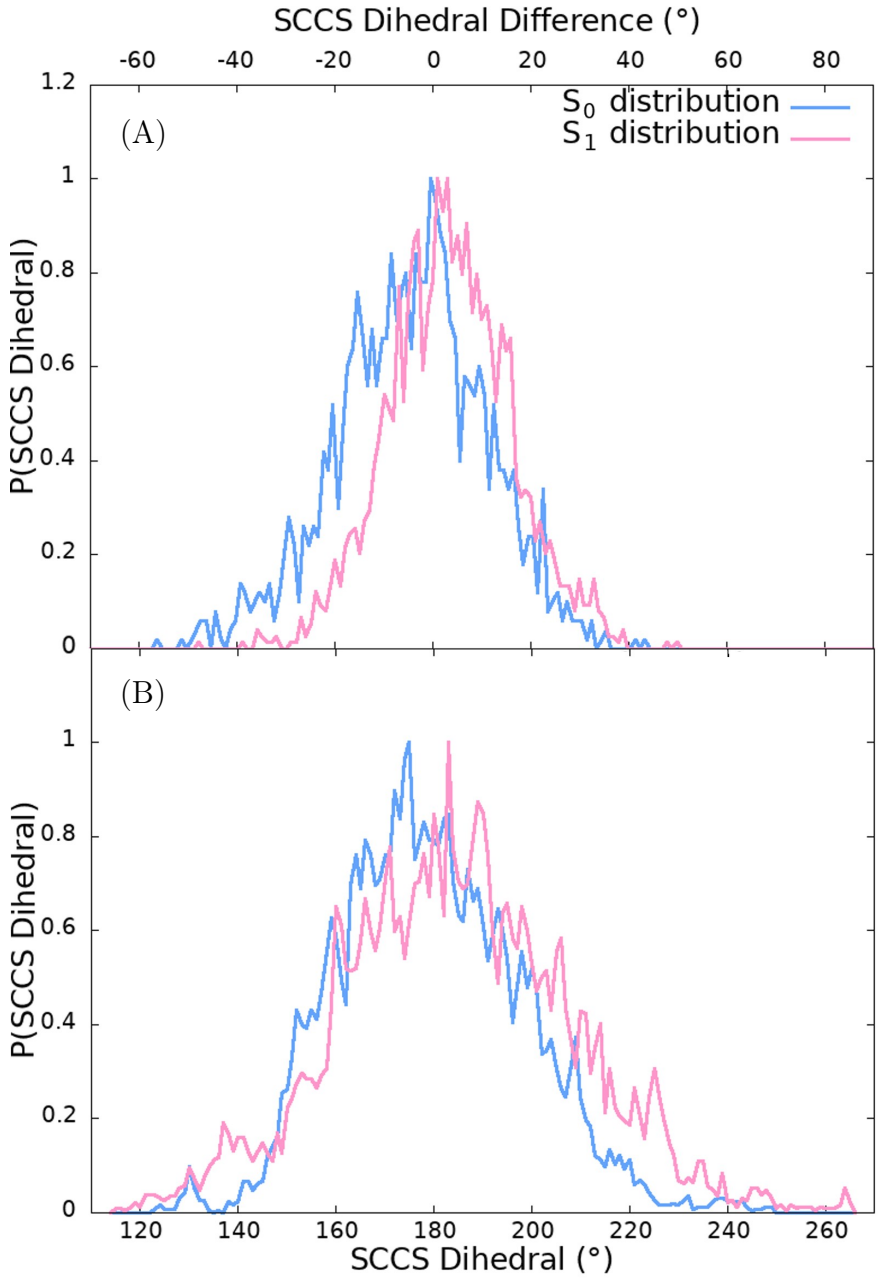


Figure 3.5: Probability distributions of the SCCS dihedral angles for the S_0 (blue) and the S_1 (pink) trajectories evolved at (A) classical and (B) QM/MM level. The secondary x-axis shows the angle difference with respect to the energy minimum.

3.3.3 Absorption and Emission Spectra from the QM/MM Trajectories

The absorption and emission spectra were obtained from the QM/MM vertical energies of the last snapshot of the 100 QM/MM trajectories evolved on the S_0 and S_1 surfaces, respectively. As it can be observed in Figure 3.6(A), the absorption spectrum presents two peaks at 2.48 eV (500 nm) and 3.32 eV (373 nm) corresponding again to the S_0 - S_1 and S_0 - S_3 electronic transitions, respectively. These energies are blue shifted with respect to those obtained from the classical trajectory (2.34 and 3.08 eV). Moreover, the relative intensity of the second peak is lower than that of the spectrum computed from the classical trajectory. In the case of the emission spectrum of Figure 3.6(B), the peak appears at 2.17 eV (571 nm), 3 nm above the value of the spectrum from the classical trajectory and 11 nm above the experimental value (560 nm). Therefore, once more, the agreement with the experiment is very good.

The effect of the torsion around the SCCS dihedral angle has been, once again, analysed by decomposing the S_0 - S_1 band of the absorption spectrum and the emission spectrum by dividing the 100 sampled geometries in 5 groups according to the dihedral angle difference with respect to the planar case of 180^{o} (each group containing 20 geometries). In this case, only the 60 geometries corresponding to the dihedral angle difference between 4.5 and 23.5^{o}, in the case of the S_0 trajectories, and the 40 geometries that present dihedral angles difference between 5.5 and 20.5^{o}, in the case of the S_1 trajectories, have been represented together since they present very similar bands (again each contribution has been scaled considering the number of geometries so that the intensities of each group is comparable). As it can be observed in Figure 3.6(A), the absorption band is red shifted when the chromophore is distorted from the planar configuration while for the emission band, the red shift only appears for very distorted configurations. In fact, small distortions of 5 to 21^{o} induce a blue shift, probably due to the effect of other factors. In addition, the torsional motion also induces a decrease in the absorption and emission intensity (hypochromism), an effect that is expected looking at the behaviour of the oscillator strength along the static scan.

The spectra shown in Figure 3.6 suggest that the OLU/luciferase complex follows the expected trend. Upon strong distortion of the dihedral angle, the energy is red shifted and the intensity is reduced, which can be rationalised when looking at the dihedral angle distribution obtained from the QM/MM dynamics displayed in Figure 3.5(B). The distributions for both ground and excited states (with a FWHM of 43 and 49^{o} for the S_0 and S_1, respectively) are broader than those of the classical trajectories, reaching values that differ from

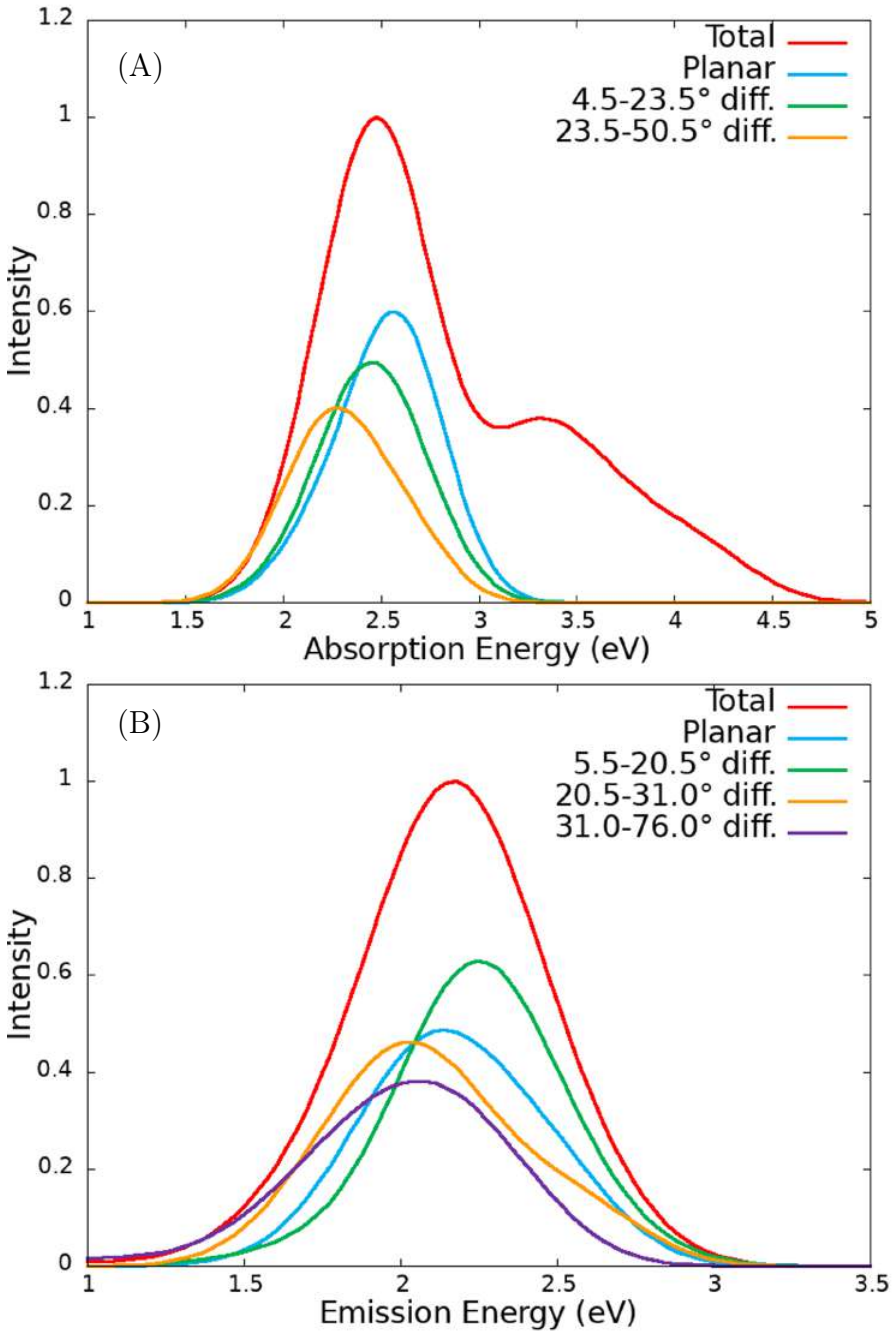


Figure 3.6: (A) Absorption and (B) emission spectra computed from the QM/MM trajectories, together with their respective contributions from the S_0 - S_1 band depending on the SCCS dihedral angle.

the planar geometry by more than 50º. According to the static scan (Figure 3.2), this is the necessary angle difference to have a significant red shift and hypochromism in the spectra. Moreover, the width of the S_1 distribution is slightly larger than that of the S_0 distribution, in consonance with the shape of the potential energy curves shown in Figure 3.2.

Summarizing, although the QM/MM vertical energy calculations based on both classical and QM/MM trajectories provide emission spectra that compare well with the experimental data, it seems that the classical trajectories underestimate the vibrational torsion of OLU, resulting in narrow angle distributions, as reported by Cerezo et al.[328] Therefore, even if energetically accurate, the classical MD protocol alone would not produce an accurate enough conformational distribution of the chromophore in the biological environment, which can be then reached once the molecule is allowed to move in a QM/MM potential.

3.3.4 Effect of the Intramolecular and Intermolecular Degrees of Freedom on the CT Character

In the previous section it has been shown that the torsional motion of the chromophore, when embedded in the protein, is large enough to induce a red shift of the absorption and emission spectra with respect to the planar geometry. Hence, here it is analysed how the CT number changes with respect to the variation of the dihedral angle. According to Figure 3.3, the distortion from the planarity should increase the CT character of the bright state. The snapshots extracted from the trajectories were grouped in 10 (classical trajectory) and in 5 (QM/MM trajectory) ensembles according to the dihedral angle difference with respect to planarity (180º), such that each ensemble contains 20 geometries. Then, the average CT number and their corresponding standard deviation for each ensemble was computed and the result is displayed in Figure 3.7. As it can be observed, the average CT number does not present significant oscillations when the SCCS angle difference increases, not even for the emission spectrum at QM/MM level for which angle differences larger than 50º are reached during the dynamics. Such a constant behaviour of the CT number can be due the combination of two factors. First, the torsional motion is not large enough to have a significant increase of the CT character. As shown in Figure 3.3(A), although a distortion of 50º already enhances the CT character of the excited state, such a large distortion is achieved only for a few snapshots in the QM/MM trajectory (Figure 3.5(B)). Second, even for these largely distorted geometries there might exist some other DOF that counteract the effect of having a large torsional angle. It is important to notice in Figure

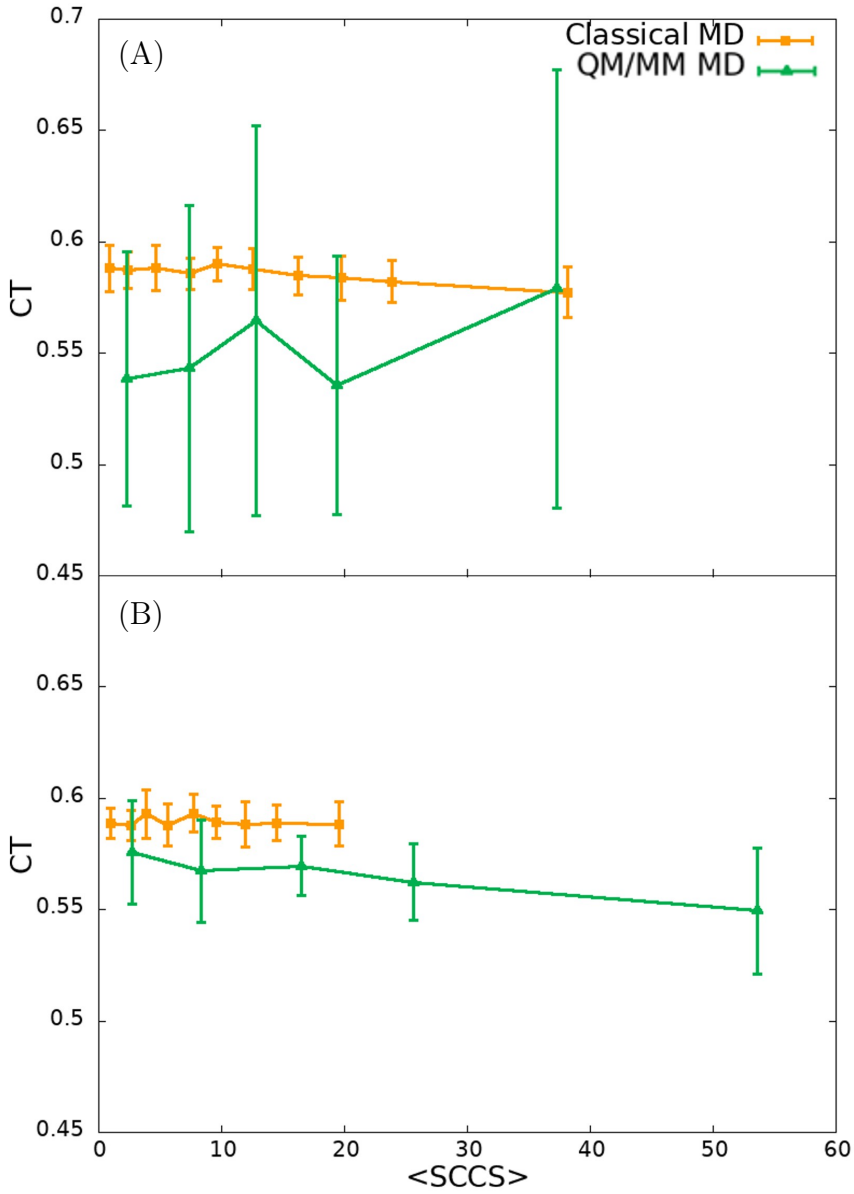


Figure 3.7: Variation of the CT number with the the SCCS dihedral angle difference from planarity for the (A) absorption and (B) emission spectra computed from snapshots taken from the classical (orange) and QM/MM (green) trajectories.

3.7 that the standard deviation of the CT number is larger for the QM/MM trajectory than for the classical one because the dihedral angle intervals employed to have the same number of snapshots in each interval is wider for the QM/MM analysis than for the classical one. This larger angle interval contains larger molecular distortions and, therefore, stronger CT number oscillations.

The relevance of other intramolecular DOF was analysed by a structural analysis. Specifically, the 20 geometries of the chromophore with both the smallest and largest CT number values were separately gathered in clusters, for which the centroids (representative structures) were obtained. These geometrical configurations were extracted from the S_0 QM/MM dynamics, which is the one that presents the most important CT oscillations as shown in Figure 3.7. The centroid geometries for each of the clusters are represented without the enzymatic environment in Figure 3.8(A) and with it in Figure B.5, and the ensemble of all the 20 geometries per cluster in Figure 3.8(B and C). As it can be seen, no significant differences are observed between the representative structures with the smallest and largest CT number. In fact, the RMSD between the geometries of each cluster and its centroid (0.41 ± 0.16 Å and 0.41 ± 0.17 Å for the cluster with the smallest and largest CT values, respectively) is larger than the RMSD between both centroid structures (0.33 Å), which means that the conformational change within each cluster is larger than the variations between them.

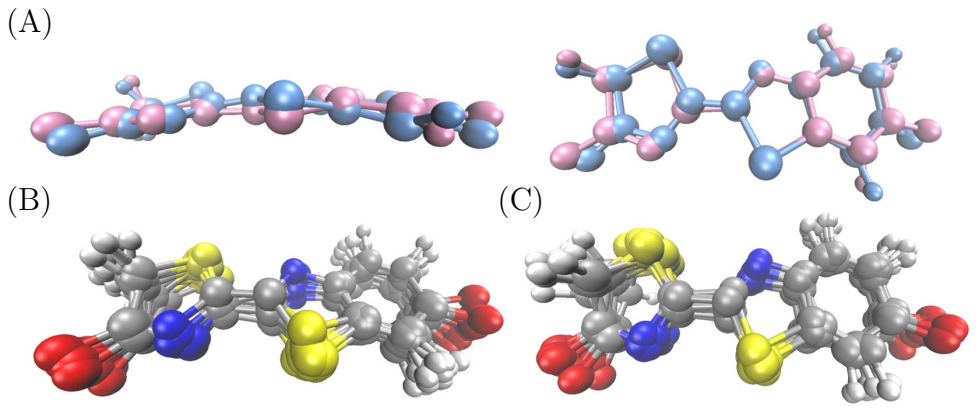


Figure 3.8: (A) Centroid geometries of the two clusters with the 20 snapshots with smallest (blue) and largest (pink) CT number values extracted from the QM/MM trajectory of the S_0 state. Clusters of the 20 geometries with (B) smallest and (C) largest CT number values.

In addition, the polar environment of the OLU for the small and large CT number clusters was analysed around the oxygen atoms of the tz (O_{tz}) and btz

(O_{btz}) fragments. As previously demonstrated,[118] the solvation of the O_{btz} is larger than that of the O_{tz}, as it can be observed in Figure B.6, where the RDF between the two oxygen atoms of the chromophore and the oxygen atom of the water molecules are represented. Moreover, it is shown that the cluster with the largest CT number values presents a larger solvation for both oxygen atoms. When looking at the number of hydrogen bonds listed in Tables B.9, B.10 and B.11, considering a donor–acceptor distance and an angle cutoffs of 3 Å and 150º, respectively, it is observed that hydrogen bonding between water and the chromophore is less important in the cases of the smallest CT number cluster. Contrary, the number of hydrogen bonds between the protein and the chromophore is very similar for both clusters. However, in the case of the smallest CT number cluster the water molecules are only present around the O_{btz}, while for the largest CT number cluster they surround both O atoms atoms. When the effect of the whole environment, the water molecules and the protein, is considered, it can be concluded that the cluster of geometries with the largest CT number presents a higher number of hydrogen bonds, mainly around the tz oxygen atom. This result is coherent with the electron transfer direction during the excitation, which happens from the btz to the tz moiety. A large positive charge around the tz fragment due to the presence of hydrogen atoms interacting with the oxygen atom of OLU stabilize the CT state. In summary, the CT character of the electronic transition can not be rationalized in terms of a set of intramolecular DOF, but it correlates well with the solvation of the oxygen atom of the tz moiety.

Additionally, in order to investigate the OLU/protein intermolecular DOF, the OLU/enzyme binding free energy was computed for the QM/MM S_0 dynamics by means of the 1A-MM-GBSA approach.[287–289] Then, the binding free energy was decomposed into amino acids contributions. In order to reduce the computational effort, the decomposition was performed for the 15 residues that are located at shortest distances from OLU along the dynamics. In addition, the ARG339 residue was also included in the analysis because it bears a positive charge that brings it to be close to the chromophore along the trajectory, as observed during a visual analysis of the dynamics, and it can be hypothesised a potential strong interaction between OLU and ARG339.

Figure 3.9 shows the binding free energy per residue and the electrostatic contribution, which should be the most relevant one for CT since in QM/MM vertical energy calculations within an electrostatic embedding only the point charges of the environment polarize the QM region by electrostatic interactions. As it can be observed, all the total binding free energies (in absolute value) are bellow -4 kcal·mol^{-1} and in most of the cases the electrostatic contribution is the most important one to the total energy, with the exception of the residues

PHE249, GLY341, LEU344 and THR345. Moreover, the residue ARG339 presents a very strong electrostatic interaction, as expected due to its positive charge and proximity to the chromophore. Additionally, similar analysis was performed on the S_1 classical trajectories with and without the AMP co-factor in order to elucidate whether the surrounding environment is the same and how the binding free energy per residue is affected. As it can be observed in Figure B.7, more than half of the interacting residues are the same in both simulations. However, the trajectory that includes AMP shows more residues present for at least 1% of the simulation time, which could mean that the presence of the co-factor induces some flexibility in the cavity and allows the interaction with more residues. However, as discussed above, the presence of these new additional residues interacting with OLU does not significantly influence the emission spectrum.

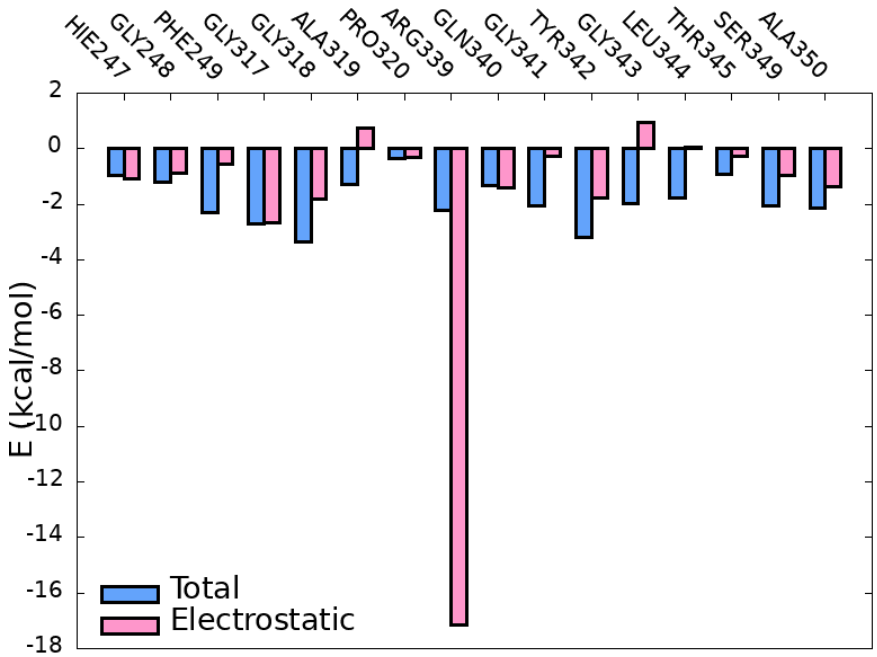


Figure 3.9: Total (blue) and electrostatic (pink) binding free energy of the chromophore with the closest residues of the protein along the S_0 QM/MM selected snapshots.

Once the most relevant protein residues have been identified, a potential correlation between their motion and the CT character of OLU was analysed. Figure 3.10(A) shows the variation of the CT number with the distance between the centres of mass of OLU and of each of the 16 residues for the 100

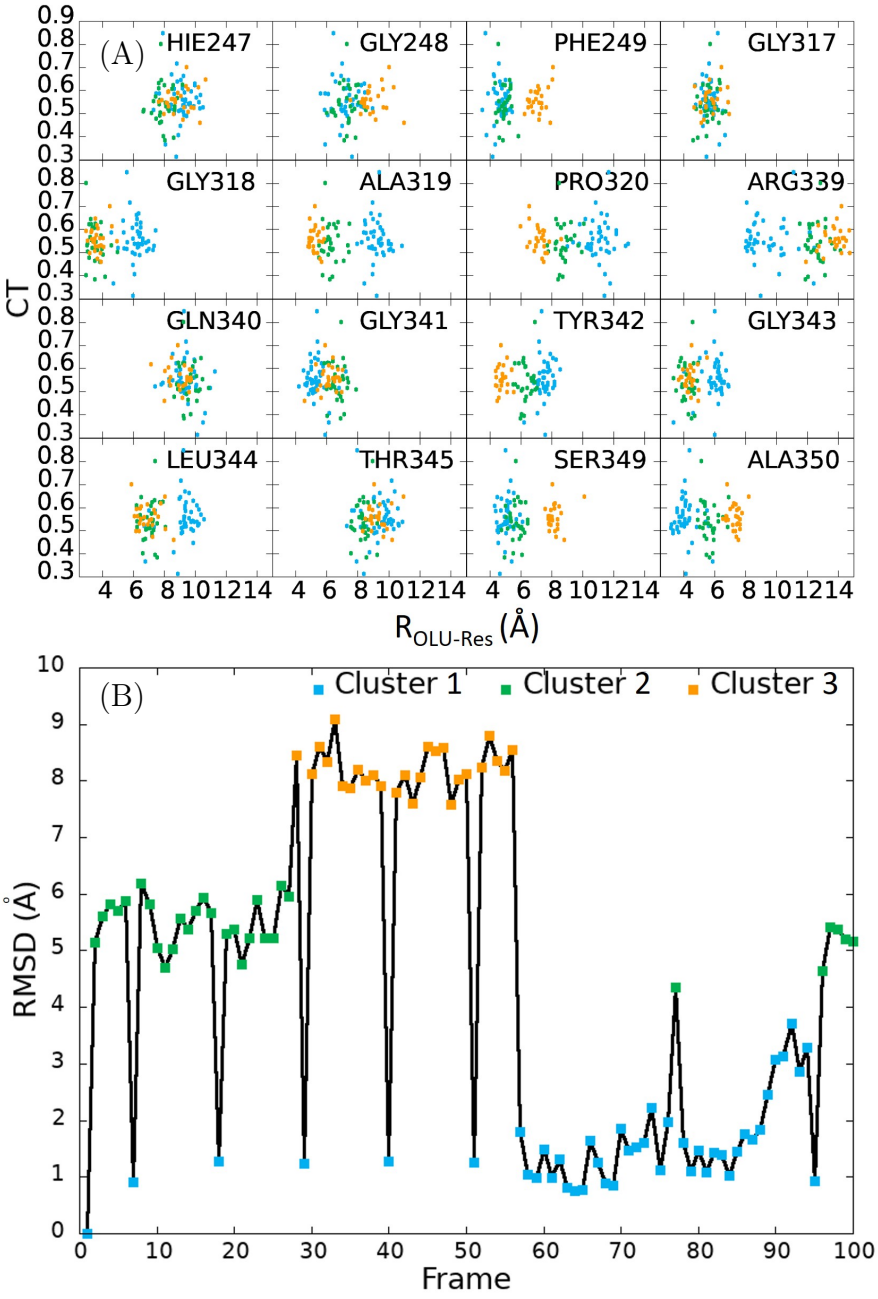


Figure 3.10: (A) CT number variation with the distance between the centres of mass of the chromophore and different amino-acids along the 100 S_0 QM/MM MD selected snapshots. (B) RMSD of the OLU inside the cavity. For both representations colours blue, green and orange represent clusters 1 to 3.

snapshots of the S_0 QM/MM trajectory. As it can be observed, the CT number does not depend on any of the distances since all the positions of the surrounding amino-acids of the protein present a very wide range of CT numbers. However, Figure 3.10(A) also shows that the position of some amino-acids is clustered into two or three regions defined by the distance from OLU. Previously, different positions of the chromophore were found for different protein species.[151] This finding indicates that the location of the chromophore inside the enzyme could be relevant to determine its electronic properties. In order to analyse further a possible relation between the position of the amino-acids and the CT of the excited state of OLU the following analysis was performed. First, the RMSD of the chromophore inside the cavity was computed and, second, a clustering analysis based on the RMSD on the selected snapshots of the trajectory was performed.

Figure 3.10(B) shows the result of the RMSD analysis and the separation in three different clusters, which are represented in Figure B.8. Moreover, colours in Figure 3.10(A) represent whether the chromophore belongs to cluster 1, 2 or 3 for each analysed structure. It can be observed that for some amino-acids, such as HIE247, GLY317, GLN340 and THR345, the distance to OLU does not depend on the cluster, while for others, for example, PRO320, ARG339 and ALA350, the clusters are well separated in terms of the distance to the chromophore. However, the different clusters do not provide different CT numbers. Specifically, the average CT numbers are 0.555 ± 0.087, 0.541 ± 0.085 and 0.559 ± 0.054 for clusters 1 to 3, respectively. In addition, when a similar analysis is performed for the previously mentioned clusters formed by the smallest and largest CT numbers, shown in Figure B.9, it can be observed that the smallest and largest values of CT are present for the same wide range of distances. Therefore, after a detailed analysis, we can exclude any dependency between the CT character of the emitting states of OLU and its position inside the cavity of the protein.

3.4 Conclusions

The intensive research on the OLU/luciferase complex performed along the last decades has shown that this biological system can be exploited in a wide range of applications. A deeper knowledge of the photophysics of the chromophore/enzyme complex will allow the tuning of its properties, leading to the development of more efficient devices with enhanced properties. The CT character of the emitting electronic state is one of the key features of the complex since it is intimately related to the emission intensity. However, a clear consensus about the nature of this electronic state was not achieved yet. Here, the bright state involved in both the absorption and emission spectra has been quantitatively characterized in terms of energy and CT character by combining MD simulations, hybrid QM/MM calculations, and transition density analysis. These results are compared with a simplified static model of OLU *in vacuum*, for which the electronic properties are computed along the SCCS torsion.

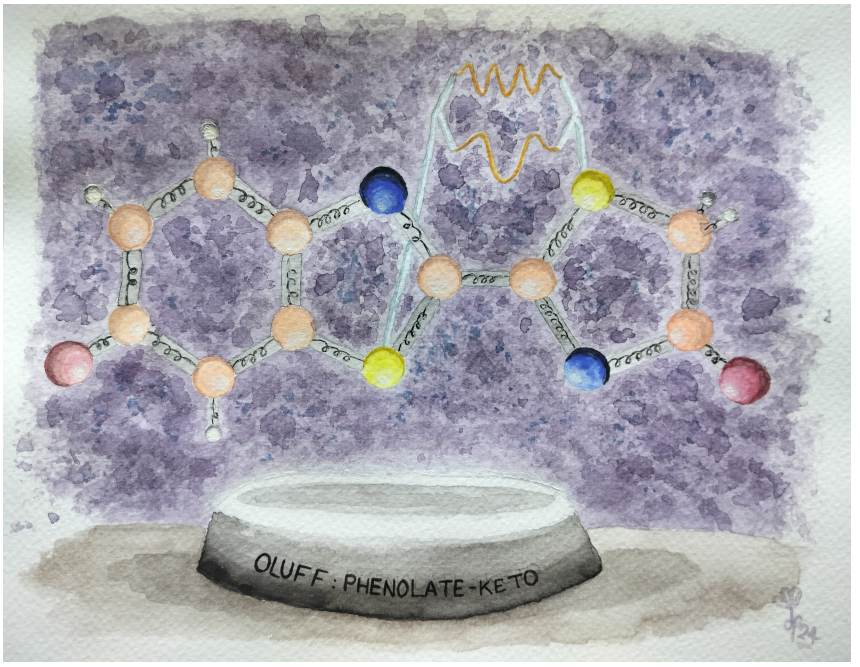
The static calculations of the chromophore *in vacuum* predict a red shift of the absorption and emission spectra when the molecule is distorted from the planar configuration. Moreover, such a distortion also induces the enhancement of the CT character and the reduction of the absorption and emission intensities. The spectra computed by QM/MM vertical energy calculations, based on snapshots sampled by classical MD, show a good agreement with the experiment. However, an energy dependency with the torsional angle was not found, since classical trajectories are not able to sample a wide range of torsional angles, which would be needed to obtain a remarkable red shift. On the flip side, when the conformational sampling is performed propagating the dynamics along the QM/MM potential energy surfaces, the torsional angle distributions are slightly broader and the absorption and emission spectra shows the expected red shift for conformations presenting larger deviations from the planarity.

The variation of the CT number with the SCCS dihedral angle, based on both classical and QM/MM MD simulations, was also investigated. However, a clear trend was not found because the interactions with the protein preclude large geometrical distortions of OLU. Nevertheless, further analyses including additional intramolecular and intermolecular DOF were performed. A clustering classification of the snapshots according to their CT number concluded that there is not correlation between the CT character of the chromophore and its intramolecular motion when OLU is embedded in the enzyme. Thus, the intermolecular OLU/protein and OLU/water DOF were also explored. Specifically, the protein residues that contribute the most to the binding free energy were identified. Then, a possible correlation between the motion of these residues

and the position of the chromophore inside the cavity with the CT character of the chromophore was searched. However, once again, a clear dependency of the CT number with the OLU/residue distance was not found. Nevertheless, when the solvation sphere around the chromophore was analysed, it was found that the presence of polar groups from the protein or water molecules interacting by hydrogen bonding with the oxygen atom of the tz moiety enhances the CT character of the electronic transition. Based on these analyses, it can be concluded that protein mutations that introduce flexibility in the protein, allowing for a larger degree of mobility of the chromophore, or increases the solvation of the chromophore will induce an enhancement of the CT character and a reduction of the emission intensity. In the same way, if the mutation leads a more rigid protein scaffold or a more non-polar environment the CT nature will be reduced, increasing the emission intensity.

Chapter IV

OLUFF: a Novel Set of Ground and Excited State FF Parameters of the Emitting Oxyluciferin Species



The modelling of the bioluminescent system of fireflies is key to understand the binding mode of the OLU/luciferase complex and its photophysical properties with the aim of developing high-efficiency devices and techniques. In this chapter, the OxyLUciferin Force Field (OLUFF) is presented, a model that is able to describe the interactions to sample the conformational space of the four possible OLU emitters in both ground and first electronically excited state. This FF has been parametrized to reproduce QM energies and geometrical parameters. Moreover, it has been validated by comparing probability distribution functions, minimized structures, infrared spectra and PC analysis obtained from OLUFF-based MD simulations with their QM counterparts. Additionally, ground state simulations have also been performed using the GAFF and compared with the OLUFF. It has been demonstrated that the OLUFF not only reproduces well the QM properties, but also improves the results from the GAFF.

* This chapter is based on the released article *Photochem. Photobio. Sci.* **2024**, 23, 2157–2171. DOI: 10.1007/s43630-024-00654-y with contributions of Henar Mateo-delaFuente[a] and Juan J. Nogueira.[a,b]

[a] Chemistry Department, Universidad Autónoma de Madrid, 28049 Madrid, Spain.

[b] Institute for Advanced Research in Chemical Sciences (IAdChem), Universidad Autónoma de Madrid, 28049 Madrid, Spain.

4.1 Introduction

Ideally, in order to systematically investigate the role of each of the amino acids surrounding the OLU chromophore, ALA mutations of the relevant amino acids identified in the previous chapter can be performed. These mutations can be very easily achieved by modifying, by hand, the original *pdb* file. The modification consists on (i) renaming the amino acid as ALA and (ii) removing all the atoms of the lateral chain but the *CB* carbon atom, which is the one directly bonded to the C_α of the amino acid. In the case the amino acid to modify is GLY, which does not have a lateral chain, the *CB* C atom can be removed as *tleap* module of AmberTools20[260] directly adds the H atoms in

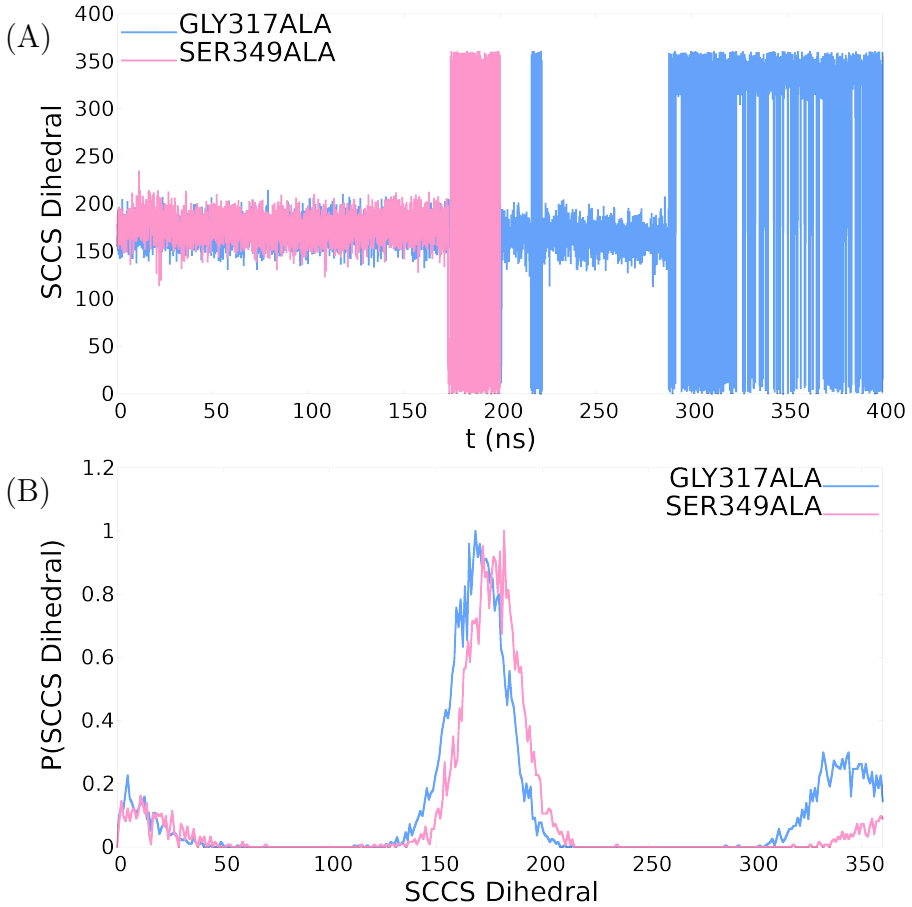


Figure 4.1: (A) Time evolution and (B) probability distribution of the SCCS dihedral angle upon the mutation of GLY317 (blue) and SER349 (pink) for ALA.

the correct position.

When the computational protocol explained in Chapter III was applied to perform MD simulations on the GLY317ALA and SER349ALA mutated systems, it was observed that the conformational sampling around the SCCS dihedral angle was not correct. As shown in Figure 4.1(A), the system, which starts in a s-trans conformation of the SCCS dihedral angle, is able to reach the s-cis conformation after a few hundred ns. Once this conformation is reached, it is not able to return to the s-trans conformation, leading to an unrealistic probability distribution of the SCCS torsion (Figure 4.1(B)) since that region of the conformational space corresponds to a maximum of the DFT-based PES (Figure 3.2).

This finding, together with the results of the previous Chapter III that already suggested that the sampling of the SCCS dihedral angle was not accurate enough to study this system by means of classical MD simulations, highlights the importance of the development of an accurate classical model for the OLU chromophore that is able to reproduce its movement, as running long QM/MM MD simulations in unaffordable. Within this context, this chapter aims to provide new set of FF parameters for the four possible OLU emitters shown in Scheme 1.6. The proposed FF, the OLUFF, is suitable for the conformational sampling of each of the four chromophores. Since standard FFs based on harmonic potentials are not able to describe bond formation and dissociation processes. Some previous studies also developed FF parameters for both S_0 and S_1 states of phenolate-keto tautomer of this system[339, 340] and a synthetic derivative.[328] However, a consistent parametrization for all the possible OLU emitters has yet not been reported.

4.2 Computational Details

The OLUFF was first parametrized by fitting QM data and, then, it was validated by comparison of QM MD simulations with classical MD simulations based on potential gradients computed by the OLUFF. Moreover, classical MD simulations were also performed using the GAFF.[261] The computational details of all these calculations are explained below.

4.2.1 Parametrization Process

The parametrization processes of each of the molecules for the S_0 and S_1 electronic states has been performed by fitting to QM data the previously described equation of the FF:

$$
\begin{aligned}
E_{total} = & \sum_{bonds} k_b(r_{ij} - r_0)^2 + \sum_{angles} k_\vartheta(\vartheta_{ijk} - \vartheta_0)^2 \\
& + \sum_{dihedrals} \sum_{n} V_n(1 + cos(n\varphi_{ijkl} - \gamma_n)) \\
& + \sum_{i<j} \left[4\epsilon_{ij} \left(\left(\frac{\sigma_{ij}}{r_{ij}} \right)^{12} - \left(\frac{\sigma_{ij}}{r_{ij}} \right)^{6} \right) + \frac{q_i q_j}{r_{ij}} \right]
\end{aligned}
\tag{4.1}
$$

where r_{ij} and r_0 represent the interatomic distance and the equilibrium distance; ϑ_{ijk} and ϑ_0, the angle between three consecutive atoms and the equilibrium angle; φ_{ijkl} and γ_n, the dihedral angle between four atoms and the phase factor associated to the position of the minima; n, the periodicity of the Fourier-series expansion for torsional terms and the number of terms of the series; ϵ, the minimum of the potential well; σ, the interatomic distance at which the value of the potential is 0; and q, the charges.

As in the previous chapter, the Seminario method[265] was applied to obtain the parameters of the harmonic terms of Equation 4.1, that is, the bond and angle potential terms and the equilibrium inter-atomic distances and angles were obtained from a geometry optimization. These optimized geometries for both S_0 and S_1 and the Hessian matrices, required to get the bond and angle force constants by the Seminario method, were obtained using the Gaussian09[341] software by means of DFT and TD-DFT *in vacuo* calculations, respectively, at the CAM-B3LYP/6-311 G(2d,p)[228, 231–233, 342–344] level of theory. In the case of the S_1 optimization, the TD-DFT calculation took into account three roots. Furthermore, the electrostatic charges were calculated with the MSK scheme[325] at the same level of theory.

As will be explained in the following sections, the only torsion degree of freedom that is truly able to rotate, the SCCS dihedral angle, has also been parametrized. This parametrization was performed with the software *PARAMFIT*[345] included in the Amber20 software.[260] The classical energy was fitted to SP QM energies, using a genetic algorithm with a mutation rate of 0.1 and a parent percent of 0.25 combined with a loosely converging simplex algorithm that is run on a random subset of the population once five generations (or iterations) had passed. The genetic algorithm is then allowed to recover for several generations before further simplex iterations are conducted. This hybrid approach dramatically speeds convergence to the global minimum, while maintaining the strengths of the genetic algorithm in searching a large, complex solution space with low sampling. Specifically, 50 iterations of possible optimizations were used together with a value of 10000 maximum number

of iterations the algorithm is allowed to run before it returns the best non-converged optimization. Additionally, the algorithm has been considered to be converged once 20 iterations in a row have passed without improvement in the best parameter set. Finally, the rest of the dihedral and improper torsions and Lennard–Jones parameters of all the chromophores were selected by means of the Antechamber *parmchk2* module of the AmberTools20 from the GAFF.[261]

4.2.2 Validation of the OLUFF

Once the FF parameters were developed, their accuracy was tested by comparison of OLUFF-based MD simulations with GAFF-based and QM MD simulations and static QM calculations. The computational details of the MD trajectories are explained in the following.

Classical Molecular Dynamics

To assess the behaviour of the OLUFF parameters, classical MD simulations for each of the OLU chromophores in both S_0 and S_1 states were performed. For each of the molecules, the initial structure optimised at QM level was classically minimized *in vacuum* by means of 5000 steps by the steepest descent algorithm[244] and for an additional 5000 steps using the conjugate gradient algorithm.[245] Consecutively, the system was progressively heated for 500 ps from 0 to 300 K and stabilized for another 500 ps at 300 K in the NVT ensemble using a 2 fs time step. The temperature was controlled by means of the Langevin thermostat[267, 268] with a 2.0 ps^{-1} of collision frequency. Finally, a production of 1 μs with a 2 fs time step and within the NPT ensemble was performed. The temperature was again controlled by means of the Langevin thermostat with a 1.0 ps^{-1} of collision frequency, while pressure was set to be 1.0 bar and controlled by means of the Berendsen barostat.[271] In these three steps, i.e., minimization, heating and production, the non-bonded interactions were computed with a cutoff of 999.0 Å and without PBC. Moreover, for both heating and production, the time evolution of the system was simulated by means of the velocity Verlet algorithm[253] and the bonds involving hydrogen atoms were constrained using the SHAKE algorithm.[274] Example inputs for the preparation of the system and MD simulations can be found in Annex C.1.

Apart from the classical MD simulations with the OLUFF, additional MD simulations of the chromophores in the S_0 state were carried out using the GAFF for all the parameters, but with charges that were calculated with the

MSK scheme and the bond and angle parameters, which were obtained by the Seminario method.[265] These simulations, that followed the previously described protocol, were used to compare the probability distribution of the SCCS dihedral angle to determine the performance of the OLUFF against the commonly employed GAFF.

Quantum Mechanics Molecular Dynamics

The accuracy of the previously explained classical (OLUFF and GAFF-based) MD simulations was evaluated by comparison with QM MD simulations. Specifically, QM MD simulations for the OLU phenolate-keto and phenol-enolate species in the S_0 were evolved using the software Orca (Version 5.0.4).[346, 347] These simulations were run with a time step of 1.0 fs using the CAM-B3LYP/6-31 G^*[342, 348–351] level of theory together with the Def2/J auxiliary basis set[352] for the resolution of identity approximation for Coulomb integrals (RI-J) and numerical chain-of-sphere integration for the HF exchange integrals (COSX).[353] Note that a smaller basis set (6-31 G^*) than that used for the SP energy calculations employed to fit the potential (6-311 G^{**}) was used, since the torsional PES is the same, as shown in Annex C.2, and performing the QM MD with the larger basis set is unfeasible. Moreover, the initial geometry was optimized at the same level of theory and the atomic velocities were randomly initialized based on a Maxwell–Boltzmann distribution at 300 K, the atomic positions were propagated with the velocity Verlet algorithm[253] in the NVT ensemble and the temperature of 300 K was controlled with the Berendsen thermostat[271] with a coupling strength of 10.0 fs. In the case of the OLU phenolate-keto, a single QM MD trajectory was run for 100 ps from the QM-minimized structure, while two trajectories were calculated for the phenol-enolate one from the QM-minimized structure (SCCS dihedral angle of $180^{\underline{o}}$) for 57 ps and another from an SCCS dihedral angle of $128^{\underline{o}}$ for 39 ps due to the computational cost.

Infrared Spectra and Principal Components Analysis

In addition to the QM MD simulations, IR spectra and the normal modes were also calculated to assess the performance of the parametrization. For this, additional classical MD simulations were preformed from the final step of the heating classical trajectory explained above. The trajectories were propagated for 200 ps with a time step of 0.1 fs at 303.15 K and 1.0 bar within the NPT ensemble and without PBC. Again, temperature and pressure were controlled with the Langevin thermostat[267, 268] with a 1.0 ps^{-1} of collision

frequency and the Berendsen barostat,[271] respectively. The non-bonded interactions were computed with a cutoff of 999.0 Å and the time evolution of the system was simulated by means of the velocity Verlet algorithm. However, no SHAKE algorithm was applied here to take into account the vibrations of the bonds involving H atoms. These trajectories were carried out, for each of the chromophores, using two sets of parameters: the OLUFF and the whole GAFF set using the MSK charges. In the case of the simulations using the GAFF, minimization and production were computed as previously described in Section *Classical Molecular Dynamics*.

For the IR spectra, both velocities, calculated during these simulations, and charges (q) of the FF were used to compute the dipole derivative autocorrelation function as described in Section 2.2.8, taking into account the separation of the MD trajectory into fragments of 1 ps to account for different time origins following Scheme 2.1 and to compute the time-averaged–trajectory-averaged dipole derivative autocorrelation function, which is related to the intensity of the IR spectrum. The IR spectra computed from the classical MD simulations were compared with the IR spectra calculated at the QM optimized geometry of each of the chromophores at the CAM-B3LYP/6-311 G(2d,p) level of theory.

For the PC analysis of the simulations, which is a classical equivalent to the QM normal modes, the 200 ps trajectories were analysed with the *diagmatrix* and *modes* commands of the *cpptraj* module of AmberTools20[354] to calculate the eigenvalues of the MW covariance matrix by means of quasiharmonic analyses. These results were contrasted with the QM frequencies.

4.3 Results and Discussion

This chapter presents the OLUFF parameters of the OLU phenolate-keto, phenol-keto, phenol-enolate and phenolate-enolate emitters of the firefly's BL system for both S_0 and S_1 electronic states. Although bonding and angle parameters are also included in the parametrization via the Seminario method,[265] this chapter aims to provide good barriers (V_n) and phases (γ_n) parameters that properly describe the torsional movement of the SCCS dihedral angle, since it is the only torsional degree of freedom that is truly able to rotate and it was shown in the previous chapter[340] that the use of the GAFF[261] does not reproduce the probability distribution obtained from QM MD simulations. In addition, the GAFF parameters do not reproduce either the PES obtained from a rigid scan of the SCCS torsion degree of freedom.

The results of this chapter are organized in the following way. First, the global fitting process workflow is explained. Then, the potential energy profile

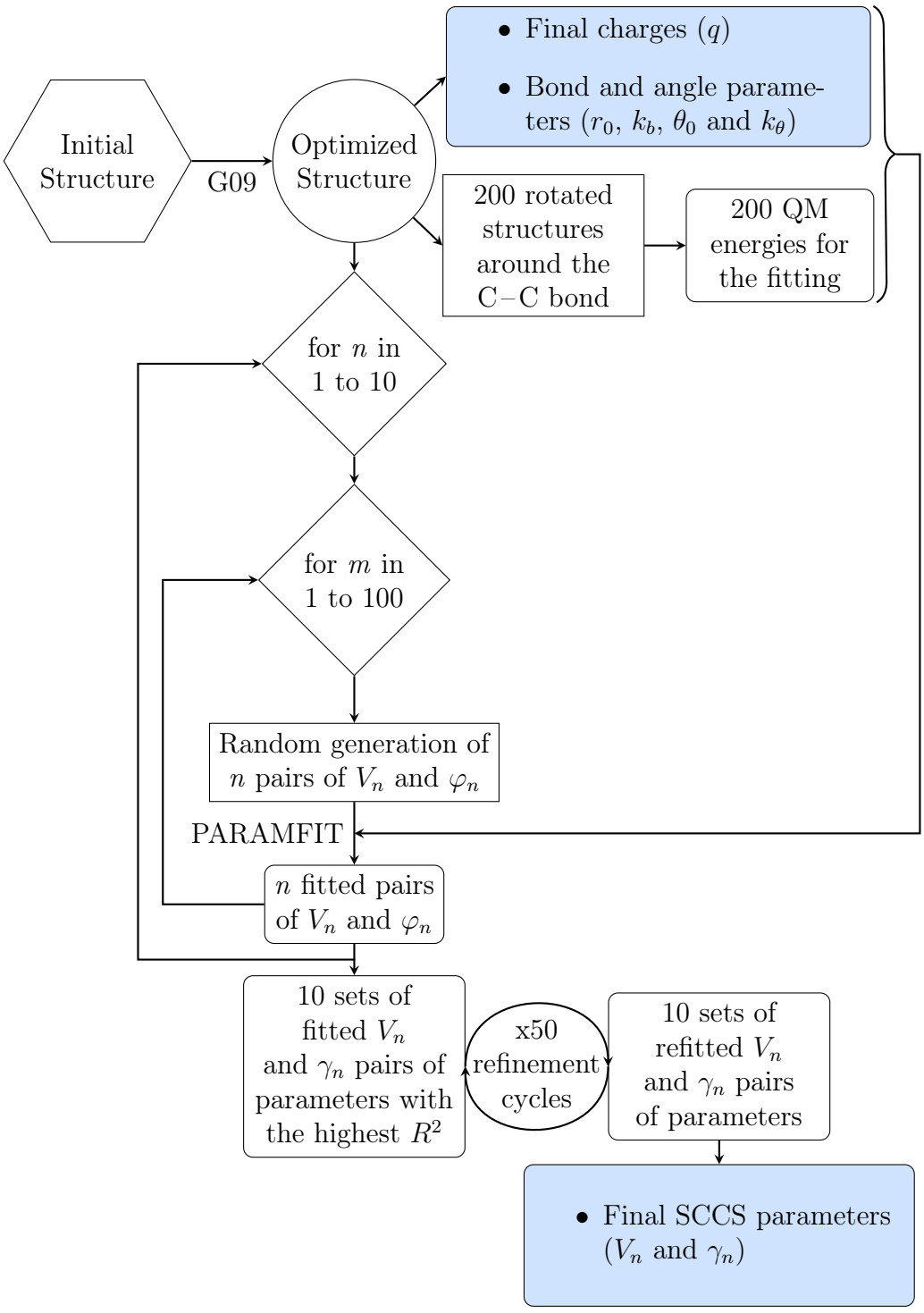
along the SCCS dihedral angle for the four chromophores is computed at the QM level. These calculations are compared with the energy profiles obtained with the GAFF parameters and with the new parameters of the developed OLUFF. Finally, the validation of the OLUFF is discussed in terms of the probability distribution of the dihedral, QM and MD minimized structures, the IR spectrum and the PC analysis.

4.3.1 Global Fitting Process

The parametrization process for all the molecules followed the steps represented in Scheme 4.1. Its aim is to fit the energies and Hessians calculated at QM level of theory to the classical energy equation shown in Equation 4.1. Firstly, the initial structure of the chromophore was optimized at QM level using the Gaussian09 software.[341] From this calculation, some of the new parameters have already been taken, namely: the atomic partial charges (q), the bond equilibrium distances (r_0) and the angle equilibrium values (ϑ_0). Moreover, from the diagonalization of the Hessian matrix, computed from the optimized geometry, bond and angle force constants (k_b and k_ϑ, respectively) were also calculated using the Seminario method.[265] Consecutively, 200 random conformations were obtained from the rotation of the optimized structure around the SCCS dihedral angle. These structures were calculated at QM level to obtain reference energies for the fitting of the parametrization process. These energies have also been represented in Figure 4.3 after the parametrization process to show the accuracy of the fitting. Moreover, the SP calculations of the phenolate-keto chromophore were used to ensure that the nature of S_1 state, which presents a $\pi\pi^*$ character, does not change along the torsion (see Figure C.2) as shown in the previous chapter.[340] This indicates that potentially problematic regions close to conical intersections, which are badly described by TD-DFT, are not present.

The next step consisted in the determination of the ideal number of terms to introduce in the description of the energy of the SCCS torsion, which corresponds to the term n in Equation 4.1. To do so, for values of n from 1 to 10, 100 sets of random values of V_n and γ_n, between 1 and 10 kcal/mol and 0 and 360º, respectively, were used as an initial guess for the *PARAMFIT* parametrization. This fitting procedure was seen to be convergent with respect to the number of terms in Equation 4.1, as depicted in Annex C.4.

After this step, Least Squares Difference (R^2) between the input quantum energy and the calculated AMBER energy for the 100 fittings for each of the ten values of n are calculated. As it can be observed in Figure 4.2, in which this value is represented (in pink) together with the number of fittings with



Scheme 4.1: Computational steps followed during the parametrisation process.

an R^2 larger than 0.7 (in blue), any of the chromophores present more than three suitable values of n for the expansion. The use of two and four sets

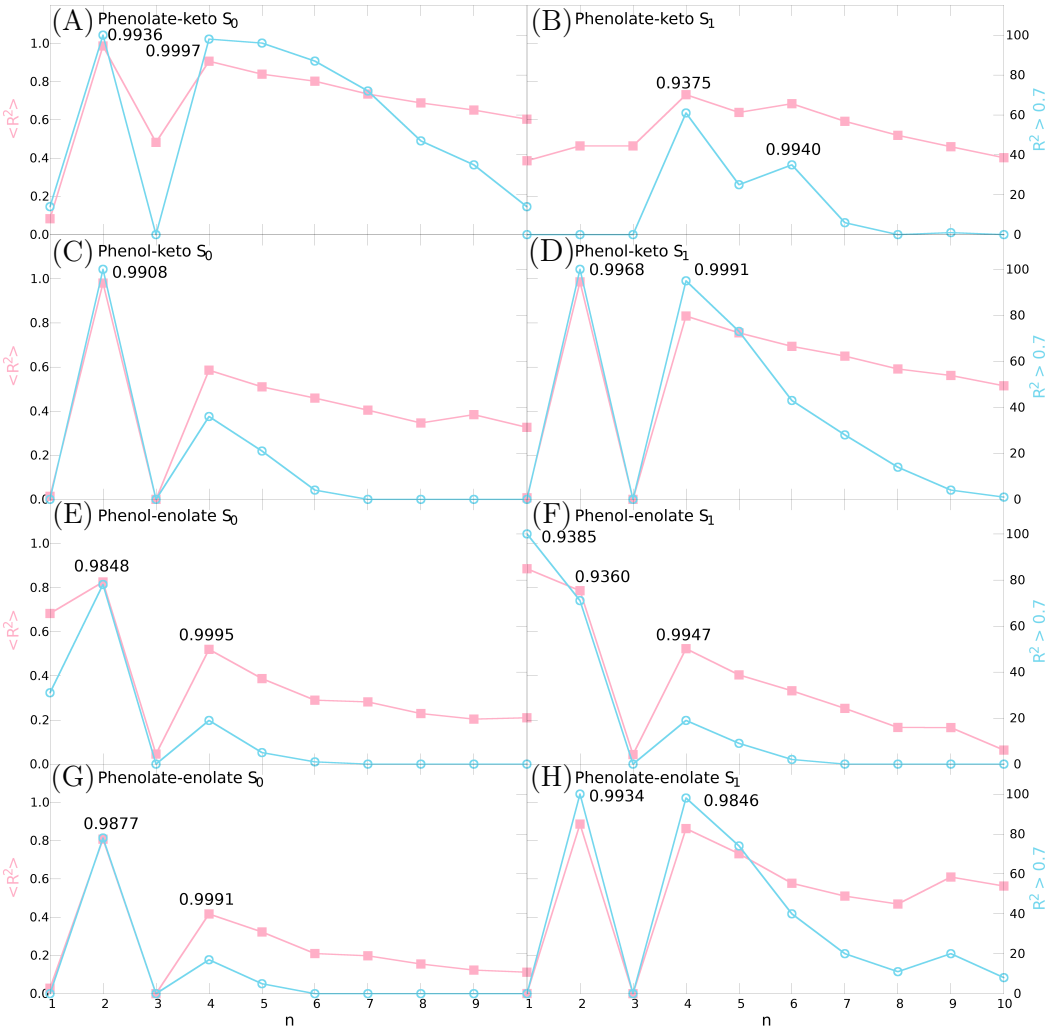


Figure 4.2: Average correlation coefficient (pink) and number of fittings m with a R^2 higher than 0.7 with respect to the number of terms in the torsion potential achieved during the initial 100 fittings of the SCCS dihedral angle parameters for the (A, C, E, G) S_0 and (B, D, F, H) S_1 states of the OLU (A, B) phenolate-keto, (C, D) phenol-keto, (E, F) phenol-enolate and (G, H) phenolate-enolate. In the plot, also the highest R^2 appear for each of the OLU chromophores.

of parameters (n=2 and 4) is suitable for all the studied systems except for phenolate-keto S_1, for which three and six sets (n=3 and 6) are a reasonable match, and phenol-keto S_0, whose only reasonable value of n is 2. Additionally, n=1 resulted in a reasonable value for the OLU phenol-enolate in the excited state. Consecutively, from these 100 fittings with random initial sets of parameters, the five sets of resulting parameters for the previously mentioned n values that present a higher value of R^2 were used as initial guesses for further refinement, with the exception of the phenol-keto in the S_0 for which ten sets of parameters were selected, since only one reasonable value of n was found. Thus, each set was used as the initial guess for another *PARAMFIT* fitting, whose result was employed as the new initial guess of a new fitting. This cyclic process, whose results are depicted in Annex C.5, was repeated 50 times for each of the chromophores. As a result of this whole process, the correlation coefficients after the refinement cycles are compared and the set of parameters corresponding to the highest R^2, represented in Figure 4.2, is selected as the V_n and γ_n of the SCCS torsion of each OLU molecule. Finally, the third term of Equation 4.1 is represented in Figure C.8 outside the fitted region, from -360 to 720º, to check the periodicity of their torsional functions outside the investigated domain and, thus, discard possible over-fitting. Since the periodicity is maintained outside the 0 to 360º region, the OLUFF torsional parameters do not suffer from over-fitting.

In summary, the parametrization process takes the charges and the bond and angle equilibrium distances from the QM optimization and calculates the one–two and one–three bonding force constants from the QM Hessian matrix at the s-trans minimum of the PES. The SCCS dihedral is obtained by fitting the third term of Equation 4.1 to the rigid PES around that torsion. The rest of the dihedral angle parameters and the non-bonded terms are taken from the GAFF. All the parameters for the four chromophores in both the S_0 and S_1 states can be found in Amber format, in the *frcmod*, *leaprc* and *mol2* files in the Supporting Information of the published article.[264]

4.3.2 Potential Energy Profiles and Probability Distributions

As initial criteria for the validation of the OLUFF potential, it is straight forward to compare the OLUFF and GAFF-based and QM rigid PESs around the SCCS torsion. Let's start by discussing the profile and distribution for the phenolate-keto specie in the ground state S_0 and S_1 states, represented in Figures 4.3(A,B), where the potential energy profiles along the SCCS dihedral angle computed by DFT and TD-DFT for S_0 and S_1, respectively, are shown

by the violet lines. For S_0, the QM potential presents the global minimum at 180^{o} corresponding to the s-trans conformation, and a partial minimum at $0^{o}/360^{o}$ (s-cis), but the latter is 6.69 kcal/mol higher in energy than the former and its population from the global minimum requires overcoming an energy barrier of 22.97 kcal/mol. This means that the population of this second minimum is not likely to occur during a MD simulation. This is corroborated by the probability distribution function obtained from the QM MD simulation evolved for 100 ps, which is shown by the violet points in Figure 4.3(A). This situation is accurately reproduced by the OLUFF parametrized in this chapter. The OLUFF potential energy profile (black line) and the dihedral angle probability distribution (black points) obtained from a classical MD employing this new potential perfectly overlap with the DFT results. Nonetheless, this does not happen when the GAFF is used. While the GAFF potential energy curve reproduces well the position of the global and partial minima at 180^{o} and $0/360^{o}$, respectively, the energy difference between them and the energy barrier that separate the two minima are 1.26 and 6.27 kcal/mol, respectively. This erroneous behaviour has two consequences in the classical MD simulation performed with GAFF: (i) a broader probability distribution than the QM one at the global minimum, and (ii) the population of the secondary minimum at $0^{o}/360^{o}$. Therefore, the OLUFF performs much better than the GAFF for phenolate-keto in the ground state.

The agreement between TD-DFT and the OLUFF is also very good for the phenolate-keto species at S_1, as seen in Figure 4.3(B). In this case, a comparison with the GAFF is not possible, since this potential was parametrized only for the ground state. Both the QM and the OLUFF present three accessible minima at the s-trans conformation (180^{o}) and at two twisted geometries with angles of 90^{o} and 270^{o}. Since these minima are virtually isoenergetic and the energy barrier between them is only 6.71 kcal/mol, all of them can be populated during the dynamics. In fact, this is exactly what happens in a classical MD simulation performed with the OLUFF, as observed in the probability distribution of the SCCS angle (black line). The QM MD simulation was not run for S_1 for the sake of computational efficiency. However, the OLUFF-based probability distribution perfectly agrees with the shape of the QM potential. In this regard, the two more intense and narrower peaks of the distribution are obtained at 90^{o} and 270^{o}, where the curvature of the potential is larger.

In the case of the other OLU forms, it can be observed in Figure 4.3 that the OLUFF parameters for the S_0 reproduce much better the potential energy profile calculated at DFT level than those of the GAFF. The only exception is the case of the OLUFF phenol-keto Figure 4.3(C), for which both the OLUFF and GAFF have similar accuracy and slightly underestimate the QM barrier

between the s-trans (180º) and s-cis (0/360º) forms and the energy of the s-cis minima. However, this difference is larger for the GAFF than for the OLUFF, specially for the barrier when going from the s-cis to the s-trans conformations, since the energy of the s-cis minimum is better reproduced

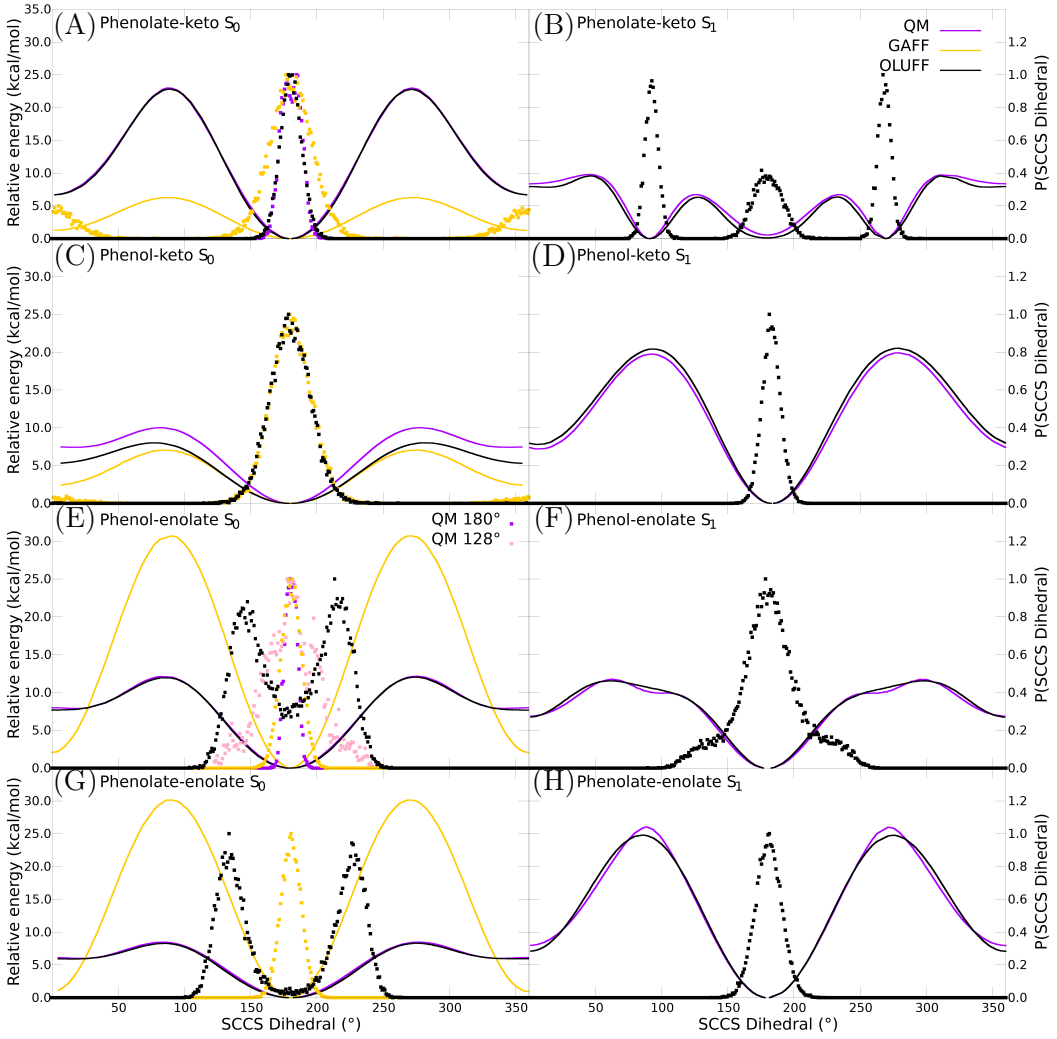


Figure 4.3: Rigid PESs (solid lines) and probability distributions (points) of the SCCS dihedral angles of the (A, C, E, G) S_0 and (B, D, F, H) S_1 states of the OLU (A, B) phenolate-keto, (C, D) phenol-keto, (E, F) phenol-enolate and (G, H) phenolate-enolate. Energies are relative to the minimum of each surface.

with the OLUFF. The fitting of the OLUFF to the S_1 QM potential of the three molecules, shown in Figures 4.3(D,F,H), is very good and classical MD simulations with the OLUFF display probability distributions peaking at the s-trans conformation (180^{o}). In the case of the phenol-enolate chromophore (Figure 4.3(F)), the distribution is broader, going from 100^{o} to 260^{o} due to the flat shape of the potential.

Finally, let's discuss the two most complex situations: the phenol-enolate and the phenolate-enolate in the ground state (Figures 4.3(E,G)). The GAFF correctly predicts the position of the global QM minimum at 180^{o}, but it largely overestimates the QM energy barrier that separates the different isomers. As a consequence, the GAFF-based dynamics provide an intense and narrow distribution centered at 180^{o}. On the contrary, the OLUFF profile again perfectly overlaps with the QM profile. However, when a classical MD is evolved with this FF, a surprising result is obtained: the probability distribution presents two maxima at 125^{o} and 225^{o}, and the population at 180^{o} (the position of the global minimum) is smaller. This can be rationalized in the following way. It is important to note that the potential energy profiles shown in Figure 4.3 are cuts of the entire PES. Therefore, the global minimum of each of these cuts is not necessarily the global minimum of the potential energy surface. Thus, different values of the SCCS angle can be populated along the trajectories, in apparent contradiction with the profiles shown in Figure 4.3. This could be corroborated by running a long MD simulation at the QM level. However, such a long simulation is not computationally feasible. Instead, two different QM MD trajectories were run starting with the phenol-enolate molecule at different values of the SCCS angle: 180^{o} and 128^{o}. These two values are representative values of the OLUFF-based dynamics probability distribution. When the QM MD starts from the s-trans conformation (180^{o}), the system is trapped in this planar conformation for the entire simulation time of 57 ps. However, when the simulation starts from the distorted dihedral angle (128^{o}), the system presents a dihedral distribution qualitatively similar in shape to that obtained with the OLUFF. Had it been possible to run a QM MD for hundreds of ns, it is very likely that the shape of the QM distribution would have converged to that of the OLUFF distribution. Moreover, Figure C.9 clearly shows the oscillations of the dihedral angle along the QM dynamics between the two minima predicted by the OLUFF-based simulations. Therefore, this analysis explains the behaviour of the SCCS torsion and confirms the accuracy of the developed parameters. Moreover, this finding is consistent with the results reported in the literature,[355] where a small minimum in the relaxed PES of the phenol-enolate OLU is described at around 167^{o}. The situation of the phenolate-enolate chromophore in the S_0 Figure 4.3(G) is

exactly the same. However, the QM MD trajectory was not run because it is extremely computationally demanding.

4.3.3 QM and OLUFF Minimized Structures

Apart from the PES, the QM and OLUFF minimized structures are compared in Figure 4.4 by means of the RMSD. As it can be observed, this value is very low for all the systems (bellow 0.350 Å), which means that both QM and OLUFF minimized structures are in good agreement, specially for the chromophores in the S_1. In the case of phenolate-keto S_0, the difference between QM and OLUFF minimized structures is due to the not strictly planar heavy atom skeleton. However, for the S_1, only the C_α of the carbonyl group is out of plane. Differences in the phenol-keto S_0 are due to a tilt of the H atoms on the C_α of the carbonyl group, while both phenol-enolate and phenolate-enolate S_0 present a slight bend of the tz moiety with respect to btz and a SCCS dihedral angle smaller than 180º (217 and 227º, respectively). Moreover, in the

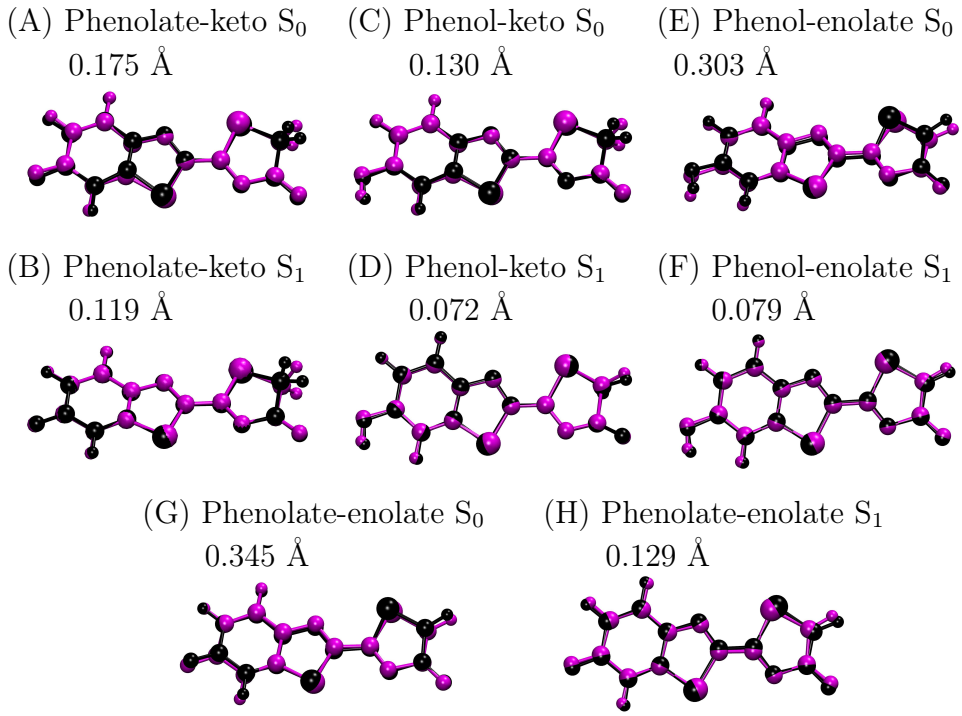


Figure 4.4: QM (violet) and OLUFF (black) minimized structures of the OLU (A, B) phenol-keto, (C, D) phenol-enolate and (E, F) phenolate-enolate of the (A, C, E) S_0 and (B, D, F) S_1 states with their RMSD values.

phenolate-enolate S_1 the QM and OLUFF differences correspond to the angles between the atoms of the tz ring. However, despite this small differences, the agreement between the QM and the OLUFF geometries is very good.

4.3.4 Dynamic Infrared Spectra

In addition to the probability distributions of the SCCS dihedral angle and the minimized structures, the IR spectra were calculated from classical MD, as specified in the computational details. As it can be observed in Figure 4.5, the spectra obtained with the OLUFF parameters lay within the same range of the one calculated at QM level for all the chromophores in both S_0 and S_1. In the following, the IR spectrum of each chromophore in each electronic excited state will be discussed in detail.

For the chromophores in the S_0 state, it is possible to observe that the shape of the QM IR is better reproduced by the OLUFF than by the GAFF. In the case of the phenolate-keto OLU Figure 4.5(A), the QM spectrum presents four peaks of high intensity and three with a medium one with the following structure: two consecutive pairs of peaks of medium and high intensity (1126 and 1211 cm^{-1} and 1305 and 1393 cm^{-1}), respectively, followed by a peak of medium intensity (1460 cm^{-1}) and two of high intensity (1514 and 1593 cm^{-1}). This pattern is also present in the IR calculated from the OLUFF: two pairs of peaks of medium and high intensity (903 and 997 and 1337 and 1470 cm^{-1}) next to a medium-intensity peak (1800 cm^{-1}) and two peaks of high intensity (1870 and 1971 cm^{-1}). However, the IR spectrum calculated from the GAFF only presents three peaks of high intensity and three with medium one since, instead of two consecutive pairs of peaks of medium and high intensity, it only presents one set of medium- and high-intensity peaks (903 and 997 cm^{-1}), followed by two peaks of medium intensity (1303 and 1429 cm^{-1}) and other two of high intensity (1635 and 1773 cm^{-1}). Moreover, both the OLUFF and the GAFF present an additional band of medium intensity at 532 cm^{-1} that might correspond with the very small peaks in the QM IR spectrum.

Regarding the phenol-keto chromophore show in Figure 4.5(C), the IR spectrum calculated from the QM single point presents two intense peaks (1164 and 1197 cm^{-1}), followed by three peaks of medium intensity (1267, 1295 and 1584 cm^{-1}) and other two of high intensity (1646 and 1866 cm^{-1}). For this chromophore, the IR computed from the OLUFF has two intense peaks (904 and 997 cm^{-1}), two peaks of medium intensity (1299 and 1370 cm^{-1}) and other two of high intensity (1635 and 1935 cm^{-1}), while that from the GAFF has two intense peaks (904 and 997 cm^{-1}), three peaks of medium intensity (1233, 1308 and 1503 cm^{-1}) and just one of high intensity (1640 cm^{-1}). Moreover, the two

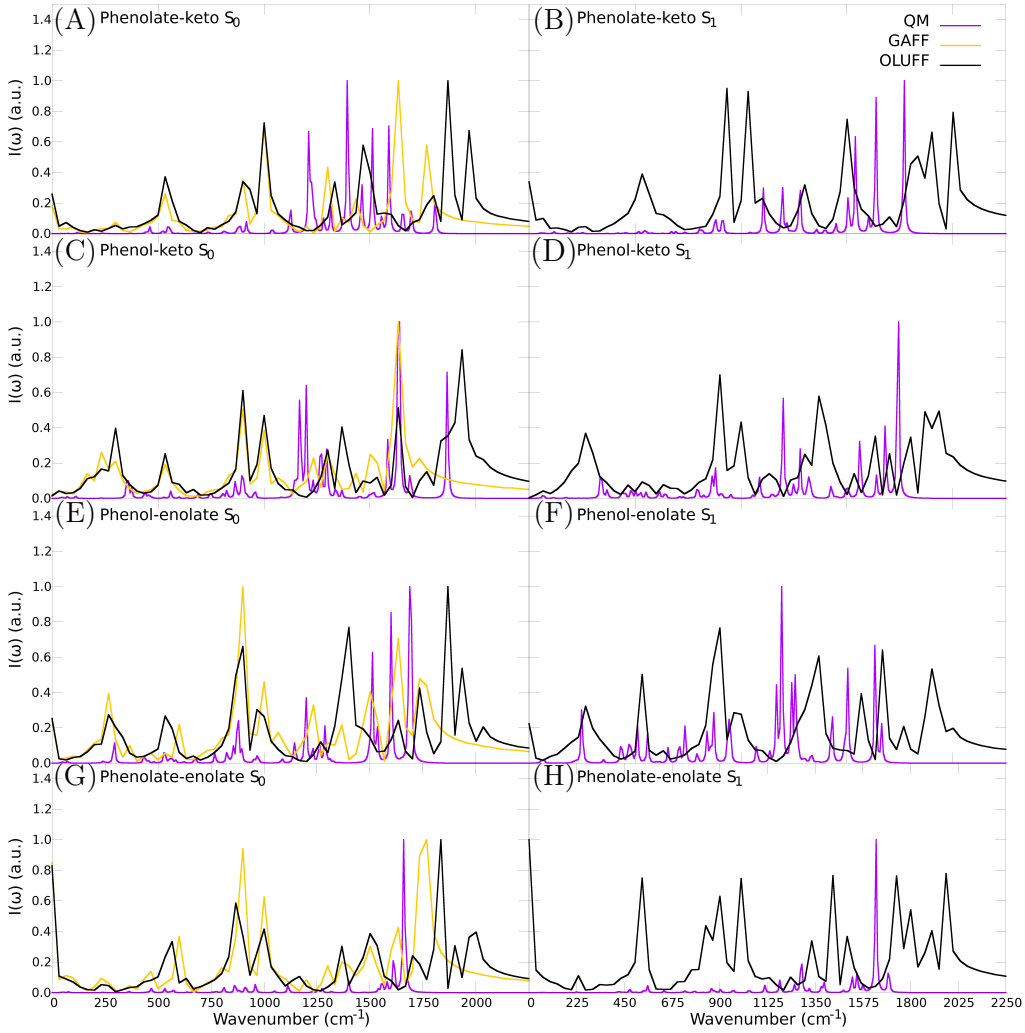


Figure 4.5: Infrared spectra of the OLU (A, B) phenol-keto, (C, D) phenol-enolate and (E, F) phenolate-enolate of the (A, C, E) S_0 and (B, D, F) S_1 states calculated from the QM geometry optimization and the MD simulations using the GAFF and the OLUFF.

IR spectra calculated from the classical MD also present two peaks of medium intensity at lower wavenumber values (around 300 and 500 cm^{-1}) that do not appear in the QM IR spectrum. In this molecule, the two FF models differ from the QM reference. However, the difference is more pronounced for the GAFF than for the OLUFF, since the former misses a peak of high intensity, and the latter one of medium intensity.

Concerning the phenol-enolate OLU depicted in Figure 4.5(E), the IR spectrum from the QM calculation presents two peaks of medium intensity (877 and 1194 cm^{-1}), followed by three strong peaks of increasing intensity (1516, 1603 and 1693 cm^{-1}). Although this structure of the spectrum is not strictly reproduced by any of the IR calculated from the classical FFs, the IR calculated from the OLUFF-based MD does present two peaks of medium intensity (269 and 536 cm^{-1}) and three peaks of increasing intensity (894, 1400 and 1875 cm^{-1}). However, additional peaks of medium intensity also appear between them (973, 1635, 1735 and 1937 cm^{-1}). On the flip side, the spectrum calculated from the GAFF-based MD presents two peaks of medium intensity (269 and 606 cm^{-1}), followed by a high-intensity peak (904 cm^{-1}), four peaks of medium intensity (1000, 1231, 1374 and 1503 cm^{-1}), a high-intensity peak (1638 cm^{-1}) and a medium-intensity peak (1731 cm^{-1}).

Finally, with respect to the phenolate-enolate OLU 4.5(G), the QM IR spectrum only presents a medium and a high-intensity peak at 1614 and 1660 cm^{-1}, respectively, while both FF-based spectra present many more. The spectrum calculated from the OLUFF MD has two peaks of medium intensity (561 and 1001 cm^{-1}) separated by an intense peak (866 cm^{-1}), followed by three other peaks of medium intensity (1364, 1501 and 1706 cm^{-1}), which could correspond to the low-intensity peaks in the QM spectrum between 1546 and 1630 cm^{-1}, an intense peak (1843 cm^{-1}) and two additional peaks of medium intensity (1911 and 2004 cm^{-1}). A similar structure can be observed in the IR from the GAFF-based simulation, with the difference that the peaks that only appear in the FF-based spectra present a higher intensity when compared with the IR spectrum from the OLUFF. It has a medium-intensity peak at 600 cm^{-1}, followed by two high-intensity peaks (894 and 1003 cm^{-1}), four small peaks of increasing intensity (1301, 1377, 1506 and 1641 cm^{-1}) and a very intense peak at 1773 cm^{-1}. These results show that the use of the GAFF for phenolate-keto, phenol-keto and phenol-enolate chromophores in S_0 fails to reproduce one normal mode of vibration active in IR and, for the phenolate-enolate OLU, it presents more peaks of higher intensity which could lead to differences in the interaction with the protein. Overall, the description of the IR spectrum for the molecules in the ground state is better when the OLUFF is employed.

With regard to the chromophores in the S_1 state and even though the dif-

ferences between the IR spectrum computed with QM and from the OLUFF MD are larger than in the S_0, it is still possible observe some similarities. The QM spectrum of the phenolate-keto, shown in 4.5(B), presents three consecutive peaks of medium intensity (1104, 1197 and 1279 cm^{-1}) and another one (1507 cm^{-1}) prior to three strong peaks of increasing intensity (1542, 1638 and 1768 cm^{-1}). However, the OLUFF-based IR presents a medium-intensity band (897 cm^{-1}), which could correspond to the low-intensity peaks of the QM IR between 866 and 920 cm^{-1}. This band is followed by two intense peaks (931 and 1036 cm^{-1}), a peak of mediumintensity at 1299 cm^{-1} and a peak of high intensity (1505 cm^{-1}), a region that could correlate to the three consecutive peaks of medium intensity of the QM IR; and three strong peaks of increasing intensity (1834, 1903 and 2001 cm^{-1}).

Regarding the phenol-keto chromophore presented in 4.5(D), the QM spectrum has a small peak at 897 cm^{-1} and a strong peak at 1201 cm^{-1}, followed by three peaks of medium intensity (1281, 1560 and 1685 cm^{-1}) and a strong peak at 1743 cm^{-1}. The IR calculated from the OLUFF MD presents a medium-intensity peak at 269 cm^{-1} that could correspond to a small peak of the QM spectrum. It also has a strong peak at 899 cm^{-1} and a peak of medium intensity at 997 cm^{-1}, followed by a strong peak (1368 cm^{-1}) and four peaks of medium intensity (1635, 1802, 1866 and 1939 cm^{-1}). Thus, the difference between the two spectra would be the apparition of an extra intense peak at 1369 cm^{-1} and the splitting of the intense band of the QM IR at 1743 cm^{-1}.

In the case of the phenol-enolate (4.5(F)), the QM spectrum presents two (247 and 511 cm^{-1}) plus three (737, 874 and 947 cm^{-1}) peaks of medium intensity, followed by two groups of one (1168 cm^{-1}) and two (1242 and 1257 cm^{-1}) peaks separated by an intense peak at 1193cm^{-1}, a medium-intensity peak at 1432 cm^{-1}, two intense peaks (1505 and 1633 cm^{-1}) and a medium-intensity peak (1663 cm^{-1}). On the flip side, the OLUFF-based spectrum has a peak of medium intensity at 269 cm^{-1}, which could correspond to the convolution of the first two peaks of the QM spectrum, and an intense peak at 535 cm^{-1}, which could correspond to the convolution of the next three peaks, followed by a high-intensity peak (901 cm^{-1}) and a medium-intensity peak (987 cm^{-1}). Moreover, it also has a high-intensity peak (1366 cm^{-1}) that is not present in the QM spectrum, in addition to a medium-intensity peak at 1570 cm^{-1}, two intense peaks (1624 and 1904 cm^{-1}) and a medium-intensity peak (2002 cm^{-1}).

Finally, for the phenolate-enolate, depicted in 4.5(H), the OLUFF-based spectrum fails to reproduce the relative intensity of the strong peak of the QM spectrum (1648 cm^{-1}), which corresponds to the synchronous stretching of all the double bonds in the chromophore. However, if that peak (1903 cm^{-1}) had the right intensity, both spectra would have the same structure. At very low

wavenumber values, the spectrum calculated at QM level presents two very weak peaks (480 and 563 cm^{-1}), which could correspond to the peaks at 233 and 540 cm^{-1} in the OLUFF-based one, since the relative intensities are similar. Consecutively, the QM spectrum presents two other weak peaks of similar intensity (1117 and 1180 cm^{-1}) that would correspond to the convolution of the next two peaks (833 and 901 cm^{-1}) and the following one (1004 cm^{-1}) of the OLUFF-based IR spectrum. At around 1300 cm^{-1}, the QM spectrum presents a peak of low intensity (1254 cm^{-1}), a peak of medium intensity (1284 cm^{-1}) and a peak of low intensity (1284 cm^{-1}). This structure of relative intensities is found in the OLUFF spectrum with high- and medium-intensity peaks (1333, 1434 and 1500 cm^{-1}). Moreover, before the very intense peak at the QM level, there are two other weak peaks at 1526 and 1551 cm^{-1} that also appear in the next part of the OLUFF spectrum (1738 and 1802 cm^{-1}) and, after that peak, another one (at 1699 and 1966 cm^{-1} in the QM and OLUFF spectra, respectively). Thus, the results from the S_1 show that, despite the difficulty of describing IR spectra with classical methodologies, the OLUFF-based spectrum reproduces reasonably well the QM vibrational features.

Additionally, the peaks corresponding to the movement of the protons are accurately described for all the chromophores with the new set of parameters as it can be observed in Figure C.10, where the same representation can be found for a wider range (up to 4000 cm^{-1}), since the position of the peaks is the same as in the QM spectra.

4.3.5 Principal Components Analysis

An additional vibrational analysis was performed to address the accuracy of the OLUFF. Figure 4.6 shows a comparison between the QM frequencies computed by DFT and the classical frequencies obtained by a PC analysis. As it can be observed, at lower values of the wavenumber (under 2000 cm^{-1}), the FF vibrations correlate linearly to the QM ones for all the OLU chromophores in both S_0 and S_1 states. Nonetheless, for the H–X normal modes that appear over 3000 cm^{-1} at the QM level, the wavenumber of the normal modes are strongly underestimated. This tendency appears for both the OLUFF and the GAFF, showing the difficulty to parametrize high-frequency and low-frequency normal modes simultaneously. However and fortunately, the X–H stretchings are very often constrained by algorithms such as the SHAKE and RATTLE ones[274, 275] in MD simulations because they are not relevant in the dynamics of the systems when structural features are analysed.

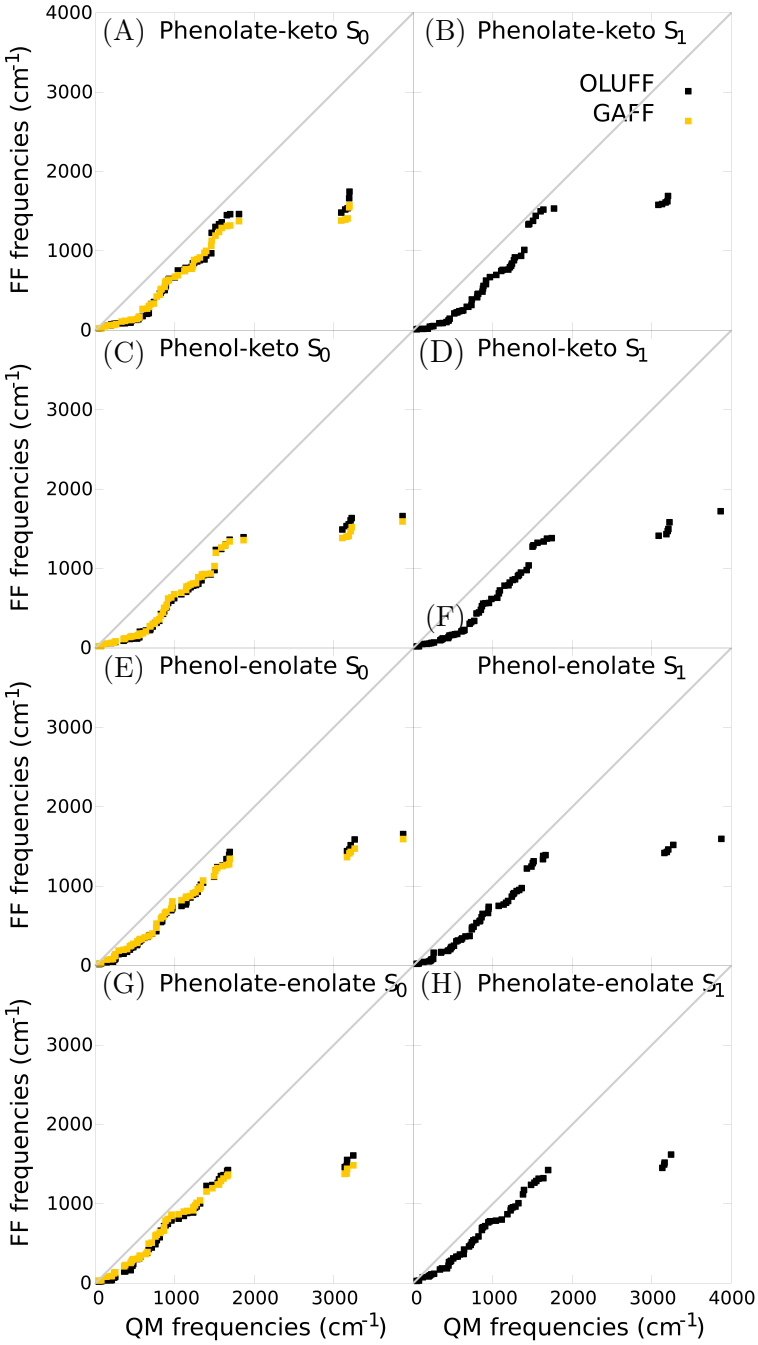


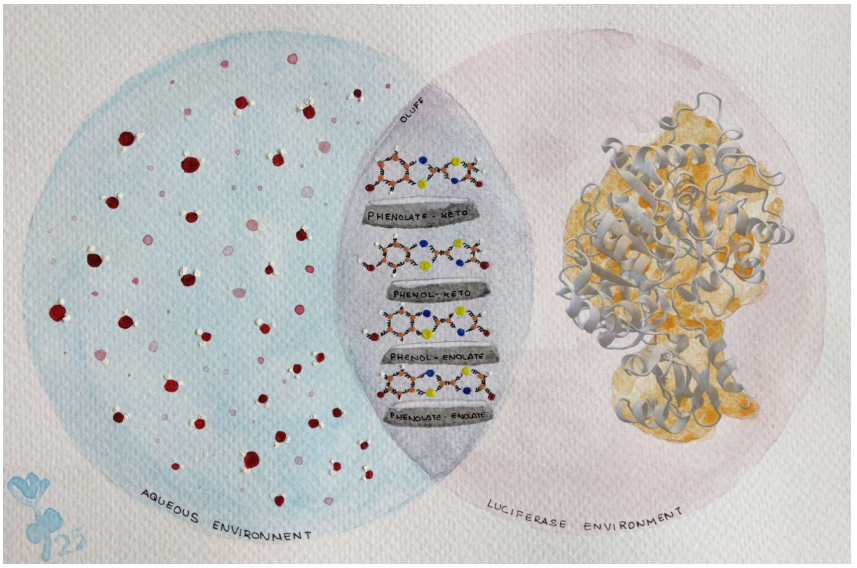
Figure 4.6: Comparison between the QM and OLUFF frequencies of the OLU (A, B) phenol-keto, (C, D) phenol-enolate and (E, F) phenolate-enolate of the (A, C, E) S_0 and (B, D, F) S_1 states calculated from the QM optimized geometries and from a PC analysis, respectively.

4.4 Conclusions

In order to deepen into the nature of the system and due to its large size, especially when interacting with the luciferase enzyme, it is important to develop models that are able to properly sample the conformational space of the system at low computational cost. This chapter presents a FF parametrization for all the four possible emitters of the chromophore in both the S_0 and S_1 states. These parameters can be used to perform classical MD simulations with each of the different chromophores inside the enzymatic media to properly sample the conformational space of each individual system. It is worth noting that the non-bonded cutoffs of the simulations should be of at least 11 Å, since the parametrization has been carried out *in vacuum*, considering all points charges of the OLU chromophores, and the largest intramolecular distance is a bit lower than this value. The parametrization is based on the Seminario method for bond and angle parameters, while the torsional potential around the single C–C' bond is fitted to QM energies using a periodic function with several terms. These parameters were validated by comparing the probability distribution functions of the dihedral angle, minimized structures, IR spectra and PC analysis computed from OLUFF-based MD simulations with analogous QM data. Based on these analyses, it is possible to conclude that the presented parametrized FF, the OLUFF, performs better than the GAFF and can be used to further investigate firefly's bioluminescent system and unravel the nature of the enzyme/substrate interactions.

Chapter V

Structure of the Oxyluciferin Emitting Species in Aqueous and Luciferase Media



Understanding the biological environment that surrounds the fireflies' BL chromophore is crucial to unravel the key factors that lead to the fine-tuning of the emitted radiation. In this chapter, the recently published OLUFF is used to sample the conformational space of the four possible OLU emitters in both the ground and first electronically excited states in aqueous media and inside the WT luciferase enzyme to describe the solvation and interaction with the protein. It is shown that, in an aqueous solution, the degree of solvation agrees with the magnitude of the atomic charges on the electronegative heteroatoms for both biological environments. However, in the enzymatic environment, the hydrogen bonding pattern with solvent molecules changes because it competes with direct hydrogen bonds with the protein and the AMP co-factor. These interactions can also be facilitated in some cases via hydrogen bond bridging through water molecules.

* This chapter is based on the preprinted article *Structural Analysis of the Oxyluciferin Emitting Species in Aqueous and Luciferase Media* with contributions of Henar Mateo-delaFuente[a] and Juan J. Nogueira.[a,b]

[a] Chemistry Department, Universidad Autónoma de Madrid, 28049 Madrid, Spain.

[b] Institute for Advanced Research in Chemical Sciences (IAdChem), Universidad Autónoma de Madrid, 28049 Madrid, Spain.

Preprint: 10.5281/zenodo.15266981

5.1 Introduction

Having developed the OLUFF, it is necessary to evaluate its behaviour not only *in vacuum*, a media in which this chromophore is never going to be found, but also in aqueous and the enzymatic media, which are the natural environments of the OLU. Hence, this chapter aims to shed some light into the structural behaviour of the four possible emitting chromophores in two different environments, an aqueous solution and the inside the enzyme, and in two different electronic states, the ground state Ş and the emitting state S_1. Moreover, the effect of the protonation and electronic state of the OLU on its chemical environment has also been studied. It has been found that the solvation around the different electronegative atoms of the system, which can act as hydrogen bonding acceptors, is related to their atomic charges and is larger in the S_1 than in the ground state, consistently with previous results.[118, 340] Moreover, in the enzymatic media, the solvation is reduced due to the limited amount of water molecules inside the active site. In addition, the presence of direct hydrogen bonds with either the protein or the AMP co-factor further reduces this solvation although, this effect competes with the possibility of having bridging water molecules between the chromophore and its chemical environment.

5.2 Computational Details

This chapter studies the four possible OLU emitters in both the S_0 and S_1 states in two different media, an aqueous solvation and within the luciferase environment, by means of classical MD simulations based on potential gradients computed by the OLUFF.[264] The computational details of all these calculations are explained below.

5.2.1 System Setup

On the one had, the optimized structures of the OLU chromophores in each electronic state were taken from the previous chapter[264] and solvated with a truncated octahedral box of water molecules that ensures a solvent shell of at least 20 Å from any solute molecule. Moreover, Na^+ cations were added to the simulation box to neutralize it (one ion in the case of phenolate-keto and phenol-enolate and two ions in the case of the phenolate-enolate systems).

On the other hand, each of the OLU/luciferase complexes was built with PyMol[324] from the X-ray diffraction structure file of the Japanese firefly luciferase (PDB ID: 2D1R[124]), *Luciola cruciata*, by replacing the structure

of the crystallized OLU by that of each of the chromophores under study and modifying the AMP residue as to achieve a net charge of -1 by adding the corresponding H atoms missing in the crystallographic structure, that would correspond to its biological state. [356, 357] Consecutively, the system was solvated and protonated using the *tleap* module of AmberTools22[358] with water molecules within a truncated octahedral box, ensuring a solvent shell of at least 12 Å from any solute molecule. Moreover, the system was again neutralized by adding Na^+, the same amount as in the case of the aqueous media since the protonation state of the protein accounts for the negative charge of the AMP.

For both media, the water molecules and Na^+ cations were described by means of the TIP3P FF,[263] whereas the OLUFF[264] was employed to describe the different OLU chromophores in their respective electronic states. Additionally, the enzyme was described using the ff19SB FF.[262] As for AMP, the GAFF[261] was used to describe the dihedral and improper torsions and Lennard-Jones parameters selected by means of the Antechamber *parmchk2* module of the AmberTools22,[358] while the rest of the bonded parameters and the charges were obtained from QM calculations performed *in vacuum* at the B3LYP/6-31G* level of theory[231, 232, 342, 349–351] and the MSK scheme[325] for the atomic charges using the Gaussian09[341] software. As in the previous chapter, the geometry optimization directly provided the equilibrium distances and angles and the force constants were calculated by means of the Seminario method that consist on the block diagonalization of the QM Hessian matrix of the rows and columns of the atoms involved in a specific harmonic potential.[265]

5.2.2 Classical Molecular Dynamics

Consecutively to the system preparation, classical MD simulations were performed for each of the 16 prepared systems (4 OLU chromophores in the S_0 and S_1 states in water and protein environments) using the same computational protocol and the CUDA version of the software Amber22[358–361]. First, the system was minimized for 5000 steps by means of the steepest descent algorithm,[244] and for additional 5000 steps using the conjugate gradient algorithm.[245] Then, it was progressively heated for 500 ps from 0 to 300 K in the NVT ensemble using a 2 fs time step and an additional simulation was evolved at 300 K for another 500 ps in the same ensemble. The temperature was controlled by means of the Langevin thermostat[267, 268] with a 2.0 ps^{-1} of collision frequency. Afterwards, a production of 1 μs with a 2 fs time step and within the NPT ensemble was performed at 300 K and 1 atm. The tempera-

ture was also controlled by means of the Langevin thermostat with a 1.0 ps^{-1} of collision frequency and the pressure by means of the Berendsen barostat.[271] The particle-mesh Ewald[272] method was used to compute the electrostatic interactions with a grid spacing of 1.0 Å, and the non-bonded interactions were computed with a cutoff of 13 Å and a switching distance of 11 Å. The bonds involving hydrogen atoms were constrained using the SHAKE[274] algorithm. Exceptionally, the trajectory for the phenolate-keto chromophore in the S_1 state in the enzymatic medium was evolved up to 1.4 μs to ensure that the protein structure was completely equilibrated, as it can be observed in Figure D.1. Finally, the different structural analysis performed in this chapter have been carried out using the *cpptraj*[354] module of AmberTools22 by taking into account the 1 μs production in the case of the aqueous simulations and the last 600 ns of the MD trajectories in the case of the OLU/luciferase complex, in order to have an equilibrated structure for the protein.

All the necessary files to reproduce the simulations, including example *tleap* and MD inputs and the set of parameters employed for the AMP, can be found in Annex D and in the Zenodo repository DOI: 10.5281/zenodo.15132416.

5.3 Results and Discussion

As mentioned above, this chapter aims to explain the behaviour of the four possible OLU emitters of fireflies' BL system in two different biological media and electronic states. Thus, the results obtained in solvent and the protein are discussed separately and, then, compared in the latter section.

5.3.1 Aqueous Trajectories

Structure of the Oxyluciferin Chromophores

In aqueous solvation, all the OLU chromophores present a large freedom of movement around the C-C' single bond connecting both rings. This can be observed in Figure 5.1(A,B), which represents the probability distribution for the SCCS dihedral angle, the only torsional degree of freedom that is truly able to rotate. On the one hand, ground state simulations show that the keto tautomers mostly present a s-trans planar configuration (around 180º) for this dihedral angle, although the distribution for the neutral form (phenol-keto) is wider than for the monoanion (phenolate-keto). Moreover, the two enolate species present a non-planar distorted configuration with maxima at 144 and 216º and 132 and 228º for the phenol-enolate and phenolate-enolate OLU, respectively. This is not a surprising result considering that the OLUFF

is specifically parametrized to reproduce this behaviour.[264] In fact, all the SCCS probability distributions, represented in Figure 5.1, show similar results to those obtained *in vacuum* in the previous chapter. On the other hand, S_1 simulations show a different picture for all the chromophores with the exception of the phenol-keto OLU for which the distribution just becomes narrower. While the phenolate-keto chromophore in the excited state is also able to populate conformations for which the tz and btz moieties are perpendicular to one-another (90 and 270º), both enolate forms are trapped in a planar s-trans configuration. Thus, in the excited state the phenolate-keto OLU presents the larger mobility among the studied systems, as also reflected by their respective RMSD shown in Figure D.2.

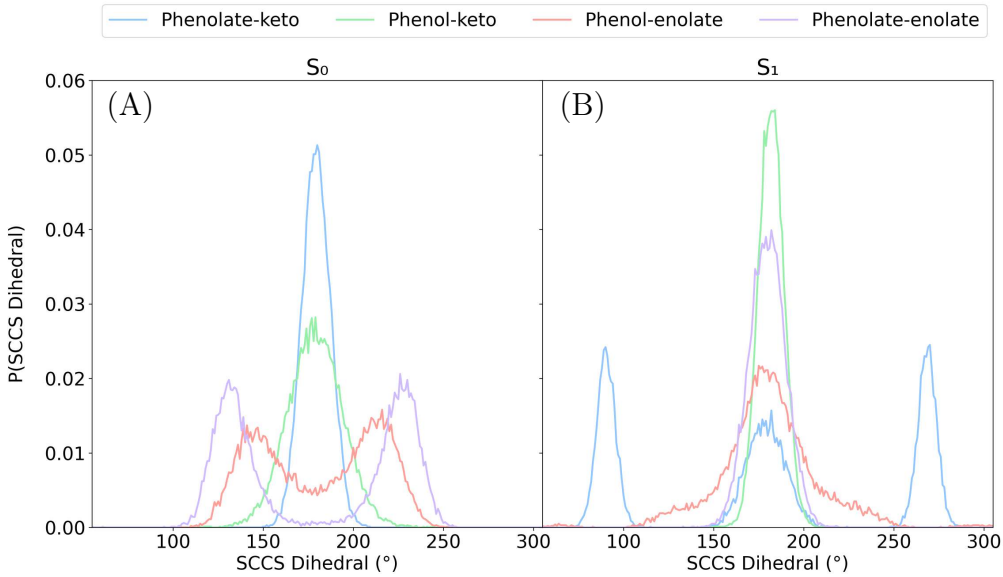


Figure 5.1: Probability distributions of the SCCS dihedral angles of each of the OLU chromophores in the (A) S_0 and (B) S_1 states in aqueous media.

Aqueous Solvation of the Oxyluciferin Chromophores

When analysing the solvation around the different OLU chromophores, it can be observed in Figure 5.2 that the trend of the RDF between the chromophores and water molecules in both electronic states coincides with the total charge of the chromophore and the atomic charge on the analysed heteroatom, taken from the OLUFF. In the case of the RDF around the oxygen atom of the btz moiety in the S_0 state (Figure 5.2(A)), the phenolate-enolate (with a total charge of -2 and a atomic charge of -0.724 on the O of the btz moiety) presents

the largest RDF followed by the phenolate-keto (with a total charge of -1 and an atomic charge of -0.623 on the O of the btz moiety). In addition, the phenol-enolate and phenol-keto chromophores (with atomic charges of -0.531 and -0.498 on the O of the btz moiety, respectively) are the OLUs with the lowest RDFs. The O of the tz moiety (Figure 5.2(C)) follows the same trend since both the phenolate-enolate and phenol-enolate (with atomic charges of -0.734 and -0.696 on the O of the tz moiety and total charges of -2 and -1, respectively) present the largest RDF while the keto OLUs present a much lower RDF. In the case of the simulations on the S_1 state (Figure 5.2(B)), a similar result can be observed for the solvation around the O of btz moiety. However, there is a significant decrease in the solvation around the tz O atom for all the chromophores, as seen in Figure 5.2(D), with the exception of the phenolate-keto (for which the RDF is larger than in the S_0) leading to a similar solvation to that of the two enolate chromophores. This result is due to the fact that it is the only OLUthat presents a more negative atomic charge on the tz O atom upon excitation. Moreover, the RDF around the two N atoms are not only very similar for all the emitters in each electronic state but also less relevant in comparison to those of the O atoms.

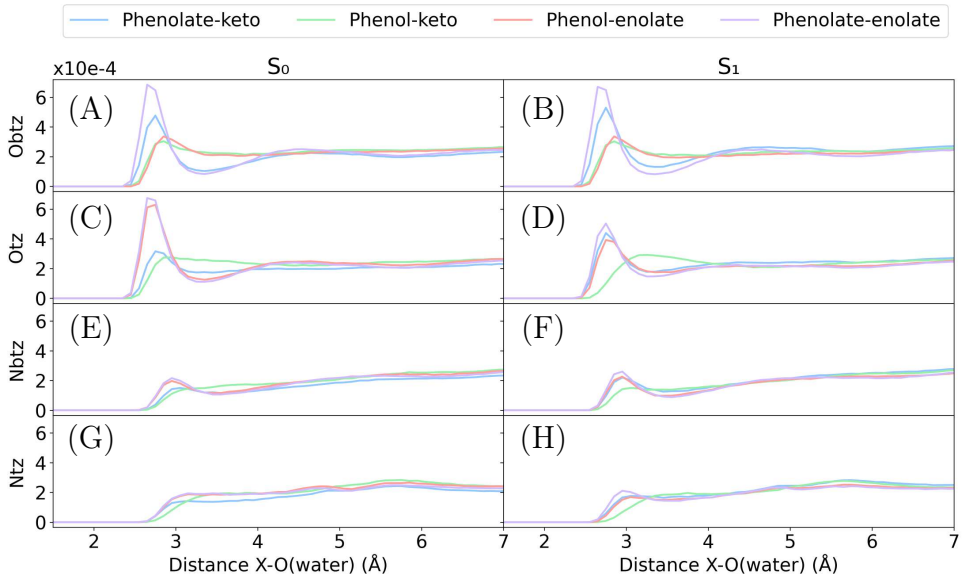


Figure 5.2: RDF of the O atoms from water surrounding each of the four electronegative sites of each of the OLU chromophores in the (A,C,E,G) S_0 and (B,D,F,H) S_1 states in aqueous media.

In agreement with the RDF results, and as it can be observed in Figure 5.3,

hydrogen bonding analysis using the default *cpptraj* angle and distance cutoffs (135º and 3.0 Å) between the hydrogen bonds donors and acceptors show that the larger the total negative charge of the molecule, the higher the total number of hydrogen bonds with the aqueous environment for all the studied hydrogen bond acceptors. This is not surprising considering that hydrogen bonding has an important electrostatic nature.[362] Comparing S_0 and S_1 hydrogen bonding patterns, it can be observed that most systems present a decrease in the total amount of hydrogen bonds in the S_1 state due to a decrease in the number of hydrogen bonds created between the solvent and the O atom of the tz moiety. This effect is especially stressed in the case of the phenol-enolate chromophore since, for the phenol-keto and the phenolate-enolate the decrease in the hydrogen bonding with the O is slightly compensated with an increase in the number of hydrogen bonds with the N atom of the btz moiety. Moreover, in the case of the phenolate-enolate OLU, the other N atom also presents a significant increase in the number of hydrogen bonds. In the case of the phenolate-keto emitter, a slight increase in the total number of hydrogen bonds can be observed when going from the S_0 to the S_1 due to an increase of the hydrogen bonding around both the O atom of the tz and the N atom of the btz moieties. The former is caused due to the fact that it is the only

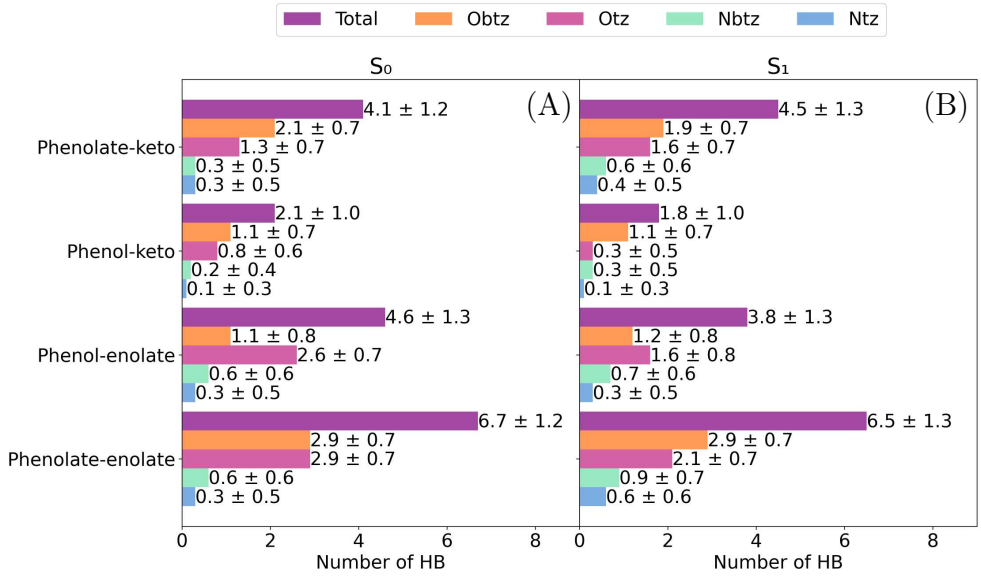


Figure 5.3: Average number of hydrogen bonds (HB) of each of the OLU chromophores with water and with each of the four electronegative atoms in the (A) S_0 and (B) S_1 states in aqueous media.

chromophore for which the FF charges of the tz O are more negative in the S_1 than in the S_0, while the latter is a general effect for all the OLU species, since all of them presents an increase in the atomic charge of this N atom upon excitation.

5.3.2 Trajectories of the OLU/luciferase Complex

Structure of the Oxyluciferin chromophores

In aqueous media the chromophores are allowed to move and rotate freely, but within the active pocket of the enzyme it is expected that this movement is constrained due to different factors such as steric hindrance and enzyme-substrate interaction. Thus, the structural properties of the OLU should be affected by the biological media in which it is embedded. While in aqueous media the chromophores were able to rotate around the SCCS dihedral angle, this torsion is not accessible inside the pocket. As can be seen in Figure 5.4, most of the chromophores adopt an s-trans conformation (with a SCCS dihedral angle around 180^{o}) specially for the S_1 state. Exceptions to this arise in the S_0 trajectories. The phenol-keto OLU, that presents an s-trans

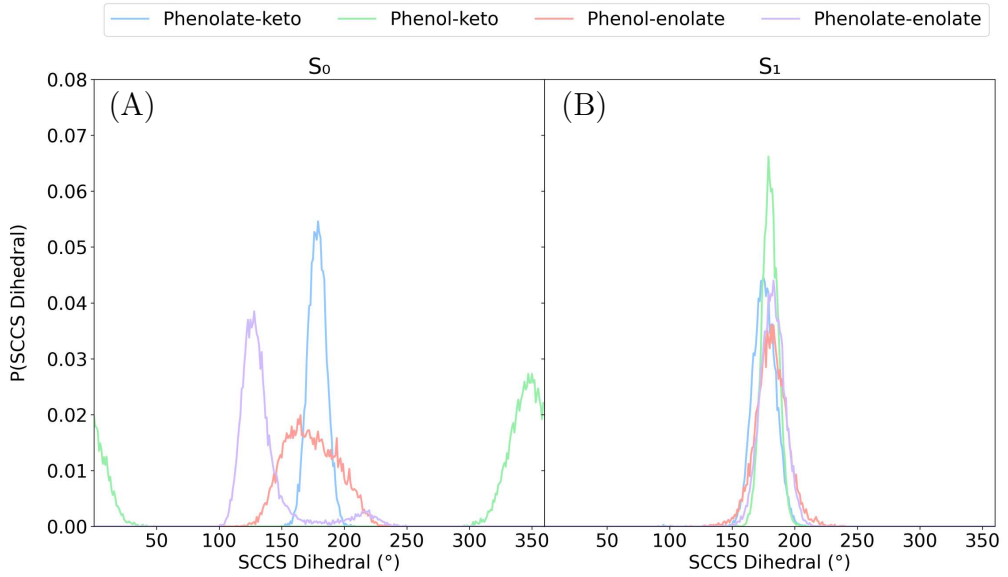


Figure 5.4: Probability distributions of the SCCS dihedral angles of each of the OLU chromophores in the (A) S_0 and (B) S_1 states in the enzymatic environment.

conformation at the beginning of the production of the trajectory, rotates around this torsion angle within the first 400 ns of the simulation; which, as previously mentioned, are not considered for the analysis since the trajectory need to be equilibrated. From then on, the chromophore remains trapped in the s-cis conformation, with a SCCS dihedral angle around 0^{o}. This motion is mediated by three water molecules and different hydrogen bonding with the surroundings of the chromophore, as explained in Annex D.3. Moreover, the phenolate-enolate is forced into one of the two distorted conformations shown in aqueous media (that around 127^{o}) and the phenol-enolate, in contrast with the aqueous situation, presents a wide distribution that corresponds to a planar s-trans conformation.

The role of water molecules

When analysing the solvation of the chromophores within the active site, the first evident difference with respect to the solvated systems in absence of the protein is the restricted amount of water molecules inside the pocket, which leads to RDFs one order of magnitude less intense than those in solvent, as seen in Figure 5.5. The RDFs of water molecules around the different heteroatoms of the OLU species in the protein present some additional differences that can be explained by taking into account the nature of the system. The O atom of the tz moiety of S_0 enolates completely loses its solvation (Figure 5.5(C)). In the case of the phenolate-enolate, this is due to the hydrogen bonding with both the AMP co-factor (Figure 5.10(C)) and the LYS531 (Figure 5.8(A)) - numbering refereed to that of the 2D1R PDB - while the phenol-enolate only presents hydrogen bonding with the protein (SER201, HIE247 and LYS531) through this hydrogen bond acceptor. Moreover, the complete lost of solvation around the tz O atom of the phenolate-keto is due to the high amount of hydrogen bonding with the AMP co-factor. In the RDF around the O atom of the btz moiety of S_1 (Figure 5.5(B)) it can be observed a relative decrease for the neutral OLU and the two enolates with respect to the phenolate-keto chromophore that can be explained with the hydrogen bonding with the protein, specifically with the ARG339 in the case of the dianion and with SER316 in that of the to phenol forms (Figure 5.8(B)). Additionally, not only the RDF around the O atom of the tz moiety of S_1 is lower than in aqueous media in general for all the chromophores, but also the ordering for the three anionic forms, which are the ones with a relevant contribution in both media, is reversed (Figure 5.5(D)). While in water the phenolate-enolate dianion presents the highest RDF and the phenol-enolate the lowest, inside the protein the chromophores with the largest and smallest RDF are the phenol-enolate

and phenolate-enolate OLU, respectively. In the case of the phenolate-enolate OLU, the RDF decrease is caused by hydrogen bonding with both ALA350 and SER349, while the phenolate-keto presents some hydrogen bonding with the co-factor and the HIE247. Moreover, the phenol-enolate presents bridging water molecules between the O atom of the tz moiety and THR345, GLY318, HIE247 and the AMP co-factor that might be stabilizing the water molecules around this hydrogen bond acceptor (Figure 5.9(B)).

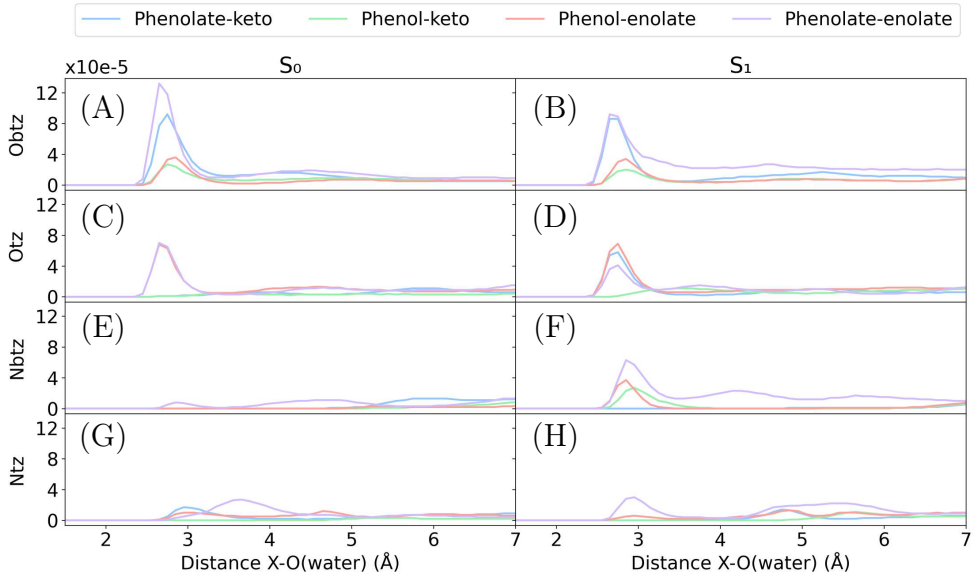


Figure 5.5: RDF of the O atoms from water surrounding each of the four electronegative sites of each of the OLU chromophores in the (A,C,E,G) S_0 and (B,D,F,H) S_1 states in the enzymatic environment.

When comparing the RDF around the N atoms for the S_0 simulations with respect to the aqueous situation, it can be observed some relative loss of solvation around them due to a restricted amount of water molecules, which prefer to form hydrogen bonds with the O atoms. Nonetheless, it can be seen some retention of the solvation structure of phenolate-keto and phenol-enolate species in the S_0 (Figure 5.5(G)) due to water bridging with SER316 (Figure 5.9(A)) and AMP (Figure 5.10(E)), respectively. This results will be explained in detail in a later section. In the case of the S_1 simulations, the RDF around the N atom of the btz moiety (Figure 5.5(F)) is maintained for all the chromophores but for the phenolate-keto, which is the only one that does not present bridging water molecules with the protein through this atom (Figure 5.9(B)). Furthermore, in the case of the solvation around the N atom of the tz

moiety in the S_1 trajectories (Figure 5.5(H)), only the phenolate-enolate OLU retains a significant solvation. Moreover, the presence of few bridging water molecules between the AMP and the N atom of the tz moiety of the phenol-enolate chromophore leads to a small retention of the solvation structure in the form of a very small peak in its RDF.

Using the same reasoning as for the RDF, the discrepancies between the number of hydrogen bonds in aqueous media and inside the cavity of the protein can be explained (Figure 5.6). As mentioned before, the limited amount of water molecules inside the active site reduces the total number of hydrogen bonds with water for all the studied systems inside the enzyme and the difference in the trends for the chromophores in each electronic state can be rationalized with the direct hydrogen bonding with the protein and the presence of bridging water molecules with either the protein or the AMP co-factor. While the presence of direct hydrogen bonding with the protein reduces the amount of hydrogen bonds with the water molecules, the possibility that water acts as a hydrogen bond bridge produces the opposite effect. This will be analysed in detail later.

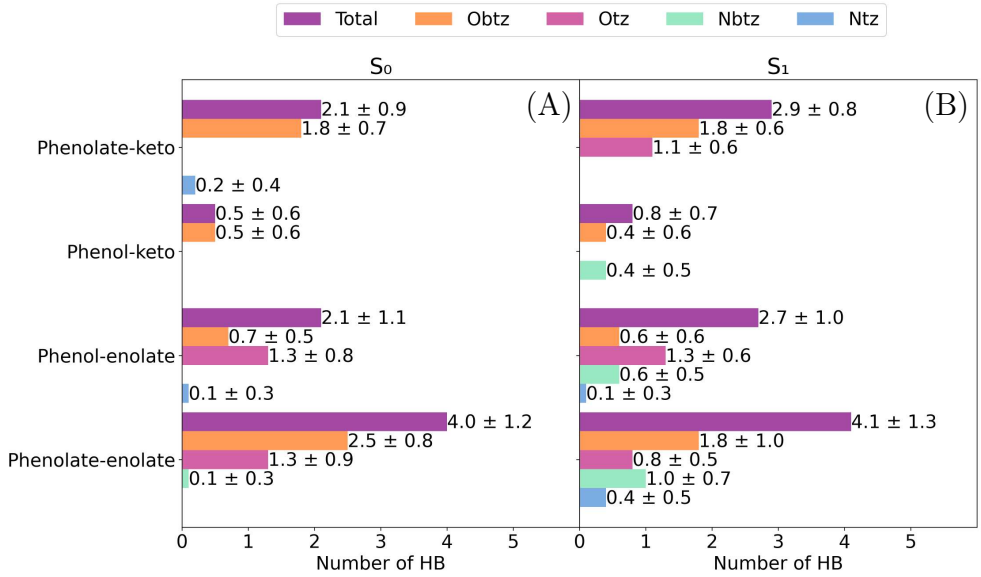


Figure 5.6: Average number of hydrogen bonds (HB) of each of the OLU chromophores with water and with each of the four electronegative atoms in the (A) S_0 and (B) S_1 states sites in the enzymatic environment.

OLU/luciferase interactions

To quantitatively assess the interactions between the different OLU chromophores in both electronic states and the protein, the amino acids with a larger amount of total contacts per residue (native and non-native) with the OLU species have been identified by means of the *nativecontacts* command implemented in *cpptraj* using a distance threshold of 5.0 Å. Figure 5.7 represents the percentage of time with contacts for those amino acids that are in contact with one of the studied chromophores for at least the 10% of the time, thus including only the relevant amino acids. On the one hand, S_0 trajectories (Figure 5.7(A)) show that the most relevant amino acid in terms of number of contacts with the chromophores is the PHE249. Moreover, it can be observed that the two phenol forms present a larger amount of contacts then the phenolate OLUs, with the exception of THR253, PHE252 and ARG220 for which the phenolate keto tautomer presents a significant amount of contacts while the other three chromophores do not. In addition, ALA350 and SER349 also have a similar amount of contacts with all the systems, although in a lower quantity. Furthermore, both phenol forms, and also the phenolate OLU to a lesser extent, present a significant amount of contacts with GLY343 and GLY341, for which the phenolate-enolate chromophore presents a much lower amount of contacts. Additionally, GLY318 and HIE247 present a significant amount of contacts with both enolate forms.

On the other hand, the S_1 state (Figure 5.7(B)) presents a different picture. While PHE249 still presents the highest amount of contacts with most of the studied chromophores, this is not the case for phenolate-enolate OLU for which the number of contacts with this amino acid and with ALA350, SER349, THR345, LEU344, GLY343, TYR342, GLY341, GLY318, GLY317, SER316, GLY248 and HIE247 is reduced in favour of contacts with ILE353, THR253, ASN231, GLY230 and ARG220, leading to a somehow more polar environment. Phenolate-keto OLU also presents a different contacts pattern, favouring the contacts with SER349, THR345, LEU344, GLY343 and PHE249 to the detriment of ALA350 TYR342 and specially THR253, PHE252 and ARG220, which means a local environment with more hydrophobic and aromatic amino acids. On the flip side, the two phenol forms, present a more similar contacts pattern with respect to the S_0. In the case of the phenol-keto OLU the most important amino acids are still ALA350, SER349, GLY343, TYR342, GLY341 and the already mentioned relevant PHE249. However, the number of contacts with ILE353, PHE252, and ARG220 is slightly reduced, and with GLY318 slightly increased. For the last of the chromophores, the phenol-enolate OLU, the number of contacts with any of the amino acids is

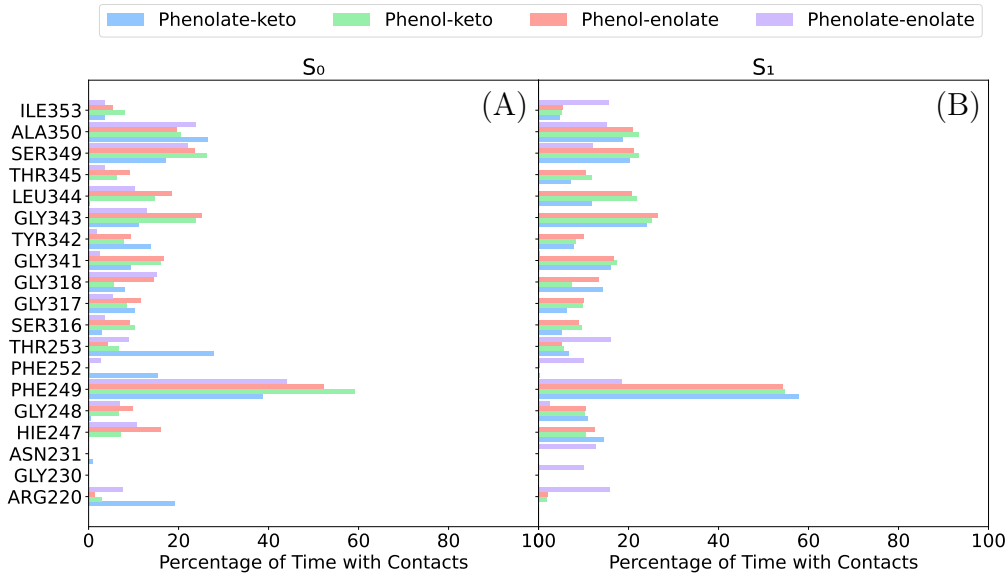


Figure 5.7: Percentage of time with contacts between the four OLU chromophores in the (A) S_0 and (B) S_1 states and specific amino acids of the enzyme along the MD trajectory. Only those amino acids with contacts for at least the 10% of the simulation time have been included in the representation.

significantly affected by the electronic state of the molecule.

In general terms, for all the studied systems in both electronic states, the most significant amino acid in terms of contacts is PHE249. The only exception to this would be the phenolate-enolate in the S_1 for which the reduction in contacts with PHE249 is compensated with an increase in contacts with other amino acids leading to a local environment of a higher polarity. In the case of the phenolate-keto OLU the change of contacts from the S_0 to the S_1 state leads to more hydrophobic pocket with more aromatic amino acids. Finally, both phenol forms present a similar environment in both electronic states since differences in contacts between them are very small.

To better understand the nature of the OLU/luciferase interactions, the direct hydrogen bonding between each OLU and each amino acid of the protein has been analysed using the default cutoffs for hydrogen bonding implemented in *cpptraj*. Figure 5.8 represents the averaged amount of hydrogen bonds for those amino acids that present at least an average of 0.1 of hydrogen bonds in at least one of the studied systems. In general, the OLU chromophores do not present a high amount of direct hydrogen bonding with the protein. In the S_0, the phenolate-keto form presents some hydrogen bonding between the

O atom of the tz moiety and GLY318; the phenol-keto form, between the O atom of the btz moiety and SER316 and ARG339; the phenol-enolate form, between the O atom of the tz moiety and SER201, HIE247 and LYS531 and between the O atom of the btz moiety and SER316; and the phenolate-enolate form has hydrogen bonding between the O and N atoms of the btz and the O atom of the tz moieties and ARG339, ALA350 and LYS531, respectively. In contrast, in the S_1 state, the phenolate-keto form presents some hydrogen bonds between the O atom of the tz and the N atom of the btz moieties and HIE247 and SER349, respectively; the phenol forms have hydrogen bonding only with the SER316 but in an opposite ratio with respect to the S_0, and the phenolate-enolate OLU presents hydrogen bonding between the O atom of the btz moiety and ARG339 and between the O atom of the tz moiety and SER349 and ALA350.

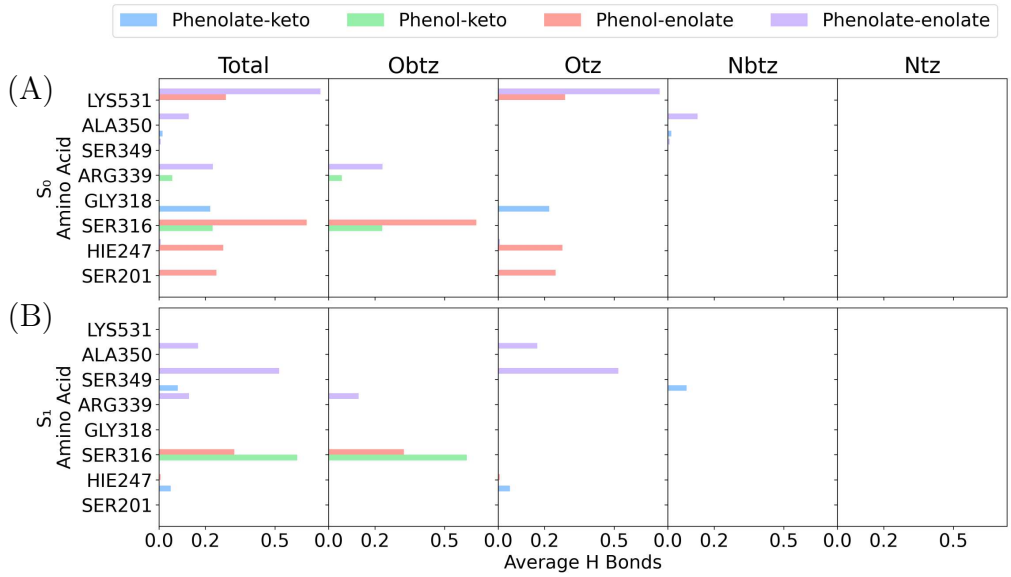


Figure 5.8: Average of hydrogen bonds (HB) between each of the four OLU chromophores in the (A) S_0 and (B) S_1 states and specific amino acids of the enzyme along the MD trajectory

In addition to the direct hydrogen bonding with the luciferase, it is also crucial for the understanding of the interactions of the OLU with the protein to characterize the amount of bridging water molecules connecting the chromophore and the enzyme through hydrogen bonding. For this purpose, Figure 5.9 represents the percentage of time that a water molecule is acting as a hydrogen bond bridge between the chromophore and a specific amino acid.

A protein residue is considered only if the interaction is present at least 20% of the simulation time. As it can be observed, each chromophore presents an unique hydrogen bond bridging pattern with the luciferase and, in addition, this pattern is different for each electronic state. In the case of the phenolate-keto OLU in the S_0 state, there is a significant amount of hydrogen bond bridging with the ARG220 and ASN231 through the O atom of the btz moiety and with SER316 through the N atom of the tz moiety. The phenol-enolate chromophore presents a significant amount of water molecules acting as hydrogen bond bridge between ARG339 and the O atom of the btz moiety and a smaller amount between HIE247, SER316, GLY318, THR345 and THR529 and the O atom of the tz moiety, except for SER316 which is connected via hydrogen bond bridging with the O atom of the btz moiety. Furthermore, the phenolate-enolate OLU has two very important hydrogen bond bridges con-

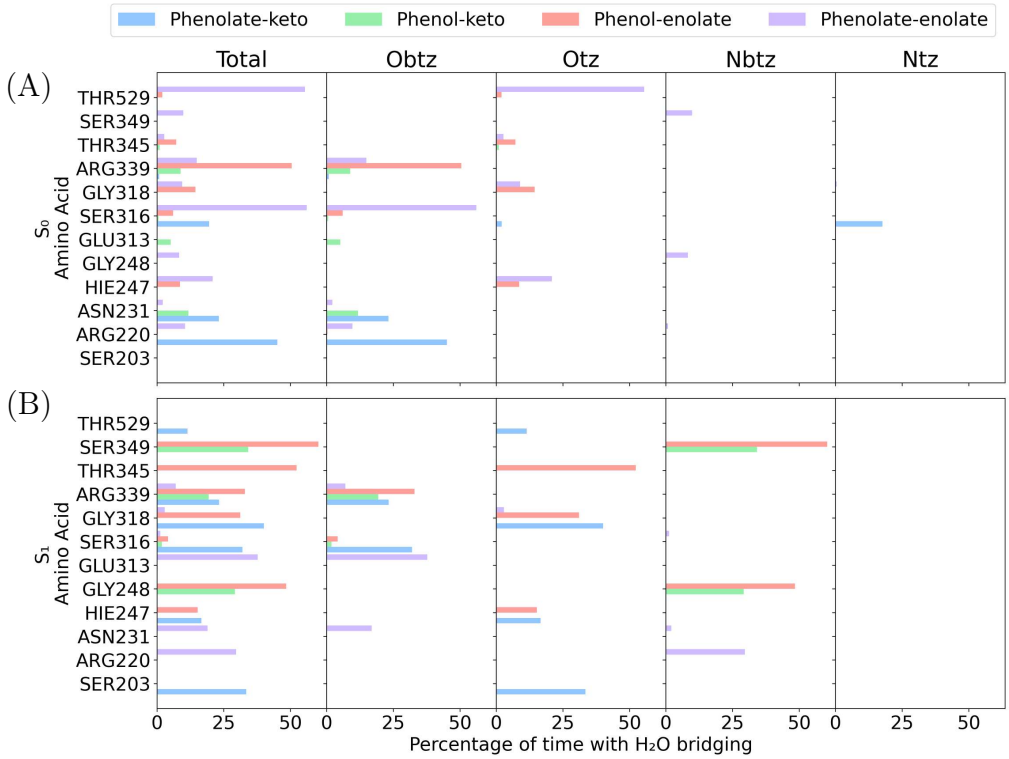


Figure 5.9: Percentage of time with a water molecule acting as a hydrogen bonds (HB) bridge between each of the four OLU chromophores in the (A) S_0 and (B) S_1 states and specific amino acids of the enzyme along the MD trajectory.

necting SER316 and THR529 with the O atom of the btz and the tz moieties, respectively. Moreover, the phenol-keto form does not have a very important hydrogen bond bridging pattern in the S_0, even though in the excited state it presents relevant bridges connecting GLY248, ARG339 and SER349 with the N or O atoms of the btz moiety. In the S_1 electronic state, the phenolate-keto OLU presents an increased amount of hydrogen bond bridging with SER316 with respect to the S_0 but through a different OLU heteroatom, the O atom of the btz moiety. Moreover, hydrogen bond bridging with ARG220 and ASN231 completely disappears and it is replaced with additional hydrogen bond bridges between SER203, HIE247, GLY318, ARG339 and THR529 and the O atoms of the tz moiety except for ARG339, that presents the hydrogen bond with the O atom of the btz moiety. In addition to the relevant bridging with ARG339 that the phenol-enolate presents in the ground state, in the S_1 this chromophore shows a larger hydrogen bond bridging pattern. Relevant contributions appear linking GLY248, GLY318, THR345 and SER349 with different hydrogen bond acceptors of the chromophore. In the case of the phenolate-enolate in the S_1, the hydrogen bond bridging pattern is produce mostly between the hydrogen bond acceptors of the btz moiety and ARG220, ASN231 and GLU313.

In general terms, it can be observed that all the studied OLUs presents a larger bridging pattern in the S_1 state than in the ground state, except for the phenolate-enolate chromophore, mostly due to extra bridging points with the O atom of the tz moiety and the N atom of the btz one.

The role of AMP

Finally, it is interesting to analyse the role of the AMP co-factor inside the active site. As it can be observed in Figure 5.10(A,B), the distance between the centre of mass of the OLU and the AMP oscillates between 7.5 and 11 Å. However, it does not correlate with the interaction with the chromophore (Figure 5.10(C-F)). In the ground state, AMP presents a large amount of direct hydrogen bonding with the O atom of the tz moiety of both phenolate ligands, while it requires an intermediate water molecule to interact with both hydrogen bond acceptors of the tz moiety of the phenol-enolate OLU. In the excited state, none of the chromophores present a significant hydrogen bonding with the AMP co-factor and only the phenolate-keto and the phenol-enolate are able to interact with the AMP via hydrogen bond bridging through the O atom of the tz moiety.

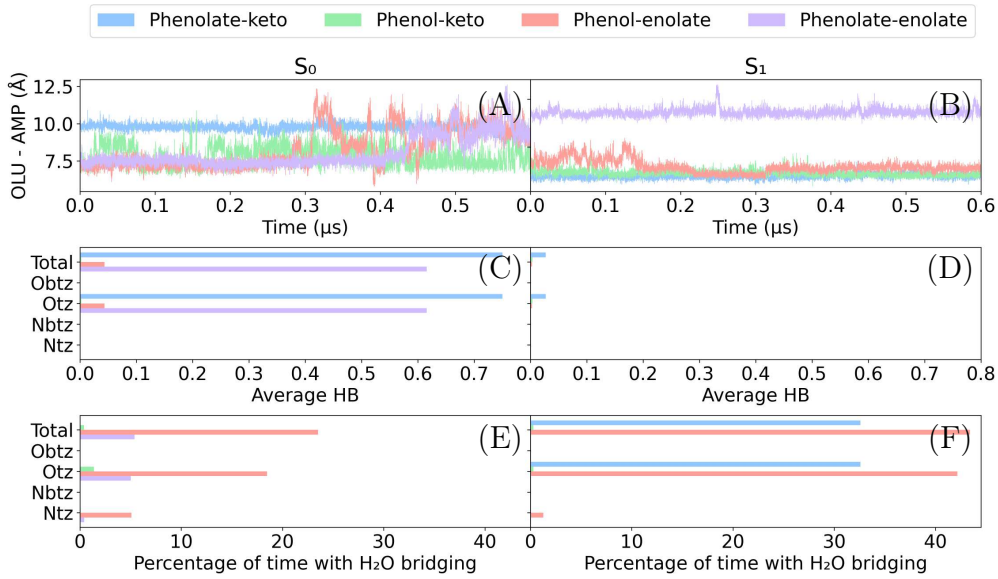


Figure 5.10: (A,B) Distance between the OLU chromophore to the AMP co-factor, (C,D) average of hydrogen bonds (HB) between each of the four OLU chromophores and the AMP and (E,F) percentage of time with a water molecule acting as a HB bridge between each of the four OLU chromophores and the AMP for the MD trajectories in the (A,C,D) S_0 and (B,D,F) S_1 states of each chromophore.

5.4 Conclusions

This chapter presents an extensive structural and intermolecular interaction analysis on two different biological environments, water solvation and the WT luciferase enzyme, of the four possible emitting forms of the OLU in both the S_0 and the S_1 states. The OLUFF, presented in Chapter IV, was employed to sample the conformational space of the different chromophores by MD simulations in order to determine their flexibility around the SCCS torsion, and to characterize the intermolecular interactions with the water molecules and protein residues through contact and hydrogen bonding analyses.

In aqueous media, the OLU presents similar probability distribution around the SCCS dihedral angle than that presented in the previous chapter[264] for the gas phase due to a large flexibility for this degree of freedom. However, inside the active site the movement is sterically constrained and the chromophores can only populate the region near the s-trans conformation. The phenol-keto OLU in the ground state presents an exception to this. At the

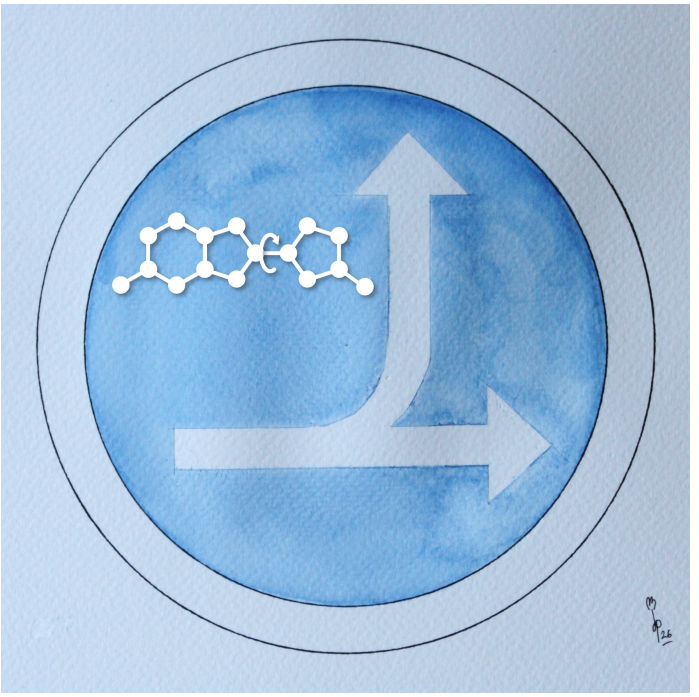
beginning of the simulation, it leaves the s-trans conformation to populate the s-cis configuration of the SCCS torsion, where it is trapped for the whole analysed simulation time.

When looking at the solvation of the hydrogen bond acceptor atoms of the OLU in aqueous media, the larger the negative atomic charge on the heteroatom, the larger the solvation. On the flip side, in the pocket of the protein, the reduced amount of water molecules leads to a lower solvation. Moreover, the solvation pattern is different from that obtained in water, as a result of the competition with direct interactions with the protein and AMP, including hydrogen bonding and bridging interactions through water molecules. In addition, non-electrostatic interactions were also analysed by a contact analysis, where the relevant amino acids have been identified. Specifically, PHE249 has been shown to be the most relevant residue for both electronic states of the chromophores.

The results shown here could be the basis for future computational or experimental work aimed at mutating the luciferase enzyme to modify the behaviour of the chromophores inside the pocket and potentially control its emitting properties. However, to achieve such a challenging goal, further research is needed to understand first the influence of the variation of hydrogen bonding and contact patterns on the photophysics of the system. To this aim, the conformational space of the phenolate-keto OLU chromophore, considered to be the red-emitting species, in the S_1 state is explored in Chapters VI and VIII to, respectively, (i) determine, *in vacuum*, the correlation between its nuclear motion and the emission energy and (ii) study, analogously to the present chapter, the structural differences between the WT and its SER286ASN mutant. Moreover, the classical sampling calculated for the current chapter has also been employed in Chapter VII to study the OLU/luciferase interaction in the S_1 state.

Chapter VI

Characterization of the Nuclear Coordinate for the Phenolate-Keto Oxyluciferin Red-Shift Emission from Quantum Mechanical Calculations



Despite the huge effort that the scientific community has put into understanding the spectral shifts of wild OLU/luciferase systems, it is still unclear which factors may be playing a relevant role on the colour of the emitted light. In order to shed some light into this conundrum, this chapter has characterised a nuclear motion that leads to the shift of the emission energy for for each of the three SCCS minima of the phenolate-keto OLU in the S_1 electronic state. This energy shift dimension, which has been expressed as a linear combination of the PCs of the system, shows that the energy difference between the S_0 and S_1 electronic states that lays within the red region of the electromagnetic spectra only appears for the two twisted conformations since, even though the s-trans dimension can be stretched over 600 nm, that region of the energy shift dimension can not be sampled due to an asymptotic energy barrier. Although it was typically believed that the chromophore could not emit from these twisted conformations due to their high CT character, the results presented along this chapter suggest that there is a vibrational motion that favours the increase of the oscillator strength thus hijacking the dampening of the intensity caused by the CT state.

* This chapter has contributions of Henar Mateo-delaFuente,[a] Adam Šrut,[b] Juan J. Nogueira[a,c] and Vera Krewald.[b]

[a] Chemistry Department, Universidad Autónoma de Madrid, 28049 Madrid, Spain.

[b] Department of Chemistry, Theoretical Chemistry, TU Darmstadt, 64287 Darmstadt, Germany.

[c] Institute for Advanced Research in Chemical Sciences (IAdChem), Universidad Autónoma de Madrid, 28049 Madrid, Spain.

6.1 Introduction

The microenvironment mechanism, first proposed by Ugarova and co-workers, [162, 167] suggests that the emission maxima of firefly's BL system is affected by the polarity of the medium. While it is true that this effect can lead to changes in the emitted colour, it has been demonstrated that even strong variations in the polarity of the solvent, from very polar (water) to non-polar (toluene), do not produce the red-shift observed in the luciferase environment.[119] Hence, it is clear that, regardless of its surroundings, there should exist a conformational change or internal coordinate of motion yielding the red-shifted emission. In agreement to this, very accurate *ab initio* QM/MM calculations, multi-layer fragment molecular orbital (MLFMO) - partially renormalized configuration interaction singles with perturbative doubles (PR-CIS(Ds))/6-31G, on the red-emitting ILE288ALA and ILE288VAL mutants were not able to reproduce the exact emission energy nor the two peaks structure (at 560 and 613 nm) of the latter.[363] However, following the exact same procedure, the authors obtained a very good agreement with the experimental emission energies, which suggests that structure that they considered for the mutants might not be the red-emitting conformation.

In order to investigate the structural changes of the OLU chromophore that lead to very different emission energies (from 560 nm in the WT luciferase to, for instance, 605 nm in the SER286ASN mutant),[124] this chapter employs the sampling generated by MD simulations in Chapter IV for the phenolate-keto OLU in the S_1 state to determine the nuclear coordinate that yields the shifts in the energy of emission. This nuclear coordinate can present contributions from different modes of movement, which in principle do not need to be chemically intuitive. Thus, inspired by the previous work of Šrut et al.,[364] the energy shift coordinate has been expressed as a linear combination of the PCs of the MD nuclear motion of the system. These contributions have been obtained from a multi-component fit to the emission energy for each of the three populated conformations of the SCCS angle (at 90, 180 and 270º). It is shown that only the two twisted conformations are able to sample regions of the S_1 PES that lead to a wavelength of emission larger than 600 nm. These regions, which were previously believed to be dark due to a strong CT character, are demonstrated in this chapter to be bright due to a combined effect of the vibrational motions of the system.

6.2 Computational Details

In order to understand the intramolecular structural changes of the OLU chromophore that lead to the differences in the emission energies, a classical MD simulation was employed to sample the conformational space of the phenolate-keto OLU in the S_1 state using the OLUFF, developed in Chapter IV. Then, a PC analysis was performed to characterise the different motions of the system, whose projections on the MW nuclear displacements along the sampled geometries were fitted to the emission energy to determine the nuclear coordinate that leads to the red shift of the system. Since this analysis relays on the harmonic approximation and the torsion is represented by a periodic potential, the trajectory was divided on three sampled regions of the SCCS dihedral angle of the S_1 PES and the analysis were performed for each of them independently.

6.2.1 Classical Molecular Dynamics

As described in Section 4.2.2, the QM optimised phenolate-keto OLU in the S_1 state was classically minimized *in vacuum* by means of 5000 steps by the steepest descent algorithm[244] followed by 5000 steps using the conjugate gradient algorithm.[245] Consecutively, the system was progressively heated in the NVT ensemble for 500 ps from 0 to 300 K and stabilized for another 500 ps at 300 K. Finally, a production of 1 μs within the NPT ensemble was performed. These simulations employed the OLUFF set of parameters for the phenolate-keto OLU in the S_1 state. Furthermore, in both the heating and the production, the time evolution of the system was simulated by means of the velocity Verlet algorithm[253] with a 2 fs time step and the bonds involving hydrogen atoms were constrained using the SHAKE algorithm,[274] as the H movement is not considered to be relevant for the red shift. Moreover, pressure and temperature were controlled by means of the Berendsen barostat[271] and the Langevin[267, 268] thermostat with a collision frequency of 2.0 and 1.0 ps^{-1}, for the heating and the production respectively.

6.2.2 Principal Components Analysis

Analogously to QM normal modes calculated from the diagonalization of the Hessian matrix in the minimum energy geometry of a system, the PC analysis allows for the characterization of the sampled movements (the PCs) along a trajectory. These PCs can be obtained from the diagonalization of the covariance matrix of the MW displacements along the dynamics, as given by

Equation 2.220. Nonetheless, the PCs are sensible to the region of the PES in which they are calculated. Hence, before the PC analysis was performed, the trajectory was divided into the three SCCS sampled conformations (at 90, 180 and 270º, were the 180º SCCS angle corresponds to the s-trans conformation) by means of a K-means clustering algorithm,[337] which analysed the trajectory every 10 snapshots starting from a random frame as implemented in the *cpptraj* module of AmberTools22.[358] Moreover, as a result of the clustering analysis, it was possible to obtain the centroid of each sampled region, which was then used as a reference structure for the PC analysis.

Before the PC analysis all the structures within each conformation were aligned to their reference structure in order to avoid translational and rotational motions. Then, for each of the conformations, the MW PCs were calculated considering only the heavy atoms of the OLU chromophore. Furthermore, in order to determine the amount of frames that are required to reach a converged sampling on all the i PCs, the MW PC were projected on the MW displacements along the dynamics, see Equation 6.1 where p_i represents the projection, N is the number of heavy atoms, C_{ij} is the transformation matrix of the MW covariance matrix after diagonalization and x_j and x_j^{ref} represent, respectively, each of the coordinates of the heavy atoms of the system and the reference structure. As it can be observed in Section E.1, a proper sampling of the PCs of each conformation is achieved after 1000 sampled frames for all the studied conformers.

$$p_i = \sum_{j=1}^{3N} C_{ij}^T M_{jj}^{1/2} \left(x_j - x_j^{ref} \right) \tag{6.1}$$

6.2.3 Fitting of Emission Energies

After completing the sampling, TD-CAM-B3LYP[231–233]/6-311G(2d,p)[313–315, 332, 333] calculations were performed for each of the 1000 randomly sampled frames for each of the conformations. These calculations, carried out by means of the Orca software v5.0.4,[365] were run considering 5 excited states and without the resolution of identity nor the TDA. Moreover, the CT character was also quantified for these frames by analysing the transition density matrix with the TheoDORe software.[249]

Then, the energy difference between the S_0 and the S_1 electronic states (ΔE), which was already converged after 300 frames as well as the rest of analysed electronic properties (see Figure E.22), was expressed, for each of the sampled frames, in terms of the projections of the MW displacements on each

of the MW PCs:

$$\Delta E = b_0 + \sum_{i}^{3N-6} b_i p_i \tag{6.2}$$

By optimising the coefficients b_i using a linear regression, it is possible to construct the dimension that leads to changes on the emission energy as a linear combination of the MW PCs:

$$\text{E Shift Dimension} = \sum_{i}^{3N-6} b_i \boldsymbol{PC}_i \tag{6.3}$$

6.2.4 Scan along the Energy Shift Dimension

Once the energy shift dimension was determined, a rigid scan along it was calculated. In order to do so, first, the MW was removed from the dimension by dividing each of the coordinates by the square root of the mass of the corresponding atom and renormalising. Then, the displacements along the dimension were added to the reference structure for the heavy atoms of each of the conformations. These displacements were obtained by multiplying the energy shift dimension by values between -0.35 to 0.35 in steps of 0.05. Consecutively, the H atoms were placed at a reasonable position using the *tleap* Amber22 module[358] and the OLUFF library[264] as geometric reference and optimised in the S_1 state at TD-CAM-B3LYP/6-31G* level of theory[231–233, 342, 348–351] with Orca software v5.0.4[365] without using the resolution of identity and taking into account three excited states. This optimisation was performed to achieve a better placement of the H atoms, while the heavy atoms remained fixed. Finally, the electronic properties were calculated at a higher level of theory using again the Orca v5.0.4[365] and the TheoDORE[249] software packages. Specifically, the basis set was increased to 6-311G(2d,p),[313–315] which has been demonstrated to reproduce well the electronic properties of the system,[311, 326, 327] and the TDA was not used.

6.3 Results and Discussion

6.3.1 Fitting of Emission Energies

Distortions on the relative position of the nuclear coordinates of the system result in changes on its electronic properties. However, assigning an specific

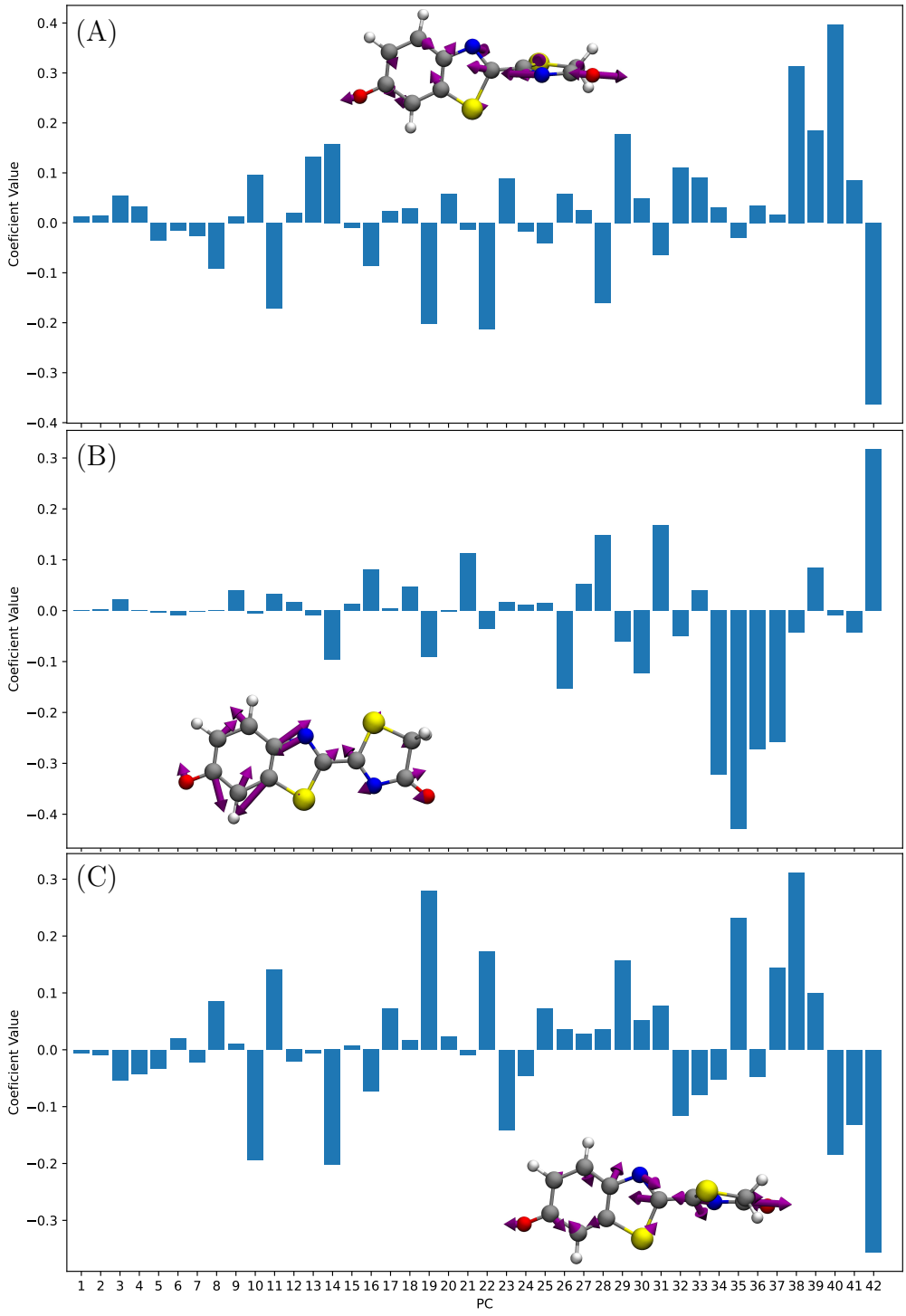


Figure 6.1: Fitting coefficients together with the resulting energy shift dimensions as linear combinations of the PCs with expansion coefficients obtained from fit for the (A) 90º, (B) s-trans and (C) 270º conformations.

distortion to a change on a certain property is not a chemically intuitive process. In order to achieve this task, the PCs of the system were linearly combined, as described in Section 6.2.3. Within the PC analysis framework, the PCs that present larger oscillations (see Figure E.23) appear as the first eigenvectors upon the diagonalization of the MW displacements covariance matrix. Moreover, as shown in Table E.1, they are associated with smaller vibrational frequencies. However, as it can be observed in Figure 6.1, in which the PC apperar in the same order; none of the studied conformations present large contributions from these motions to the energy shift coordinate. In fact, for all of them the most relevant contributions appear for PCs associated with larger vibrational frequencies, which lay in the region of the double bond tensions. As reflected by the in-plot graphical representation of the energy shift dimension, for all the conformations this coordinate consists on distortions in the plane of either the btz or the tz moieties of the chromophore without major contributions of out-of-plane bending motions.

Even though it was expected to find some contribution of the twisting of the SCCS dihedral angle, this motion was not found to be a relevant contribution to the energy shift because the rigid PES for this motion is symmetric and therefore both positive and negative displacements with respect to the reference structure on this direction would lead to either a decrease in the emission energy in case of the s-trans conformation or to an increase in the emission energy in case of the two twisted conformations. As a consequence of this symmetry, the effect of the torsion would cancel out during the linear fitting. Nonetheless, it has been possible to retain the effect of the twisting by dividing the total sampled trajectory into the three sampled regions. This way, three different motions, one for each of the PES wells, have been characterised.

6.3.2 Scan along the Energy Shift Dimension

As evidence that the dimensions found above correspond to that of the energy shift, a rigid scan was performed along the three vibrational coordinates shown in Figure 6.1. While it is true that the OLUFF PES does not exactly reproduce that of the QM S_1 state (see Figure 6.2(A-C)), it can be observed that it reproduces well the position and curvature of the minima, even though it was not parametrised considering this specific DOF and, thus, proving once again the robustness of the parametrisation carried out in Chapter IV. Moreover, it is shown that, for all the conformations, the S_0 and S_1 minima do not appear at the same point of the energy shift coordinate, leading to an increase or decrease in the emission energy for positive or negative values of the projection, respectively. Regardless of the conformation, the scan of the energy shift

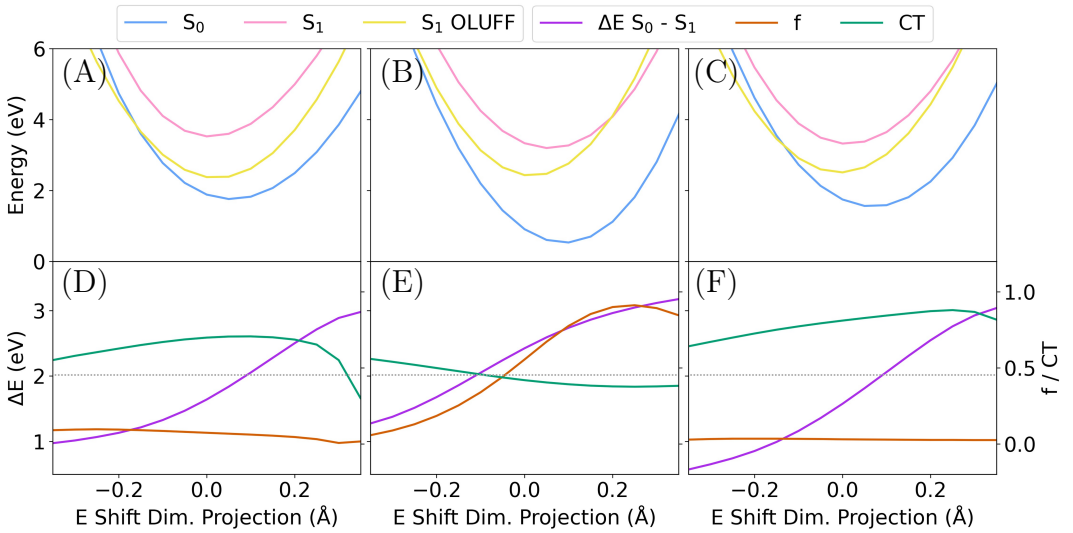


Figure 6.2: (A-C) PES and (D-E) emission energy (ΔE), oscillator strengths (f) and CT for geometries on (A,D) the 90º, (B,E) s-trans and (C,F) 270º conformations along the scanned energy shift dimension. The dashed horizontal line represents the the 600 nm threshold for the ΔE emission energy.

dimension shows that the emission energy can, in principle, go from 1 to 3 eV (Figure 6.2(D-F)). However, in practice, only the twisted conformations are able to reach emission energies bellow 2.07 eV or, equivalently, 600 nm as the s-trans conformation is not allowed to explore regions of the energy shift dimension that are located far from the reference geometry, as shown in Figure 6.3(A-C). Hence, the red emission observed inside some luciferases can not be produced from the planar s-trans conformation.

Furthermore, when looking at the oscillator strength, it can be observed that it presents strong correlation to the energy shift dimension for the s-trans conformation (Figure 6.2(E)), for which the oscillator strength can go from 0 at negative values of the projection to near 1 at 0.2 of the projection. Moreover, although the increase of the oscillator strength for this conformer is correlated with the decrease of the CT character, the CT character does not change a lot along the energy shift dimension of the s-trans conformer. In contrast to the planar situation, the twisted conformations present a CT character increase along the energy shift dimension while the oscillator strength for the two twisted conformations is very low and uncorrelated to the energy shift dimension, specially for the 270º conformation. Interestingly, for the 90º conformation, it appears that the oscillator strength slightly increases (up to 0.096) when the emission energy decreases. As a consequence of the very low

oscillator strength along the scan, emission within these twisted conformations should not be observed. However, when looking at the dynamic picture in Figure 6.3(D,F), the oscillator strength of some sampled structures is very high (over 0.3 for some sampled structures from both distorted conformations). Thus, it is possible to detect a red emission (over 600 nm) coming from the twisted conformations and this red emission should be facilitated by a combined effect of the vibrational motions along other PCs of the system in each of the two minima.

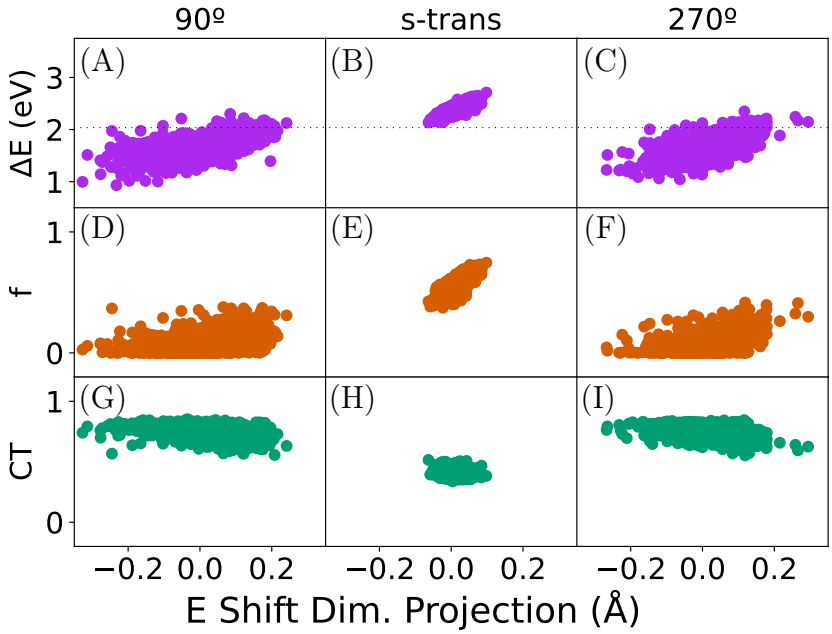


Figure 6.3: (A-C) Emission energy, (D-F) oscillator strengths and G-H) CT for the sampled geometries on (A,D,G) the 90º, (B,E,H) s-trans and (C,F,I) 270º conformations along the energy shift dimension. The dashed horizontal line represents the 600 nm threshold for the ΔE emission energy.

6.4 Conclusions

Firefly's BL system consists on an D-luciferin/luciferase complex that is able to emit light in the visible region after several biochemical reactions. Most WT luciferases present a green-yellow colour of emission (around 560 nm). However, a few of them and also some mutated luciferases are able to emit over 600 nm. This huge red-shift has been tried to be rationalised in many different ways, including conformational and polarity changes on the chromophore and the enzyme environment, respectively.[119] Nonetheless, there is still no consensus on the factors that lead to the colour modulation and role in this process.

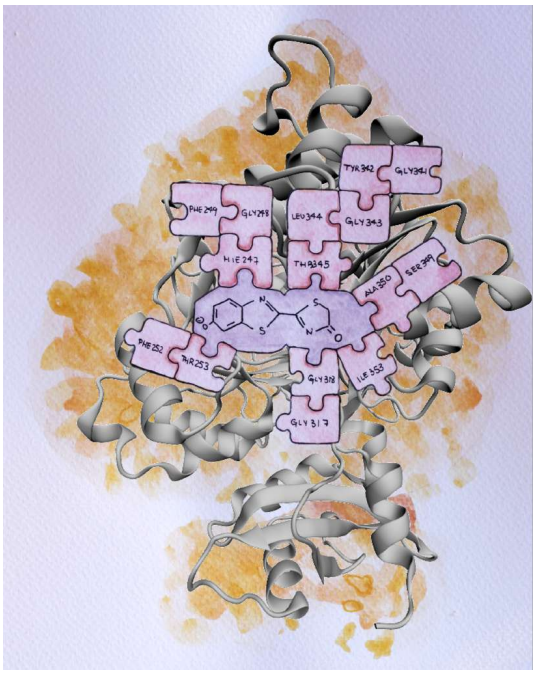
In this context, this chapter is aimed to understand the conformational changes that lead to the energy shifts of the emitted light. In order to do so, *in vacuum* classical MD sampling of the phenolate-keto OLU in the S_1 electronic state is combined with QM calculations to provide three energy shift dimensions, expressed as a linear combination of the PCs of the system. Each of these dimensions have been characterised for a different SCCS conformation namely, the planar s-trans and the 90 and 270^{o} twisted conformations.

It has been observed that, for each of the three studied conformations, the energy shift dimension consisted on in-plane tensions within each of the two moieties of the OLU chromophore. Moreover, and even though the scan along the energy shift dimension shows that, in principle, the emission energy can go from 1 to 3 eV for the tree sampled conformations, it appears that only the two twisted conformation are, in fact, able to populate regions of the S_1 PES that lead to a wavelength of emission larger than 600 nm. Furthermore, it has been proved that, in the twisted conformations, the dampening of the intensity caused by the CT state can be overcame by a combined effect of the different vibrational motions of the system, thus, increasing the brightness of the transition in a region which was previously believed to be dark.

Future work should focus on characterizing the nuclear coordinate that leads to the increase of the oscillator strength which, in combination with the energy shift dimension characterised here, could be employed to predict the emission energy and the brightness of the transition inside different luciferase enzymes without the need of performing additional QM calculations. Hence this could reduce the computational cost of the research process.

Chapter VII

Classical and QM/MM Energy Decomposition Analysis of Oxyluciferin/Luciferase in the Electronic Excited State



Unravelling the nature of the interaction energy on the complex formed by the OLU chromophore in the first electronically excited state and the luciferase enzyme can lead to a more profound understanding of fireflies' BL, which is a natural process with a large amount of applications in medicine. This chapter discusses the interaction energy of the phenolate-keto OLU chromophore and the luciferase protein and stresses the significant role of electrostatic terms on the total interaction energy. Nonetheless, non-electrostatic contributions are non-negligible due to the high presence of π-system-containing amino acids within the BL pocket. Furthermore, a methodology for performing energy decomposition analysis in the excited state has been outlined using a multiscale hybrid electrostatic-embedding QM/MM approach, stressing that the OLUFF is not suitable for this task from a FF-based approach. Moreover, the proximity of PHE249 and SER349 on the different terms of the interaction energy has also been analysed, unravelling that the most relevant terms of their contributions are the polarization and electrostatic ones, respectively. This strategy has been further extended to investigate the interactions of the other three potential emitters within the OLU/luciferase complex. Such analyses may help clarify the factors governing colour modulation in the BL process and provide valuable insights for tuning the OLU/luciferase complex to enhance its applicability.

---

* This chapter is partially based on the preprinted article *Characterization of the Phenoale-Keto Oxyluciferin/Lucifrase Interactions in the S_1 State by QM/MM Energy Decomposition Analysis* with contributions of Henar MateodelaFuente,[a] Nuria Anguita Ortiz,[a] Marcos Mandado,[b] and Juan J. Nogueira.[a,c]

[a] Chemistry Department, Universidad Autónoma de Madrid, 28049 Madrid, Spain.

[b] Department of Physical Chemistry, University of Vigo, Lagoas-Marcosende S/N, S36310 Vigo, Galicia, Spain.

[c] Institute for Advanced Research in Chemical Sciences (IAdChem), Universidad Autónoma de Madrid, 28049 Madrid, Spain.

Preprint: 10.5281/zenodo.17638190.

7.1 Introduction

In order to improve the performance of the OLU/luciferase system by, for example, being able to control colour of the emitted light, and to extend its already large number of applications, it is important to fully understand its nature. Chapter V discussed the structural behaviour of the OLU chromophores inside the luciferase pocket.[366] However, in order to achieve a better understanding of the emission process, it is quite relevant to characterize the nature of the ligand/protein interactions as detailed binding affinities have not been determined yet for all OLU analogues.[119] Thus, in this chapter the interaction energy between the two monomers of the system, the OLU chromophore and the luciferase protein, is calculated classically and quantum mechanically by the 1A-MM-GBSA and the QM/MM-EDA approaches, respectively. The EDA protocol provides much more accurate results than a simple energy analysis based on FF contributions, as discussed along this chapter. Furthermore, this chapter provides a suitable computational protocol for the calculation and analysis of the interaction energy for systems in electronically excited states, which may be as well applied to other chromophores interacting with complex environments. Hence, this chapter starts discussing only the results for the phenolate-keto OLU (Figure 7.1(A)) in order to stablish the computational protocol for the calculation and to analyse the interaction energy for system in the S_1 state and then, the proposed protocol is applied to the remaining three OLU emitters. It is shown that all the emitters interact with almost the same amino acids, evidencing the rigidity of the active pocket, regardlessness of the protonation state. Moreover, it has been demonstrated that the FF-based 1A-MM-GBSA approach presents a very different picture than the EDA. This difference arises as a result of an underestimation of the electrostatic interaction and overestimation the van der Waals term, the latter most likely due to an underestimation of the Pauli repulsion by the FF that is consistent with results already reported in literature.[258]

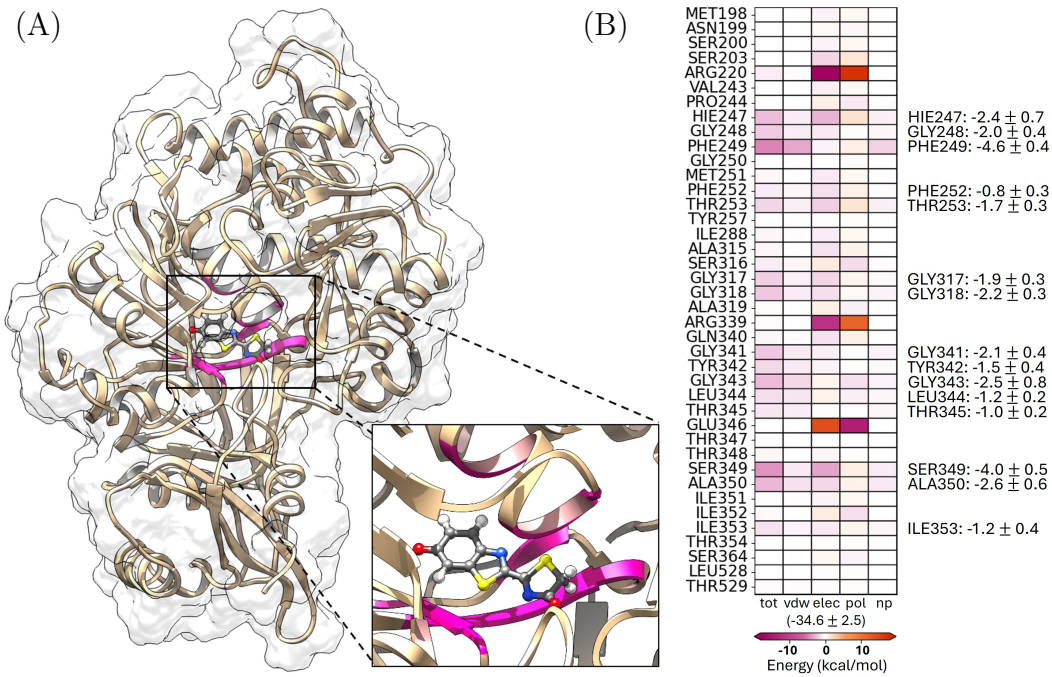


Figure 7.1: (A) Phenolate-keto OLU/luciferase complex with the 15 amino acids with the largest pairwise total binding free energy in pink. (B) Pairwise residue decomposition based on the 1A-MM-GBSA approach of total binding free energy (ΔG^{tot}) and the van der Waals (ΔG^{vdW}), electrostatic (ΔG^{elec}), polar solvation (ΔG^{pol}), non-polar (ΔG^{np}) solvation terms including the 40 residues that are located at shortest distances from OLU along the simulation. At the bottom of the the plot, between parenthesis, the total binding free energy of the chromophore appears and, at the right-hand side, the binding Gibbs free energy per residue of the 15 amino acids with the largest ΔG is represented.

7.2 Computational Details

This chapter employs the S_1 classical MD simulations calculated in Chapter V in order to obtain a thermally averaged ensemble of geometries for the analysis of the interaction energy with both the 1A-MM-GBSA and the QM/MM-EDA. As explained before, these simulations were computed for a 1.4 μs using potential gradients based on the OLUFF for the chromophore,[264] a reparametrized version of the GAFF[261] for the AMP, and the ff19SB46[262] FF for the Japanese firefly luciferase (PDB ID: 2D1R)[124] and the TIP3P[263] model for the Na^+ cations and water molecules. A full description of the

computational protocol of the dynamics can be found in Chapter V.

7.2.1 Interaction Energy

1A-MM-GBSA Approach

To study the nature of the total interaction energy in the S_1 state, the last 600 ns of the MD simulations were used to generate the ensemble of geometries of the system. In a first stage, these 600 ns were used to calculate, every 3 ns, the binding free energy by means of the 1A-MM-GBSA approach[287, 288] using the *MMPBSA.py* tool.[335] This binding free energy was decomposed into (i) interaction (electrostatic and van der Waals) and solvation (polar and non-polar) terms and (ii) amino acid contributions to characterize the nature of the interactions and to determine which amino acids present a higher contribution to the interaction energy, as shown in Figure 7.1(B). In order to reduce the computational effort, the decomposition was performed for the 40 residues that are located at shortest distances from OLU along the MD dynamics.

QM/MM-EDA Approach

Consecutively and in order to obtain a more accurate description of the interaction energy, the classical sampling was employed to perform the QM/MM-EDA along the simulation, using an scheme based on electron deformation densities[302, 367, 368] and treating the interaction between the QM and MM regions by an electrostatic embedding scheme. To do so, it is required that (i) the DFT functional and basis set is suitable for the calculation and (ii) the number of amino acids included in the QM is converged. For that, a representative structure of the 600 ns of the phenolate-keto trajectory in the S_1 was obtained by means of a hierarchical agglomerative (bottom-up) clustering analysis using the *cpptraj*[354] module of the AmberTools22.[358] Starting from an initial random frame. This algorithm analysed the trajectory every 20 snapshots, finishing the clustering when minimum the average distance between members of two clusters is greater than 50.0 Å. This convergence criteria ensures that only one structure was obtained.

In the case of the selection of the proper functional and basis set, monomer 1, treated at QM level, corresponds to the phenolate-keto OLU chromophore while monomer 2, treated at QM/MM level, represents the protein. Moreover, only PHE249, identified as the largest contributor to binding energy as shown in Figure 7.1(B), was included in the QM region of monomer 2, and the rest of the protein was described by a FF within a a MM electrostatic

embedding scheme. The QM region was described with different combinations of the CAM-B3LYP[231–233, 316] and the M062X[317] DFT functionals and 6-31G(2d,p),[342, 348–351] 6-31+G(2d,p),[342, 348–351, 369, 370] 6-311+G(2d,p)[342–344, 369, 370] and 6-31G(2df,p)[342, 343, 348, 349, 351, 371] basis sets with and without empirical dispersion treated by the D3BJ version of Grimme's dispersion with Becke-Johnson damping[372] (for CAM-B3LYP) or by the D3 version of Grimme's dispersion with the original D3 damping function[373] (for M062X) due to implementation constrains. As discussed in Annexe F, CAM-B3LYP is not a suitable DFT functional for the calculation of interaction energies and the most suitable functional/basis set combination for calculating binding energies is the M062X functional with the 6-31G(2d,p), as the inclusion of f polarization functions only increases the computational time without improving the results and the use of diffuse functions leads not only to a completely wrong description of dispersive terms and but also to convergence problems.

Once a suitable DFT functional was selected, the convergence with the number of amino acids in the QM region of the protein was studied for the phenolate-keto OLU by increasing the number of amino acids in the QM region of the luciferase from one to twenty and performing the QM/MM-EDA. The amino acids were added following the energetic ordering obtained from the 1A-MM-GBSA approach, reaching convergence for all the energy terms when fifteen amino acids were included in the QM region. Subsequently, the convergence with respect to the number of geometries was studied. For this, the fifteen amino acids were included in the QM region of the protein, and the number of equidistant frames to perform the QM/MM-EDA was increased up to two hundred and fifty, achieving convergence after one hundred and twenty five frames. Hence, for the remaining three OLU chromophores the QM/MM-EDA was directly calculated including only 15 amino acids on the QM region of monomer 2 and using a sampling of 123 frames for the phenol-keto OLU, 125 frames for the phenol-enolate OLU and 84 frames for the phenolate-enolate. In the case of the phenol-keto, this reduction of frames was due to CT values larger than 0.15 (see Figure F.4(B)) while for the phenolate-enolate, due to a convergence failures of the QM/MM-EDA. Despite this reduction on the number of sampled frames, the QM/MM-EDA of these three chromophores is converged, as can be observed in Figure F.5.

All the QM calculations were performed with the software Gaussian09[341] using the CP correction to overcome the BSSE.[307, 308] Moreover, in order to calculate the interaction energies in the S_1 state, the electron densities of the monomer 1 and the complex in the first electronic excited state were described with TD-DFT. On the flip side, the electron density of monomer 2

(the protein) was directly obtained at DFT level, using the same functional as for the chromophore, since the protein does not participate in the excited state of the system. This is clearly reflected on the analysis of the one-electron transition density matrix between the ground state and the S_1 states using the TheoDORE software.[249] The input preparation for the QM calculations was automatically performed by means of the *MoBioTools* toolkit,[329] which employs the link atom approach to describe the QM/MM interface,[374, 375] and the QM/MM-EDA was carried out using the QM/MM version of the *EDA-NCI* program.[376] Importantly, as the the *EDA-NCI* program does not support excited state calculations by itself, the SCF matrix and energy were manually deleted from the *fchk* file of the complex and the OLU monomer and the *CI Rho(1)* matrixes and energies were renamed to SCF so the program would read the correct descriptors.

7.3 Results and Discussion

As previously mentioned, understanding the OLU/luciferase system is crucial for modulating its photochemical properties. This chapter focusses on the computational characterization of OLU/luciferase interaction in the excited state from two different approaches: (i) the classical 1A-MM-GBSA approach, that calculates the interaction energy from the different terms in the FF, and (ii) the QM/MM-EDA approach, that uses a scheme based on electron deformation densities.

7.3.1 Phenolate-Keto Oxyluciferin

1A-MM-GBSA Approach

From a classical perspective, the interaction energy can be calculated from the different terms of the FF by means of the 1A-MM-GBSA,[287, 288] which further allows the pairwise per residue characterization of the different energetic terms, namely: van der Waals, electrostatic, polar solvation and non-polar solvation terms. Within this classical framework, most of the binding free energy comes from the van der Waals contribution, as shown in Table 7.1. This means that the stabilization of the OLU inside the active pocket would come from dispersion and inductive forces, while the electrostatic terms would not play a significant role in the stabilization. Moreover, as it can be observed in Figure 7.1(B), the 15 amino acids with the largest total binding free energy are, in decreasing order of interaction, PHE249, SER349, ALA350, GLY343, HIE247,

GLY318, GLY341, GLY248, GLY317, THR253, TYR342, LEU344, ILE353, THR345 and PHE252. Among them, only PHE249 and SER349 present a binding free energy below -4 kcal/mol, which are dominated by van der Waals and electrostatic terms, respectively. Furthermore, there are two amino acids that present a strong electrostatic stabilization (ALA220 and ARG339) and another with a significant polar solvation (GLU346). However, this stabilization is countervailed by either the polar solvation or electrostatic terms, leading to a very small contribution to the total binding free energy of the OLU/luciferase complex.

QM/MM-EDA Approach

In order to obtain a more accurate description of the intermolecular interactions responsible for the stability of the OLU/luciferase complex, the interaction energy and its different components were obtained at QM level. As explained in Section 2.4.2, the total interaction energy is divided into three different terms: (i) the electrostatic energy, accounting for the presence of effective charges and multipole moments; (ii) the Pauli repulsion energy, arising from the Pauli exclusion principle and further divided into attractive (intermolecular exchange) and repulsive (repulsion) terms; and (iii) the polarization energy depending on the polarization density, which can be approximated as a summation of induction and dispersion energies using a second-order perturbation expansion of the polarization density.

In the case of biological systems, and due to computational constrains, accurate and meaningful results of the substrate/protein interactions at QM level can only be achieved by means of a QM/MM framework. Herein, the QM region of the calculation includes a representative amount of the biomolecule, in this particular case the amino acids of the active site, while the rest of the biomolecule is replaced by a MM electrostatic embedding. Thus, the point charges of the MM region are added to the Hamiltonian during the QM calculation of the complex and fragments as well as to the subsequent calculations of the different interaction energy terms. Nonetheless, the selection of the amino acids that should be included in the QM region is not trivial. Furthermore, biological systems also require a large sampling from the MD simulation so that converged results are obtained in the energy analysis.

In this section, first the convergence of the EDA with respect to the number of amino acids in the QM region of the proteic monomer is analysed, followed by the convergence with the number of geometries selected from MD simulation. Moreover, for the selection of the amino acids of the QM, the proposed workflow relies on the 1A-MM-GBSA approach to determine which

amino acids, among the closest during the MD trajectory, present a larger contribution to the total interaction energy, as represented in Figure 7.1(B). However, since the objective is to investigate the interactions in the S_1 state, this approach is not ideal because, even though the FF is parametrised to perform the sampling in the excited state, the only reliable non-bonding term of the OLUFF is the electrostatic one, as Lennard–Jones parameters are taken from the General Amber FF,[261] which is parametrized for the ground state.

Convergence with the Number of Amino Acids -

Initially, the twenty amino acids with the largest binding Gibbs free energy according to the 1A-MM-GBSA approach were included one by one in decreasing order of contribution in the QM region for the TD-DFT calculations and then, the QM/MM-EDA is performed for the representative structure of the trajectory. As it can be observed in Figure 7.2(A), Pauli, induction and dispersion are fully converged when fifteen amino acids, those stressed in Figure 7.1(B), were included in the QM region while the electrostatic contribution is still slightly decreasing. This is not surprising due to the fact that electrostatic energy presents a strong long-range character and thus is more difficult to converge. However, since the difference compared to a QM region of twenty amino acids is only 3.4%, all the terms are considered to be converged when fifteen amino acids are included in the QM region.

It is important to mention that the interaction energy calculation, and therefore the EDA decomposition, in electronically excited complexes is only reliable if the electronic excitation is located on just one monomer (in this case, the phenolate-keto OLU chromophore). Otherwise, if the excitation entails substantial CT between monomers, an electronic excited state localized on a single monomer would fail to adequately represent the charge densities of the isolated monomers. In these cases, a weighted superposition of excitations on both monomers would provide a more appropriate description. To investigate this issue, the transition density matrix was analysed to calculate the number of transferred electrons between the two monomers and the mean position of the orbital before and after the electronic transition (POS). POS values of 1 or 2 indicate that electron is localized on monomer 1 or 2, respectively, both before and after the electronic transition, whereas a value of 1.5 would indicates delocalization between the two monomers. As it can be observed in Figure 7.2(B), the protein environment does not participate in the excited state because there is not any CT character and the POS is localized in the OLU chromophore. Thus, the interaction energy calculation and the subsequent EDA can be easily performed using the OLU excited state and the protein ground state as references for the isolated densities of the monomers.

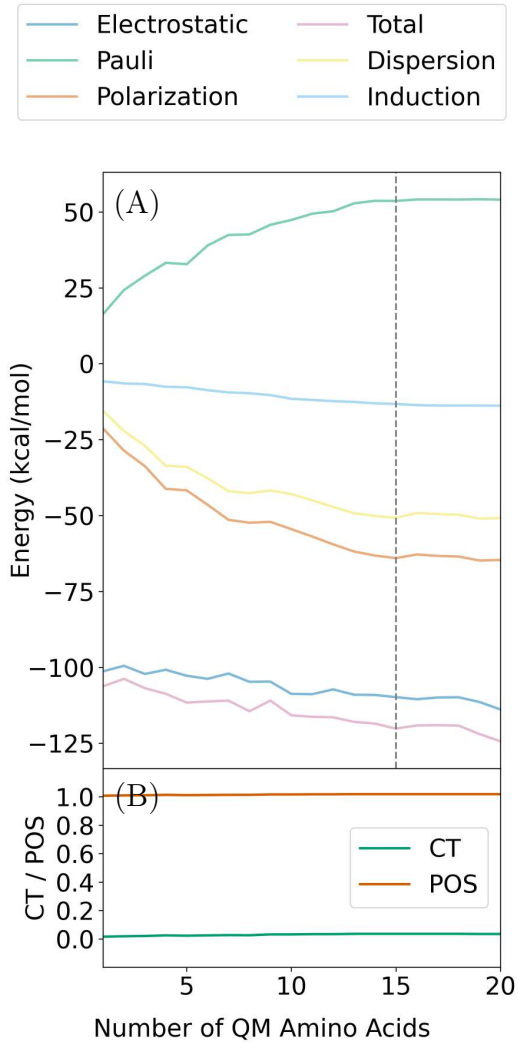


Figure 7.2: (A) Convergence of the interaction energy and the EDA components as the number of amino acids in the QM region increases. (B) Number of transferred electrons between the two monomers (green) and averaged position of the orbital before and after the electronic transition (orange).

Convergence with the Number of Frames - As mentioned before, not only the number of amino acids in the QM region has to be converged, it is also important to have a good conformational sampling of the system, for instance, a sufficiently large number of snapshots for the analysis of the properties. Thus, EDA calculations were performed on 1, 2, 5, 10, 25, 50, 125 and 250 equidistant geometries from the MD trajectory. This analysis was per-

formed on electronically excited complexes localized on a single monomer (see Figure F.4). As it can be observed in Figure 7.3, all the interaction energy the components are converged when 125 snapshots were considered in the analysis. Therefore, the 250 snapshots for which the energies were computed are enough to obtain reliable results.

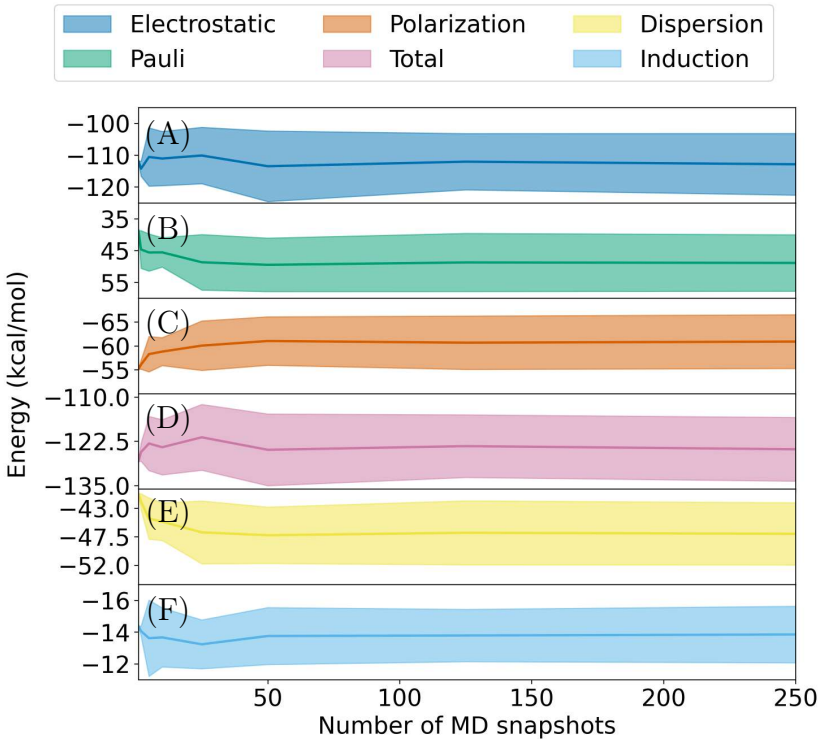


Figure 7.3: Convergence of the interaction energy and each of the EDA components with respect to the number of sampled geometries.

Once convergence of the energetic terms is guaranteed, let's focus on the analysis and discussion of the results. First, the probability distributions of the total interaction energies and of each of the energy components are shown in Figure 7.4. It can be observed that the different energy components are not distributed similarly as their distributions present different broadness. Furthermore, as previously demonstrated,[377] the broadness of the distribution is associated with the magnitude of the average value (see Table 7.1). These results show that the phenolate-keto OLU/luciferase interaction in the S_1 state is very favourable and that the electrostatic character is predominant over Pauli and polarization. These results are radically opposite to the results obtained with the 1A-MM-GBSA approach. Thus, the classical model strongly under-

estimates the electrostatic interaction (17.0% vs 90.4%) and overestimates the van der Waals term (83.0% vs 9.6%), corresponding to the sum of the Pauli and polarization terms in the EDA framework. This overestimation of the van der Waals contribution is likely a consequence of the underestimation of the Pauli repulsion by the FF, as previously observed.[258] A proper description of the Pauli repulsion would partially cancel out the attractive interactions and, as a result, the van der Waals energy would be less important at classical level. Therefore, while the 1A-MM-GBSA approach erroneously states that the interaction is mostly due to van der Waals forces, the QM/MM-EDA shows that the interaction is, in fact, of electrostatic character.

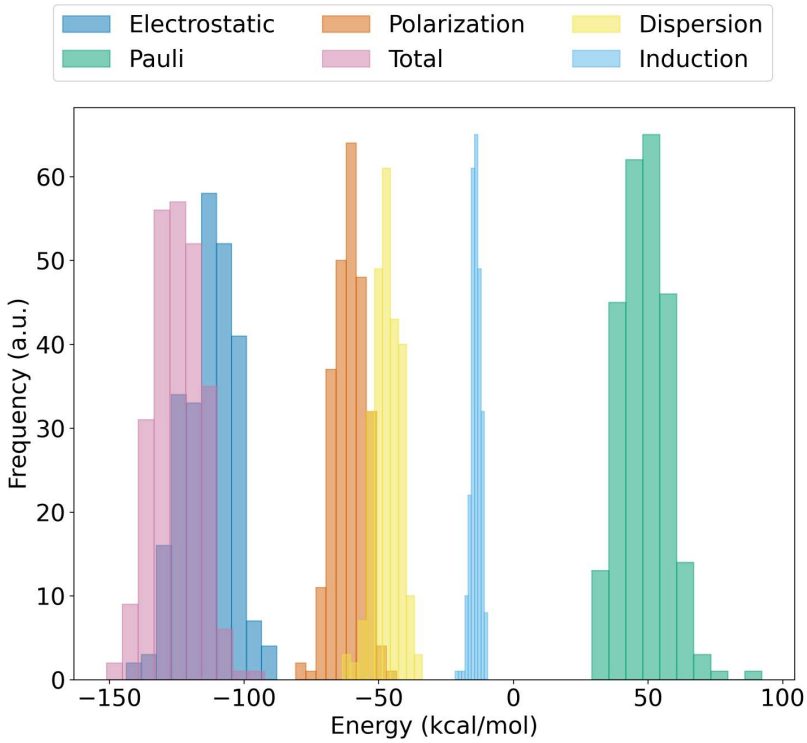


Figure 7.4: Energy distributions of the contributions to the total interaction energy calculated from the QM/MM-EDA on 250 geometries.

Table 7.1: Averages and standard deviations of the 1A-MM-GBSA and EDA components distributions calculated from the FF and the QM/MM-EDA on 250 geometries. Percentages are relative to the sum of electrostatic and van der Waals terms (1A-MM-GBSA) and electrostatic and pauli+polarization (EDA). Energies are in kcal/mol.

1A-MM-GBSA			EDA		
Contribution	Energy	%	Contribution	Energy	%
Electrostatic	-8.5 ± 8.3	17.0	Electrostatic	-112.9 ± 9.7	90.4
			Pauli	48.9 ± 8.8	-
van der Waals	-41.4 ± 2.0	83.0	Polarization	-60.9 ± 5.6	-
			van der Waals	-12.0 ± 4.9	9.6
Polar	19.2 ± 8.0	-	Dispersion	-47.1 ± 4.9	-
Non-polar	-3.86 ± 0.11	-	Induction	-13.8 ± 1.8	-
Total	-34.6 ± 3.5	-	Total	-124.8 ± 9.0	-

Finally, the effect of the proximity of PHE249 and SER349 amino acids has been analysed (Table 7.2). To do so, the 250 geometries selected for the QM/MM-EDA were divided according to the distance between the centre of mass of the OLU chromophore and that of each amino acid in two sets of 100 frames, each with the shortest and the largest distances. The selection of these two amino acids was not arbitrary, as they are the two amino acids that interact more strongly with the OLU according to the 1A-MM-GBSA approach. Moreover, in Chaper V, it is already demonstrated the significant presence of PHE249 during the MD simulations, which may affect the attractive polarization component of the interaction energy. In fact, in the case of the PHE249, the total interaction energy of OLU with the environment is very similar regardless of the proximity of this amino acid (-124.8 ± 9.6 kcal/mol for a distance of 3.92 ± 0.09 Å and -124.7 ± 8.7 kcal/mol for a distance of 4.3 ± 0.1 Å) as a consequence of the larger decrease in the Pauli repulsion (7.1 kcal/mol), which counteracts the decrease of the attractive electrostatic and polarization terms (7.2 kcal/mol). Furthermore, it can be observed that the decrease in the electrostatic energy is less pronounced (2.5 kcal/mol) than the decrease in the polarization term (4.7 kcal/mol) when this amino acid is farther. Hence, it can be concluded that OLU-PHE249 interaction presents a large polarization contribution, which is consistent with the aromatic nature of the system, where the stacking interactions must be relevant.

On the flip side, the presence of SER349 in the vicinity of the phenolate-keto OLU (4.98 ± 0.08 Å) does contribute to a the complex stabilization as a decrease of 3.0 kcal/mol, in absolute value, in the total interaction energy is

observed when this amino acid is farther (5.3 ± 0.1 Å) from the chromophore. This decrease in the interaction energy is mostly due to a decrease in the electrostatic contribution (5.7 kcal/mol in absolute value), which is larger than that of the polarization term (2.4 kcal/mol in absolute value). In this case, the decrease of the Pauli repulsion (5.2 kcal/mol) is not able to fully counteract the most favorable attractive interactions. In contrast to PHE249, when SER349 moves farther from the OLU, the electrostatic contribution decreases more than the polarization term. Thus, it can be concluded that the OLU-SER349 interactions present an electrostatic nature, which is consistent to the presence of the alcohol group in its lateral chain.

Table 7.2: Averages and standard deviations of the EDA components distributions calculated from 100 frames out of the 250 EDA calculations in which the amino acids PHE249 and SER349 were closest or farthest to OLU. Energies are in kcal/mol.

	PHE249		SER349	
Contribution	Closest	Farthest	Closest	Farthest
Electrostatic	-114.2 ± 10.2	-111.7± 9.5	-116.0 ± 9.9	-110.3 ± 9.1
Pauli	52.7 ± 8.3	45.6 ± 8.7	51.7 ± 8.9	46.5 ± 8.3
Polarization	-63.4 ± 5.2	-58.7 ± 5.3	-62.2 ± 5.7	-59.8 ± 5.6
Dispersion	-49.2 ± 4.6	-45.2 ± 4.4	48.0 ± 5.1	-46.0 ± 4.7
Induction	-14.1 ± 1.7	-13.5 ± 1.9	-14.2 ± 1.9	-13.7 ± 1.6
Total	-124.8 ± 9.6	-124.7 ± 8.7	-126.4 ± 8.7	-123.4 ± 8.8

7.3.2 Phenol-Keto, Phenol-Enolate and Phenolate-Enolate Oxyluciferins

As for the phenolate-keto OLU, this section is dedicated to the characterization of the OLU/luciferase interactions of the phenol-keto, phenol-enolate and phenolate-enolate tautomers following the procedure mentioned before.

1A-MM-GBSA Approach

Similarly to the phenolate-keto OLU, the main contribution to the binding free energy of both the neutral an dianionic tautomers comes from the van der Waals contribution, while for the phenol-enolate OLU these two terms present a similar contribution due to a larger electrostatic interaction, as shown in Table 7.3. Moreover, as reflected by the polar solvation energy, the phenolate-enolate OLU presents a very low polar solvation penalty suggesting that, upon

binding to the luciferase, many interactions with the solvent are retained. This is consistent with the RDF presented in Figure 5.5, which shows that the solvation sphere is only retained for this chromophore. On the flip side, the polar solvation energy for the phenol-keto and phenol-enolate tautomers is very high, as they completely loose their solvation upon interaction with the protein. Furthermore, the non-polar contribution to the solvation energy remains similar across all the chromophores and appears to play a minor role in the binding energy.

Table 7.3: Averages and standard deviations of the MM-GBSA distributions calculated from the FF. Percentages are relative to the sum of electrostatic and van der Waals terms. Energies are in kcal/mol.

	Phenol-keto		Phenol-enolate		Phenolate-enolate	
Contribution	Energy	%	Energy	%	Energy	%
Electrostatic	-9.7 ± 3.9	18.5	-44.4 ± 9.4	51.5	3.7 ± 25	12.0
van der Waals	-42.8 ± 2.2	81.5	-41.8 ± 2.1	48.5	-27.1 ± 2.5	88.0
Polar	24.9 ± 3.1	-	51.0 ± 8.5	-	4.1 ± 23	-
Non-polar	-4.02 ± 0.11	-	-4.00 ± 0.14	-	-3.52 ± 0.21	-
Total	-31.6 ± 3.3	-	-39.2 ± 3.0	-	-22.8 ± 4.0	-

These results can be further analysed upon the pairwise residue decomposition of the binding free energy, shown in Figure 7.5. As for the phenolate-keto OLU, these three chromophores present a very strong interaction with PHE249 due, mostly, to the van der Waals contribution, although for the phenolate-enolate OLU is not the strongest one. Moreover, they also interact strongly with SER349. Nonetheless, the nature of this interaction changes with the total charge of chromophore. While the interaction with the phenolate-enolate is exclusively electrostatic, the interaction with the the phenol-keto presents a larger van der Waals contribution. Moreover, the 15 amino acids with a larger contribution for the phenol-keto (PHE249, SER349, SER316, ALA350, THR345, GLY341, LEU344, HIE247, GLY343, TYR342, ILE353, THR253, GLY317, GLY318, GLY248 in decreasing order on interaction), the phenol-enolate (PHE249, SER349, ALA350, GLY343, LEU344, GLY248, GLY317, GLY341, SER316, GLY318, TYR342, HIE247, ILE353, THR253, THR348 in decreasing order on interaction) and phenolate-enolate (ARG220, SER349, ALA350, ILE353, THR253, PHE252, ARG339, PHE249, GLY341, SER316, GLY248, ALA315, GLN340, GLY317, THR354 in decreasing order on interaction) are almost the same, which evidences the rigidity of the active pocket independently of the protonation state of the chromophore.

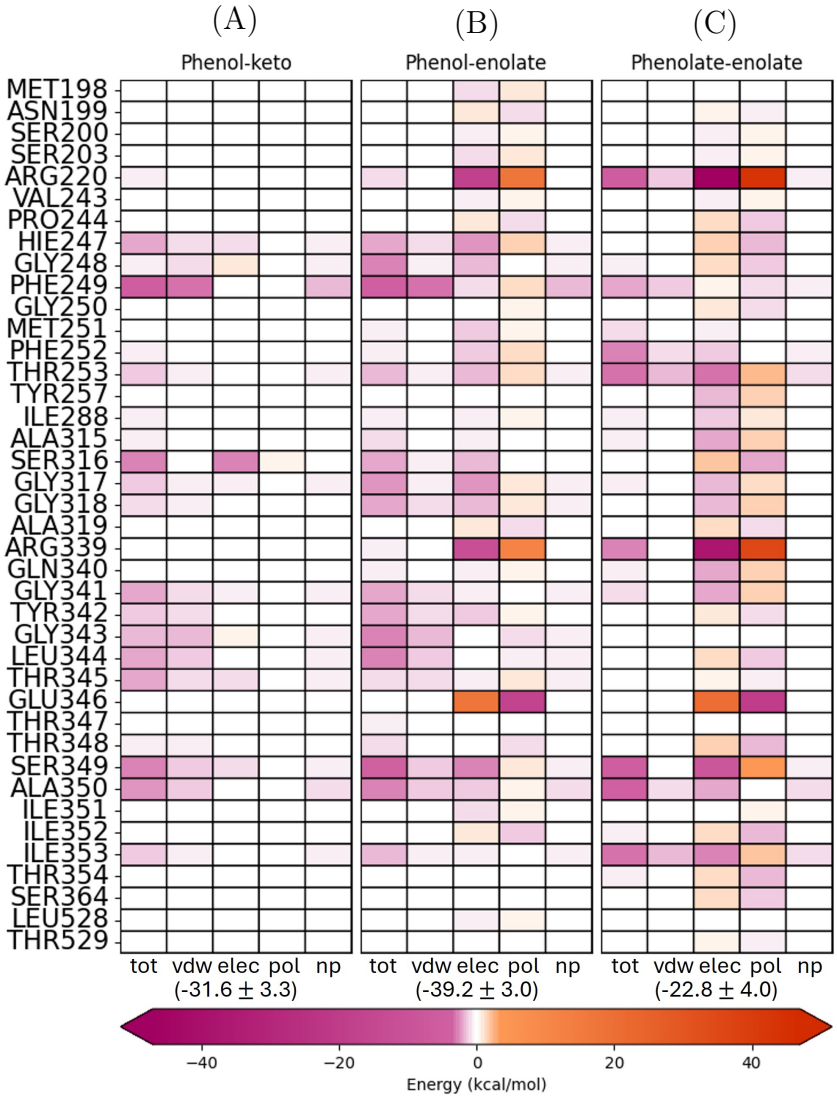


Figure 7.5: Pairwise residue decomposition based on the MM-GBSA approach of total binding free energy (ΔG^{tot}) and the van der Waals (ΔG^{vdW}), electrostatic (ΔG^{elec}), polar solvation (ΔG^{pol}), non-polar (ΔG^{np}) solvation terms including the 40 residues that are located at shortest distances from the (A) phenol-keto, the (B) phenol-enolate and the (C) phenolate-enolate OLUs along the simulation. At the bottom of the the plot, between parenthesis, the total binding free energy of the chromophore appears.

QM/MM-EDA Approach

In order to obtain more accurate results for the interaction energy, the QM/MM-EDA was also performed for the phenol-keto, phenol-enolate and phenolate-

enolate OLUs. As it can be observed in Figure 7.6, again all the probability distributions of the total interaction energies and each of the energy components obtained from the QM/MM-EDA are not distributed similarly and their respective broadness are associated with the magnitude of the average value (Table 7.4).[377]

Table 7.4: Averages and standard deviations of the EDA distributions calculated from the QM/MM calculations. Percentages are relative to the sum of electrostatic and pauli+polarization terms. Energies are in kcal/mol.

	Phenol-keto		Phenol-enolate		Phenolate-enolate	
Contribution	Energy	%	Energy	%	Energy	%
Electrostatic	-40.0 ± 5.4	87.9	-114.6 ± 8.6	86.8	-456 ± 33	79.0
Pauli	58.8 ± 9.6	-	59 ± 11	-	430 ± 90	-
Polarization	-63.3 ± 6.6	-	-76.8 ± 7.6	-	-309 ± 64	
van der Waals	-4.5 ± 4.0	12.1	-17.4 ± 6.0	13.2	121 ± 31	21.0
Dispersion	-50.1 ± 5.2	-	-58.7 ± 5.9	-	-183 ± 41	-
Induction	-13.3 ± 2.0	-	-18.2 ± 2.8	-	-126 ± 25	-
Total	-45.4 ± 4.5	-	-131.9 ± 8.9	-	-335 ± 18	-

In the case of the phenol-keto OLU, it can be observed that, although the polarization is larger, both electrostatic (-40.0 ± 5.4 kcal/mol) and polarization (-63.3 ± 6.6 kcal/mol) terms present a large contribution to the total interaction energy (-45.4 ± 4.5 kcal/mol). Similarly to the phenolate-keto tautomer presented in the previous section, the phenol-enolate chromophore presents a predominant electrostatic character with a large contribution from the polarization and Pauli terms. In the case of the dianion, all the EDA energy terms are larger because the interactions are stronger since the molecule is doubly charged. Evidently, the increase of the total charge of one of the interacting monomers leads to a larger electrostatic term (-456 ± 33 kcal/mol). This effect is specially stressed by the fact that this system presents two positively charged ARG amino acids in the vicinity of the phenolate-enolate OLU, which does not happen in any of the other three systems. Moreover, Pauli repulsion energy (430 ± 90 kcal/mol) also increases as a consequence of the extra negative charge which, at the same time, can be more easily polarized by the environment, thus leading to a larger polarization contribution (-309 ± 64 kcal/mol). Furthermore, as happened in the case of the phenolate-keto OLU, these results contrast to the results obtained with the 1A-MM-GBSA approach as again the classical model underestimates the electrostatic interaction (18.5% vs 87.9%, 51.5% vs 86.8% and 12.0% vs 79.0%) and overestimates

the van der Waals term (81.5% vs 12.1%, 48.5% vs 13.2% and 88.0% vs 21.0%) as a consequence of the underestimation of the Pauli repulsion by the FF, as previously discussed.

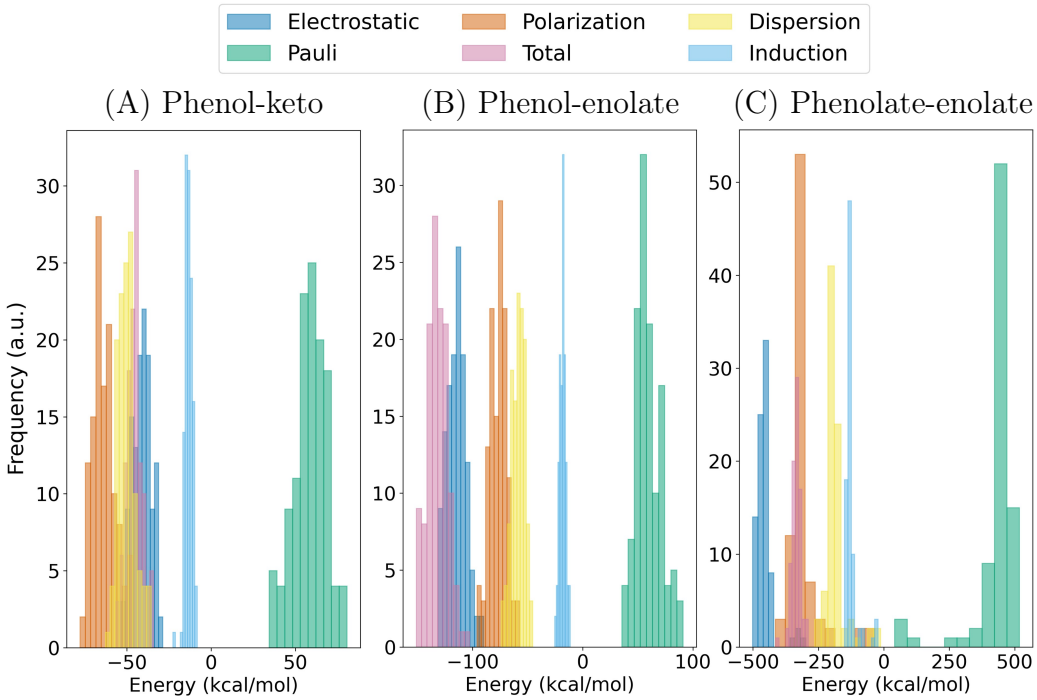


Figure 7.6: Energy distributions of the contributions to the total interaction energy calculated from the QM/MM-EDA sampled geometries from the trajectories of the (A) phenol-keto, (B) phenol-enolate and (C) phenolate-enolate OLUs.

7.4 Conclusions

This chapter employs both the 1A-MM-GBSA and the QM/MM-EDA to characterise, for all the possible OLU emitters, the OLU/luciferase interaction in the S_1 state. It has been evidenced: (i) the clear need of a QM treatment for the calculation of interaction energies as the OLUFF fails to reproduce the importance of the different energy terms, (ii) the importance of considering a sufficiently large QM region (15 amino acids of the protein plus the OLU chromophore) and (iii) the relevance of a good conformational sampling since wide energy distributions are obtained.

It was observed by the QM calculations and the QM/MM-EDA that the interaction energy in the electronically excited state is governed mostly by the electrostatic term for all the OLU emitters, specially for the phenolate-enolate chromophore due its double negative charge, that induces the presence of two positively charged ARG amino acids in the active pocket. Moreover, non-electrostatic contributions are not negligible due to the presence of amino acids with π systems inside the active site of the protein, being PHE249 the most important one. In addition, the relevance of the SER349 residue, which interacts mainly electrostatically with the chromophore, was highlighted.

The methodology presented here can be applied to explore not only the interaction between the OLU and the WT luciferase enzyme, but also additional complexes with mutated proteins. This would help to elucidate the factors that affect the colour modulation of the bioluminescent process. Moreover, different complexes with locally excited chromophores could also be analysed.

Chapter VIII

Structure of the Phenolate-Keto Oxyluciferin Emitting Species inside the SER286ASN Mutated Luciferase

Studying the changes on biological environment that surrounds the fireflies' bioluminescent chromophore upon single mutations on the luciferase enzyme is crucial to unravel the key factors that lead to the fine tuning of the emitted radiation. This chapter employs the OLUFF to sample the conformational space of the phenolate-keto OLU emitter in the first electronically excited state of the red-shifted SER286ASN mutant luciferase to describe the solvation and the interactions with the protein, in comparison to the WT luciferase. It is shown that the chromophore inside the mutant is able to populate three different regions of the SCCS dihedral angle, a s-trans planar structure and two twisted conformations. On the one hand, the s-trans region presents similar solvation than inside the WT protein resulting from the compensation of the reduction of the direct hydrogen bonds with solvent molecules and the increase of the amount of water bridges between the chromophore and both the enzyme and the AMP cofactor. On the other hand, the twisted conformations present a different solvation pattern resulting on the significant reduction of direct hydrogen bonding with the enzyme and the increase of the water bridging with the both the enzyme and AMP. As a result, the active pocket of the mutant presents an environment that is slightly more polar than in the WT luciferase, specially for the two twisted conformations.

8.1 Introduction

Up to this point, this thesis has been focused exclusively on the description of the complex formed by the OLU chromophore and the WT luciferase. Nonetheless, it is also relevant to characterise the behaviour of a red emitting complex, in comparison to the WT. Studying these two situations might lead to a deeper understanding of the structural changes that accompany the red shift of the emitted light. It has been reported in the literature that, upon the SER286ASN mutation, the emission of the OLU/luciferase complex is red shifted from 560 to 605 nm.[124] In that work, the authors reported the relevant role of the ILE288 in the colour modulation as they observed that this mutation might prevent conformational change of the luciferase during the adenylation reaction. Specifically, they observed that the side chain of the ILE288 does not move towards the chromophore in the SER286ASN mutant, suggesting that the hydrophobicity of the microenvironment and the rigidity of the chromophore is key to understand the BL colour. However, when they compared the SER286ASN structure with the adenylated OLU intermediate with the WT structure in complex with AMP/OLU, which would correspond to the structure after the emission, they observed that the direction of the lateral chain of ILE288 is similar in both systems. The ILE288 movement towards the chromophore was later confirmed by means of MD simulations on the adenylate intermediate for the WT and SER286ASN, SER286GLY, and SER286ILE mutants, which further suggested that the colour modulation depends on both the rigidity of the active site and the hydrogen bonding network around the O atom of the btz ring.[378] Moreover, despite the very accurate emission energies achieved by *ab initio* QM/MM calculations on the WT system[363] and following exactly the same procedure, Tagami et al. obtained emission energies that were blue-shifted around 30 nm for the ILE288ALA red-emitting mutant and were not able to reproduce the two peaks structure (at 560 and 613 nm) of the ILE288VAL. These independent works hence suggest that there should exist other structural differences that explain the red-shifted emission of the SER286ASN mutant.

In order to assess this issue, in this chapter the behaviour of the phenol-keto tautomer is characterized inside the SER286ASN mutant of the luciferase enzyme and compares it with that of the WT complex, that has been already disused in Chapter V. Moreover, and analogously to Chapters III and VI, the phenolate-keto OLU has been employed along this chapter as it is not only considered to be one of the emitting species but specifically, the only red-emitting one.[321–323] It is shown in this chapter that the OLU inside the active pocket of the SER286ASN mutant is able to populate the twisted conformations with

a SCCS dihedral angle of 90 and 270º while in the WT luciferase the structure is confined in the s-trans region. The predicted population of these regions of the PES was very low and, even though this could be a consequence of classical MD not presenting the same sampling capability that a QM-based distribution such as the Wigner sampling,[379] it could also be due to the bias introduced by assuming that the excited state OLU is obtained in the s-trans conformation, as suggested by a static *in vacuum* mechanistic study at the second-order multi-configurational perturbation (CASPT2).[65] Nonetheless, this study did not considered the environment of the chromophore. As commented before, the ILE288 does not move towards the chromophore after the adenylation reaction in the SER286ASN mutant. Hence, the active pocket of the enzyme is less hindered during the oxidation step than in the WT luciferases which could result on a (partially) twisted conformation of the resulting excited state OLU upon further interactions with the environment.

These twisted conformations of the SER286ASN mutant not only present a different solvation than the s-trans conformation, but also a completely different hydrogen bonding patter with both the protein and the AMP cofactor, as will be discussed in Section 8.3.

8.2 Computational Details

This chapter employs three replicas of MD simulations of the phenolate-keto OLU/luciferase complex in the S_1 state with the WT and the SER286ASN

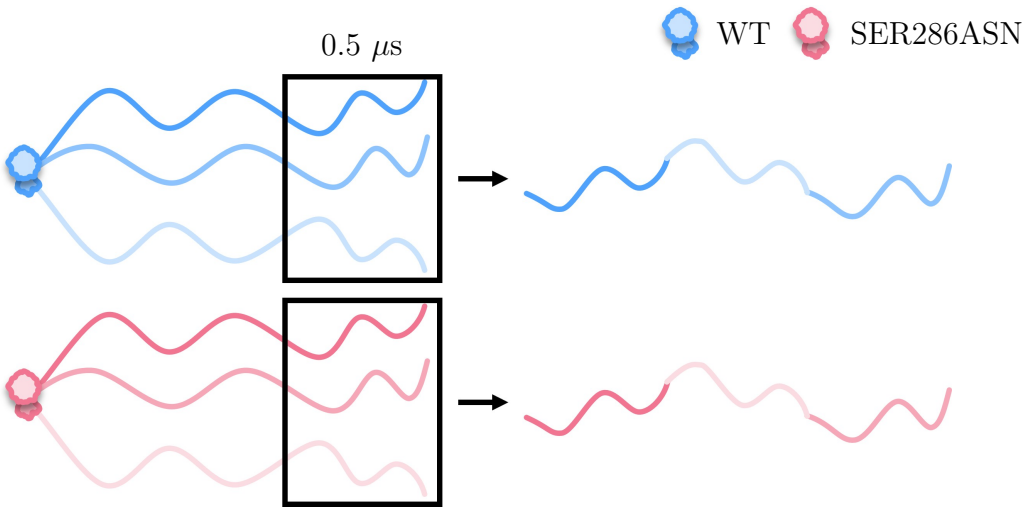


Figure 8.1: Scheme of the computational protocol followed along Chapter VIII.

mutant. Once the protein is stabilised, the last 0.5 μs of the simulations for each system are packed together and employed in the different structural analysis, as graphically represented in Figure 8.1.

8.2.1 System Setup

Similarly to Chapter V, both OLU/luciferase complexes were built with PyMol[324] from the X-ray diffraction structures file of the Japanese firefly, *Luciola cruciata*, luciferase published in the PDB (PDB ID: 2D1R for the WT protein and 2D1T for the SER286ASN mutant).[124] The chromophore was manually docked inside the protein pocket by replacing the structure of the crystallized OLU and the AMP residue was modified as to achieve a net charge of -1, that would correspond to its biological state,[356, 357] by adding the missing H atoms in the crystallographic structure. Consecutively, the system was solvated and protonated using the *tleap* module of AmberTools22[358] with water molecules within a truncated octahedral box, ensuring a solvent shell of at least 12 Å from any solute molecule. Moreover, the *tleap* module was also used to neutralize both systems by adding one Na^+ ion.

Furthermore, these two systems employ the same set of FF parameters as the phenolate-keto simulation in the S_1 state in Chapter V. The water molecules and Na^+ cations were described by means of the TIP3P FF;[263] the pehnolate-keto OLU in the S_1 electronic state, by the OLUFF;[264] the enzyme, by the ff19SB FF[262] and the AMP, by the GAFF[261] for the dihedral and improper torsions and Lennard-Jones parameters and by parametrised bonded parameters and charges.

8.2.2 Classical Molecular Dynamics

Then, the two systems were minimised for 5000 steps by means of the steepest descent algorithm,[244] and for additional 5000 steps using the conjugate gradient algorithm.[245] From the minimised structure of each of the systems, the three aforementioned replicas were run. Each replica consisted on a progressive heating from 0 to 300 K in the NVT ensemble for 500 ps using a 2 fs time step followed by additional 500 ps in the same ensemble and a production with a 2 fs time step within the NPT ensemble at 303.15 K and 1 atm. On the one hand, simulations on the WT system were run for 1.4 μs (replica 1), 1.7 μs (replica 2) and 1.5 μs (replica 3) while, on the other, two of the simulations on the SER286ASN mutant were run for 1.5 μs and the other for 1 μs. This difference in production times is necessary to ensure an stable conformation on the protein, as shown in Figure G.1.

Moreover, for all the simulations temperature was controlled by means of the Langevin thermostat[267, 268] with a collision frequency of 2.0 ps^{-1} and 1.0 ps^{-1} for the heating and production, respectively. Furthermore, the pressure in the NPT ensemble was controlled by means of the Berendsen barostat,[271] the particle-mesh Ewald[272] method was used to compute the electrostatic interactions with a grid spacing of 1.0 Å, the non-bonded interactions were computed with a cutoff of 13 Å and a switching distance of 11 Å, and the bonds involving hydrogen atoms were constrained using the SHAKE algorithm.[274] Finally, the different structural analysis performed along this chapter have been carried out using the *cpptraj*[354] module of AmberTools22 by taking into account the last 500 ns of the MD trajectories of each of the two systems to consider only the stable conformation of the protein. Specifically, the MW RMSD of the protein was analysed by means of the *rmsd* command after the protein were aligned along the whole trajectory; the torsion around the SCCS dihedral, by means of the *dihedral* and *hist* commands; the RDF, by means of the *radial* command; the hydrogen bonding and water bridging, by means of the *hbond* command using the default distance and angle cutoffs (3.0 Å and 135º, respectively); and the OLU-AMP distance, by means of the *distance* command taking into account the centre of mass of each of the molecules.

In addition to the WT and SER286ASN trajectories, the latter simulation was further divided into the three sampled regions of the SCCS dihedral of the chromophore by means of a clustering analysis with the hierarchical agglomerative (bottom-up) algorithm using the SCCS dihedral value as distance between frames and sieving every 10 frames. After clustering, the rest of the frames were be added to the corresponding cluster. As only three different regions of the SCCS dihedral angle were sampled, the algorithm was considered to be converged when three clusters remained. These three fragments of the SER286ASN trajectory (with 28246, 936 and 819 frames, respectively for the s-trans, 90º and 270º conformations) were also analysed to compare the geometrical features in each of these regions.

Moreover, it has been observed that the classical MD simulation underestimates the sampling in comparison to a Wigner distribution because it does not take into account the zero point energy of the system.[379] Hence, in order to achieve a sampling equivalent to that of a Wigner distribution for the SER286ASN mutant, the system was progressively heated form the last step of the first replica from 303.15 K to 328.15 K using the heating ramp represented in Figure G.2(A). Initially, the temperature was progressively increased 5 K for 25 ns and then the system was allowed to stabilise for another 50 ns at 308.15 K. This procedure was repeated five times until a temperature of 328.15 K is reached to avoid the denaturalization of the protein (see Figure G.2(B)).

8.3 Results and Discussion

As already mentioned, this chapter aims to explain the structural behaviour of the phenolate-keto OLU in the S_1 state in two different media, the WT *Luciola cruciata* luciferase and its SER286ASN mutant. Following the same structure than in Chapter V, first the geometrical configuration of the chromophore is discussed, followed by its interactions with the different components in the media: (i) the water molecules inside the active site, (ii) the amino acids of the luciferase and (iii) the AMP co-factor. Moreover, as the chromophore inside the the mutant is able to reach different regions of the conformational space, the SER286ASN trajectory has been divided into three different regions for a deeper discussion of the structural changes with respect to the simulation in the WT enzyme.

8.3.1 Structure of the Oxyluciferin Chromophores

As it was already observed in Figure 5.4, inside the WT protein the OLU chromophore is trapped in the s-trans conformation due to the steric hindrance and enzyme-substrate interactions (see Figure 8.2). However, and although the larger population is present for the s-trans conformation, the SER286ASN trajectory presents a non-negligible population of the two twisted conforma-

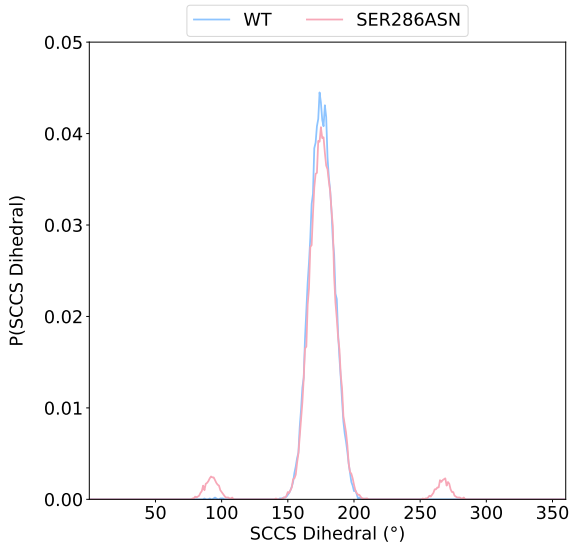


Figure 8.2: Probability distributions of the SCCS dihedral angles of the phenolate-keto OLU in the WT (blue) and SER286ASN (pink) systems.

tions (at 90 and 270º), that is sampled in all the replicas (see Figure G.3) in contrast to the WT system. Moreover, when the sampling is performed at higher temperatures, which would lead to a sampling more similar to that of a Wigner distribution,[379] the probability distribution of the twisted structure also increases (see Figure G.2).

As mentioned in Chapter III, the twisted conformations seem to be those yielding the red-shifted emission. Hence, the SER286ASN trajectory has been divided into three different fragments namely: (i) s-trans, (ii) 90º and (iii) 270º twisted structures, in order to analyse the structural differences among them.

8.3.2 The Role of Water Molecules

When comparing the solvation of the phenolate-keto chromophore within the active site of the WT and SER286ASN luciferases, it can be observed that the RDFs around the four electronegative atoms of the OLU in Figure 8.3(A-D) is very similar for both enzymatic media, although the RDF around the O atom of the btz moiety is larger for the WT system. Nonetheless, this result is biased by the larger amount of the s-trans conformation sampled during the MD simulation and therefore, it is crucial to analyse the mutant results for each of the SCCS conformers independently.

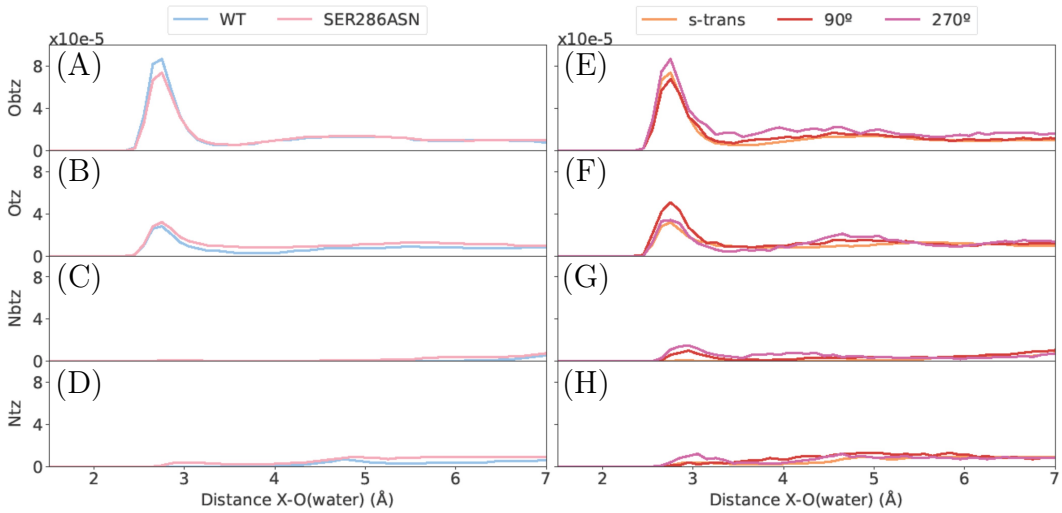


Figure 8.3: RDF of the O atoms from water surrounding each of the X four electronegative sites of phenolate-keto OLU (A-D) in the WT and SER286ASN systems and (E-H) in the different SCCS conformations in the SER286ASN system.

After analysing the SER286ASN system for each of the different SCCS conformations, some slight differences can be observed as shown in Figure 8.3(E-H). On the one hand, the 270º and 90º conformations present, respectively, a larger RDF around the O atom of the btz and tz moieties. The former is due to the lower amount of hydrogen bonding with GLY342, the only amino acid that presents hydrogen bonds with the O atom of the btz fragment (Figure 8.6(B)), while the latter, due to the combined effect of the bridging water molecules to the HIE247 (Figure 8.7(B)) and to the AMP co-factor (Figure 8.8(F)). On the other hand, both 90 and 270º conformations present some radial distribution of water molecules around the N atom of the btz moiety that arises as a consequence of the bridging water molecules with SER349 and GLY343, respectively (Figure 8.7(B)). Moreover, the 270º conformation also presents bridging water molecules with the GLY343 through the N atom of the tz moiety, which stabilises this twisted conformation by keeping the N atoms closer, and with AMP via the the N atom of the btz moiety, leading to an increase of its RDF around these two N atoms (Figure 8.3(G,H)).

As already mentioned in Section 5.3.1, RDF results are consistent with the number of hydrogen bonds between the water molecules and each of the electronegative atoms of the OLU represented in Figure 8.4. Both WT and SER286ASN systems present a similar amount of hydrogen bonds for all the O and N atoms. However, within the three different SCCS conformations

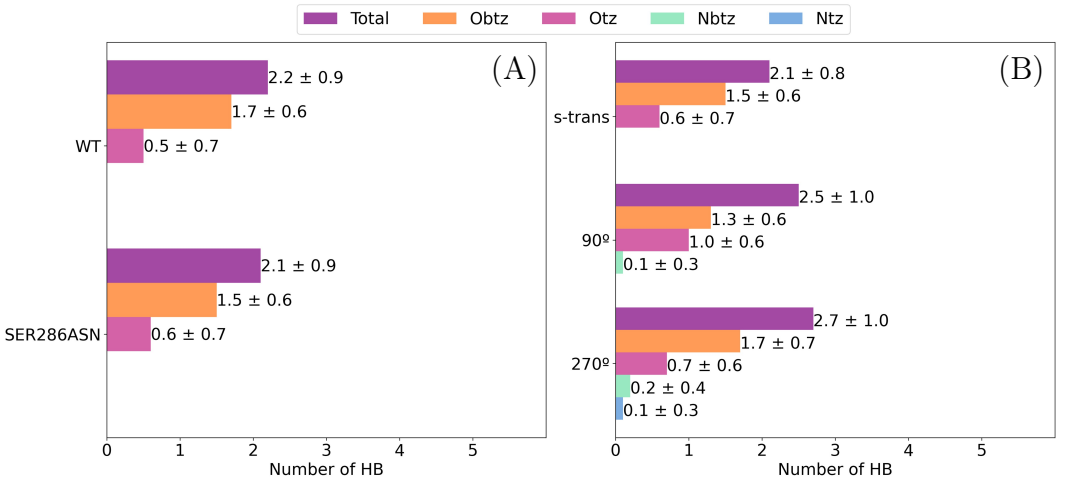


Figure 8.4: Average number of hydrogen bonds (HB) of the phenolate-keto OLU chromophore with water for each of the four electronegative atoms (A) in the WT and SER286ASN systems and (B) in the different SCCS conformations in the SER286ASN system.

in the SER286ASN system, the two distorted conformations present a larger amount of hydrogen bonds than the planar s-trans conformation. In the case of the 90^{o} configuration this increase in the hydrogen bonding is mostly due to a larger amount of hydrogen bonds with the O atom of the tz moiety as a consequence of water bridging with the AMP co-factor while for the 270^{o} configuration, which presents the largest amount of hydrogen bonds among the three of them, is due to the combination of the larger amount of hydrogen bonds with the O atom of the btz fragment and the two N atoms resulting from the water bridging with both the protein and the AMP.

8.3.3 OLU/luciferase Interactions

When looking at the amount of contacts between the phenolate-keto OLU and the WT luciferase and the SER286ASN mutant in Figure 8.5(A) to assess the OLU-luciferase interactions, it can be observed that the contacts pattern is very similar for both of them. Remarkably, the number of contacts of the chromophore with PHE249, the most relevant amino acid in terms of contacts, is lower in the mutant (49.8%) than in the WT luciferase (57.7%) due to a reduced number of contacts in both the s-trans (50.5%) and twisted conformations (43.9 and 32.3%). Nonetheless, as shown in Figure 8.5(B), this reduction is larger for the two twisted conformations than for the s-trans one. Moreover, the SER286ASN mutant also presents some other smaller differences with respect to the WT system in the contacts pattern coming from the other amino acids: (i) an increase in the contacts with (a) the polar SER349 (from 21.4 to 22.3%), THR345 (from 6.7 to 9.3%), TYR342 (from 7.3 to 8.7%) and GLN340 (from 0.3 to 2.9%) and (b) the non-polar LEU344 (from 12.4 to 16.0%), GLY318 (from 15.4 to 16.6%), GLY317 (from 6.5 to 8.3%), PHE252 (from 0.7 to 1.6%) and GLY248 (from 8.6 to 9.6%) and (ii) a decrease in the contacts with (a) the polar THR253 (from 8.0 to 7.0%), (b) the non-polar ALA350 (from 19.7 to 19.0%) and ILE288 (from 1.1 to 0.5%) and (c) the positively charged HIE247 (from 12.4 to 10.7%) and ARG220 (from 2.3 to 1.4%). These differences yield an environment that presents a larger amount of contacts with polar amino acids and a lower amount with non-polar and positively charged amino acids in the mutant (46.6, 158.9 and 14.7%, respectively) than in the WT luciferase (41.5, 173.7 and 17.6%). Hence, the OLU is surrounded by a more polar environment in the mutant than in the WT luciferase, which could favour the presence of the water molecules inside the active pocket.

This increase in the relative polarity of the active pocket is even more relevant for the twisted conformations than for the s-trans structure during the simulation of the SER286ASN mutant as the amount of contacts with non-

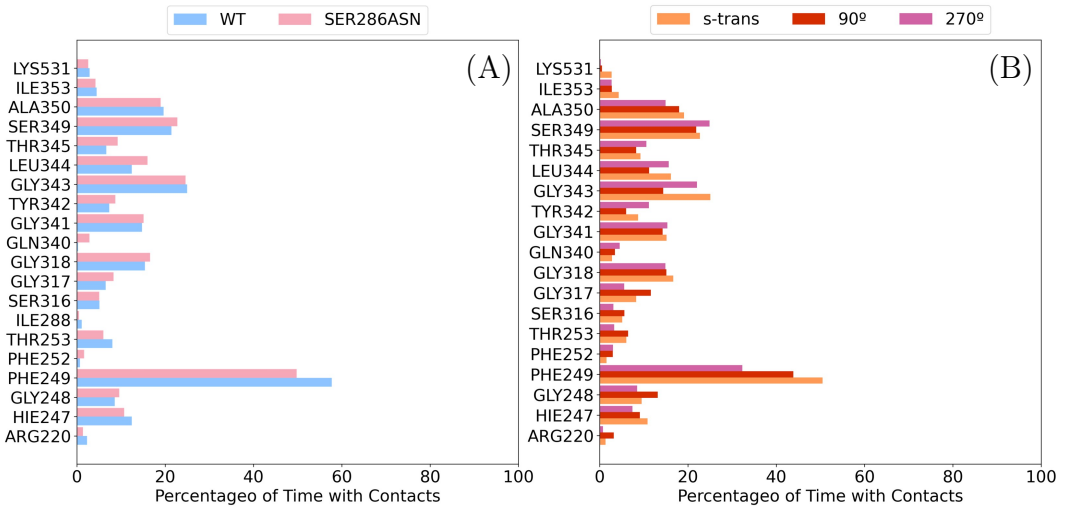


Figure 8.5: Percentage of time with contacts between (A) the phenolate-keto OLU and the WT and the SER286ASN luciferases and (B) the phenolate-keto OLU in the different SCCS conformations and the SER286ASN luciferase along the MD trajectory. Only those amino acids with contacts for at least the 1% of the simulation time have been included in the representation.

polar and positively charged amino acids in the twisted conformations (154.3 and 12.8% for the 90º conformation and 154.1 and 8.3% for the 270º conformation) is lower than with the s-trans conformation (174.7 and 15.0%). Changes on contacts with polar amino acids are compensated within each of the twisted conformations. However, this is not the case for with non-polar and positively charged amino acids. On the one hand, the 90º conformation presents, when compared to the s-trans conformation, a decrease in the contacts with the non-polar ILE353 (from 4.3 to 2.7%), ALA350 (from 19.1 to 18.0%), LEU344 (from 8.7 to 6.0%), GLY343 (from 25.0 to 14.4%), TYR42 (from 4.3 to 2.7%), GLY341 (from 15.1 to 14.3%), GLY318 (from 16.6 to 16.1%), GLY317 (from 4.3 to 2.7%) and PHE249 (from 50.5 to 43.9%) amino acids. Moreover, it also presents (i) an increase in the contacts with (a) the non-polar GLY317 (from 8.3 to 11.6%), GLY248 (from 9.5 to 13.1%) and PHE252 (from 1.5 to 3.0%) and (b) the positively charged ARG220 (from 1.4 to 3.2%) and (ii) a decrease with the positively charged LYS531 (from 2.7 to 0.5%) and HIE247 (from 10.9 to 9.1%). On the other hand, the 270º conformation presents a decrease in the contacts with all the positively charged amino acids (from 2.7 to 0.2% with LYS531, from 10.9 to 7.4% with HIE247 and from 1.4 to 0.7% with ARG220) and most of the non-polar amino acids (from 4.3 to 2.7% with

ILE353, from 19.1 to 14.9% with ALA350, from 25.0 to 22.1% with GLY343, from 16.6 to 14.9 with GLY318, from 8.3 to 5.5 with GLY317, from 50.5 to 32.3 with PHE249 and from 9.5 to 8.5% with GLY248).

The analysis of direct hydrogen bonding between the OLU and each amino acid of the WT and SER286ASN mutant enzymes in Figure 8.6(A) shows a different hydrogen bonding patter for each of the systems. While the chromophore presents a large amount of hydrogen bonds with LYS531 via the O atom of the tz moiety regardless of the enzyme, the mutant presents a significantly larger amount of hydrogen bonds with SER349, GLY341 and HIE247 with the N atom of the btz moiety, O atom of the btz moiety and the O atom of the tz moiety, respectively. This increase in the hydrogen bonding with respect to the WT system explains the slightly smaller RDF around the O atom of the btz moiety since a larger affinity to form hydrogen bonds with the protein reduces the availability of the electronegative atoms for hydrogen

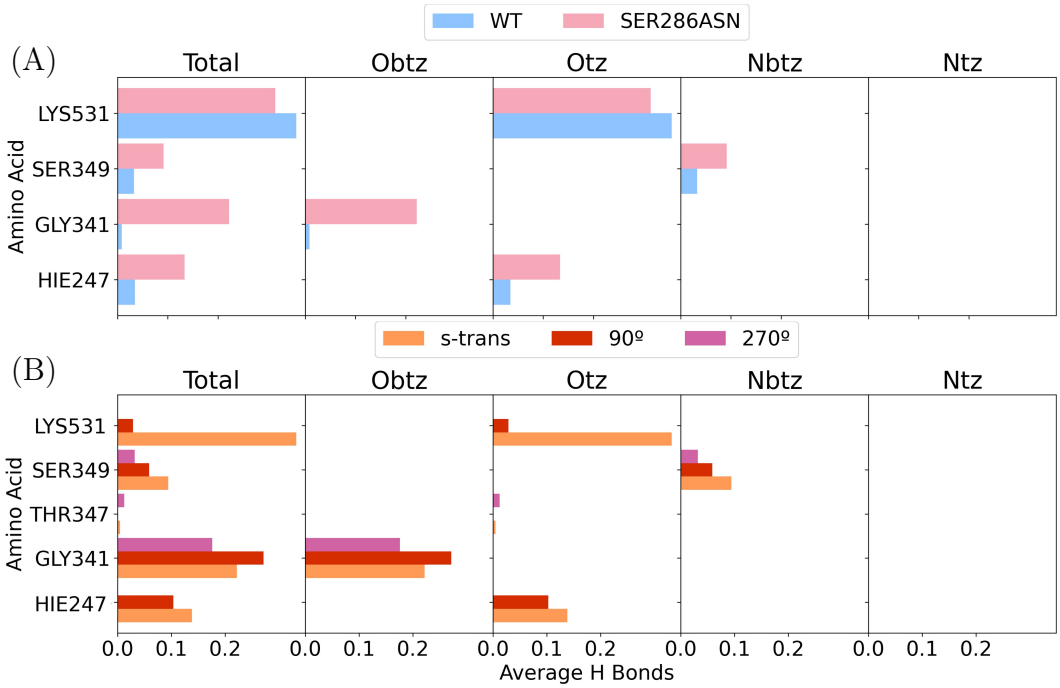


Figure 8.6: Average of hydrogen bonds (HB) between the phenolate-keto OLU with the enzyme for each of the four electronegative atoms (A) in the WT and SER286ASN systems and (B) in the different SCCS conformations in the SER286ASN system. Only those amino acids with at least an average of 0.1 of hydrogen bonds withe the OLU in at least one of the studied systems have been included in the representation.

bonding formation with the solvent.

Nonetheless, the hydrogen bonding pattern is very different for each of the SCCS conformers, as shown in Figure 8.6(B). Upon torsion, the hydrogen bond with LYS531 disappears and decreases with SER349. Moreover, in the case of the 270º conformer, also the hydrogen bonding with GLY341 is reduced and with HIE247 completely avoided. On the flip side, the 90º conformation presents a larger amount of hydrogen bonding with GLY341 than the s-trans conformation and some hydrogen bonding with HIE247 is retained. As previously discussed in Section 8.3.2, this reduction of direct hydrogen bonding for the twisted conformations with the protein with respect to the s-trans situation allows for a lager amount of hydrogen bonds between the OLU chromophore and the water molecules of the solvent resulting in a larger RDF for the non-planar configurations.

In addition to the direct hydrogen bonding with the two luciferases, and analogously to Chapter V, the amount of bridging water molecules connecting the chromophore and the enzyme through hydrogen bonding has also been characterized to gain a deeper understanding of the OLU/luciferase interactions. As shown in Figure 8.7(A), which represents the percentage of time that a water molecule is acting as a hydrogen bond bridge between the chromophore and a specific amino acid, the amount of bridging with the WT luciferase is larger than that with the SER286ASN mutant for all the represented amino acids with the exception of SER316, for which the water bridging with the mutant is slightly larger than with WT luciferase. Recalling the results in Figure 8.6(A), it can be said that the electrostatic interaction that implies the hydrogen bonding between the phenolate-keto OLU and the WT luciferase is mediated by bridging water molecules, while the SER286ASN mutation allows for a direct OLU/luciferase electrostatic interaction in the s-trans configuration.

However, the two twisted conformations present a different bridging pattern (Figure 8.7(B)) that connects the chromophore with different amino acids of the enzyme and reduces the amount of hydrogen bond bridges with ARG339 and SER316, which are the most important connections in the s-trans conformation. Specifically, the 90º conformation presents some bridging between the N atom of the btz moiety and SER349 and between the O atom of the tz moiety and HIE247, while the 270º conformation presents a large bridging with GLY343 via the two N atoms of the OLU chromophore, thus stabilising the twisted structure, and a small bridging between GLY318 and the O atom of the tz moiety. Hence, the smaller amount of direct hydrogen bonds with the protein observed in Figure 8.6(B) for the twisted conformations is compensated with a larger amount bridging connecting the chromophore and the

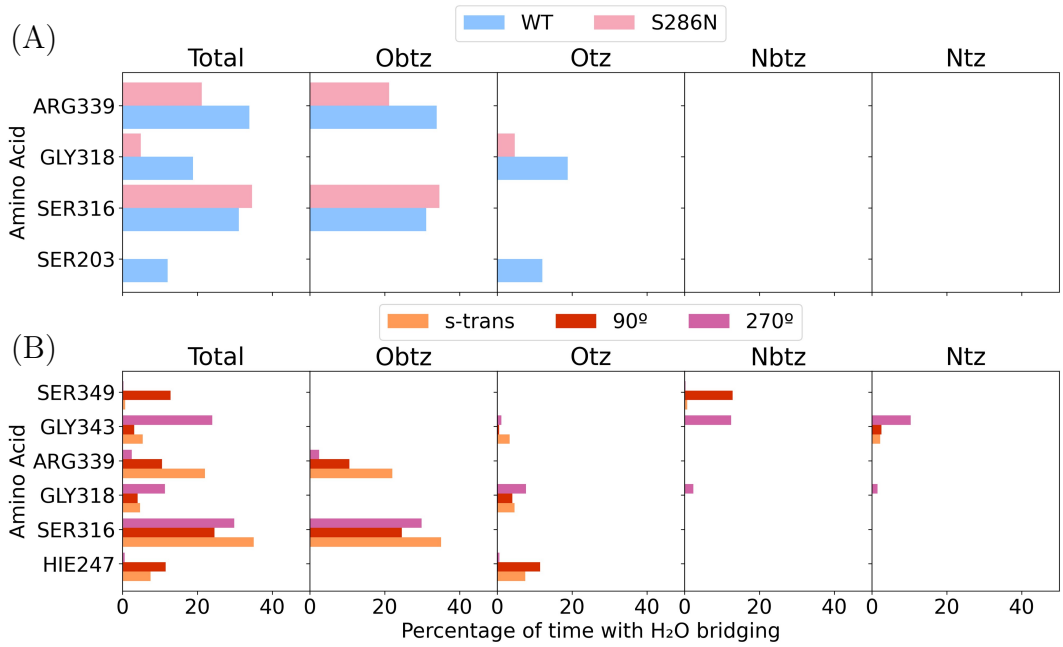


Figure 8.7: Percentage of time with a water molecule acting as a hydrogen bond (HB) bridge between OLU chromophore an the enzyme (A) in the WT and SER286ASN systems and (B) in the different SCCS conformations in the SER286ASN system. Only those amino acids with a bridging water molecule for at least a 20 % of the simulation time of one of the studied systems have been included in the representation.

enzyme via indirect electrostatic interactions.

8.3.4 The Role of AMP

Finally, it is also interesting to analyse the role of the AMP co-factor inside the active site. As it can be observed in Figure 8.8(A), the distance between the centre of mass of the OLU and the AMP is around 7 Å. However, the co-factor can move farther away in the active site of the mutant, regardless of the SCCS conformation (Figure 8.8(B)). Moreover, Figure 8.8(C) shows that the number of direct hydrogen bonds between the OLU and the AMP co-factor is slightly larger in the SER286ASN mutant while the percentage of time with a bridging water molecule is smaller (Figure 8.8(E)) than in the WT system. However, this result is once again biased by relative amount of sampled geometries on each conformation. On the one hand, both twisted structures, specially the 270º conformer, present a significantly larger direct hydrogen bonding with the

co-factor via the O atom of the tz moiety than the s-trans conformer (Figure 8.8(D)), which in fact presents a smaller hydrogen bonding in the mutant than in the WT luciferase. On the other hand, they also present a larger percentage of time with a bridging water molecule via the O atom of the tz moiety and, in the case of the 270º conformer, also via the N atom of the btz moiety (Figure 8.8(F)). Hence, in the SER286ASN mutant, the OLU chromophore is almost always connected either directly or via a bridging water molecule to the AMP co-factor.

Moreover, the significant increase in the direct hydrogen bonding with the AMP for the twisted structures compensates, specially for the 270º conformer, the loss of the direct hydrogen bond with the LYS531 amino acid (Figure 8.6(B)). However, the large amount of water bridging increases the RDF around the O atom of tz fragment for the 90º conformer and retains some solvation sphere around the N atom of btz fragment for the 270º conformer.

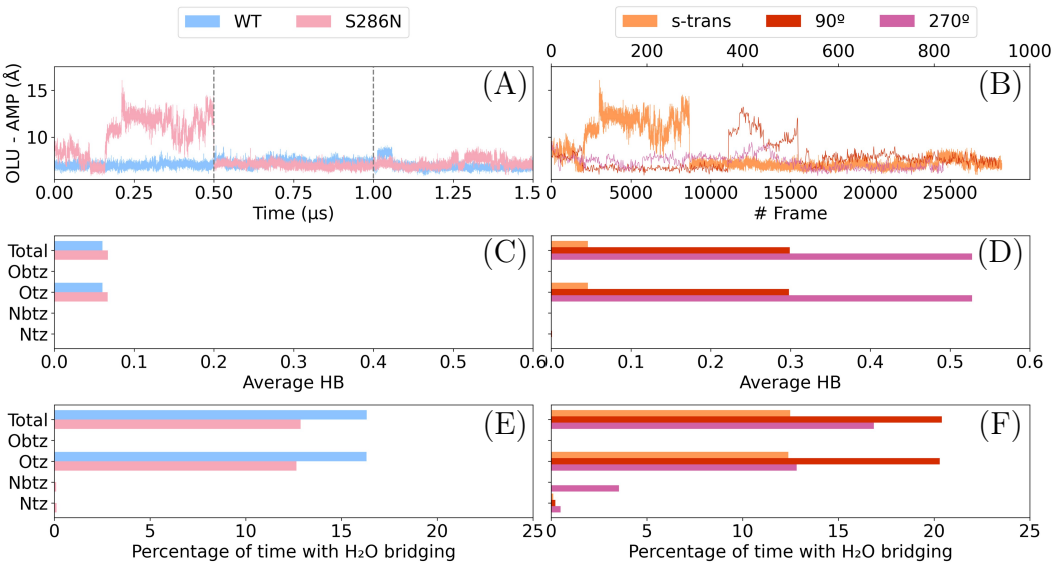


Figure 8.8: (A,B) Distance between the OLU chromophore to the AMP co-factor, (C,D) average of hydrogen bonds (HB) between the OLU chromophore and the AMP and (E,F) percentage of time with a water molecule acting as a HB bridge between the OLU chromophore and the AMP. Results for the MD trajectories in the (A,C,D) WT and SER286ASN systems and (B,D,F) the different SCCS conformations in the SER286ASN system. Vertical dashed lines in (A) represent the separation into the three different replicas. Secondary x axis in (B) applies to 90 and 270º conformations.

8.4 Conclusions

Despite the huge interest of the scientific community on firefly's BL emission, there are still many unknowns about the system responsible of it. Due to its valuable contribution to a wide variety of applications, it has been a strong effort to unravel the biochemical explanation behind the colour of the emitted light, which has led to the hypothesis discussed in Section 1.2.3. Within this context, this chapter presents an extensive structural and intermolecular interaction analysis on two different luciferase environments of the phenolate-keto OLU in the S_1 state by means of a MD sampling based on the OLUFF to determine the flexibility around the SCCS torsion of the chromophore, and to characterize the intermolecular interactions with the water molecules, protein residues and the AMP cofactor through contact and hydrogen bonding analyses.

It has been shown that the OLU inside the active pocket of the SER286ASN mutant is able to sample the twisted conformations with a SCCS dihedral angle of 90 and 270º, while in the WT luciferase the structure is confined in the s-trans region. These twisted conformations of the SER286ASN mutant present a different solvation pattern than the s-trans conformation, for which it is very similar to that of the WT system. Despite this similarity, it can be observed that the solvation on the WT system is slightly larger as a result of the lower amount of direct hydrogen bonds and the larger amount of water bridges between the chromophore and both the enzyme and the AMP cofactor. In addition, the twisted structures show a larger solvation interaction around the O atom of the btz moiety and the O atom of the tz moiety for the 270º and 90º conformations, respectively. In contrast to the s-trans conformation of both the WT and SER286ASN mutant luciferases, the twisted structures also present some water molecules around the N atom of the btz moiety and, in the case of the 270º conformer, also around the N atom of the tz moiety. This increase in the solvation pattern for the twisted conformations is caused by the significant reduction of direct hydrogen bonding with the enzyme and the increase of the water bridging with the both the enzyme and AMP, that is able to not only to compensate but to overcome the significative increase on the direct hydrogen bonding with the cofactor.

Remarkably, it has also been observed that the number of contacts, which qualitatively reflects the van der Waals non-bonding interactions, of PHE249 with the chromophore is lower in the mutant than in the WT luciferase, specially for the twisted conformations. This result, together with the smaller changes in the contacts pattern, leads to an environment that is slightly more polar in the mutant than in the WT luciferase, which stabilises the presence

of the water molecules inside the active pocket.

The results presented along this chapter demonstrate the structural differences between the WT and SER286ASN mutant luciferases employing a dynamical approach. Although a quantitative characterization of the OLU/protein interactions and the calculation of the emission spectra could complement this structural analysis, this chapter clearly shows that the active site leads to a larger flexibility of the SCCS torsion and a more polar microenvironment around the chromophore on the SER286ASN mutant, in agreement with the existing literature.[124, 363]

Moreover, these results are obtained on the assumption that the excited state OLU is obtained in the s-trans conformer. However, the different interaction pattern with the enzyme and the more flexible active site of the SER286ASN mutant, as a consequence of the ILE288 not moving towards the chromophore after the adenylation reaction,[124, 363] could lead to an initial (partially) twisted conformation of the resulting excited state OLU after the oxidation reaction and so to a larger population of the twisted conformation in the emitting state. In order to prove this hypothesis, the mechanism of the oxidation reaction should be studied at QM/MM level of theory for both the WT and the SER286ASN mutant, which is out of the scope of this chapter.

Chapter IX

Conclusions

This PhD thesis has focused on the computational characterization of firefly's BL system, the OLU/luciferase complex, with the objective of improving the understanding of its colour tuning mechanism. To do so, a wide variety of computational modelling techniques have been employed. Specifically, classical MD simulations have been used to sample the conformational space of the chromophore in four different media, namely: (i) *in vacuum*, (ii) in aqueous media and in the active site of (iii.a) the WT luciferase and (iii.b) the SER286ASN mutant. Moreover, QM and QM/MM calculations have been employed to study the electronic properties of the system and the OLU/protein interactions. Firstly, in Chapter III the GAFF-based sampling was compared to the QM/MM-based one, stressing the need of a specific FF of the OLU chromophore. Hence, OLUFF has been parameterised in Chapter IV to perform the sampling of the system. This new FF, obtained from *in vacuum* QM calculations, has been employed to dynamically characterise the active pocket of (i) the WT enzyme for the four possible emitting tautomers in Chapters V and VII and (ii) the SER286ASN mutant for the red-emitter phenolate-keto specie in Chapter VIII. Moreover, the *in vacuum* OLUFF sampling has also been employed in Chapter VI to study the nuclear motion of the phenolate-keto chromophore that yields the energy shift, stressing the role of the typically disregarded TICT mechanism. From these analysis the following conclusions of the present thesis can be drawn:

- When comparing GAFF- and QM/MM-based samplings for the phenolate-keto OLU in Chapter III, it has been shown that classical trajectories are not able to properly sample the range of the SCCS torsional angle. However, this motion presents a relevant role for the red shift of the absorption and emission spectra. Therefore, for the success of this thesis, it was crucial to develop a set of FF parameters that could accurately sample the conformational space of the chromophore. This parametrization was performed, following the procedure discussed in Chapter IV, using QM calculations as energetic reference for the obtention of the torsional periodic function with several terms. As a result, the OLUFF has been presented and validated for the four possible emitters of the chromophore in both the S_0 and S_1 electronic states. The validation, which consisted on the comparison of probability distribution functions of the dihedral angle, minimized structures, IR spectra and PC analysis computed from FF-based MD simulations with analogous QM data, has shown that the OLUFF performs better than the GAFF and that it can be used to further investigate the OLU/luciferase complex and unravel the nature of the enzyme/substrate interactions.

- The simplified static model of the phenolate-keto OLU *in vacuum* presented in Chapter III, in which three differentiated minima of the S_1 state were found at 90, 180 (s-trans) and 270º, has shown that red shift of the S_0/S_1 electronic transition can be achieved upon the torsion of the SCCS dihedral angle. At the same time, it has also shown that this torsional distortion leads to an increase of the CT character and a decrease of the oscillator strength of the electronic transition. Nonetheless, this correlation has not been not found when analysing both the classical and QM/MM dynamics simulations, as CT character in the s-trans conformation is not strongly correlated to the red shift of emission, as shown in Chapter VI. Upon further investigation of the conformational changes that lead to the emission energy shifts by fitting the projection of the MW displacements of the PCs of the system to the emission energy, it has been shown that only in-plane tensions within each of the two moieties of the phenolate-keto OLU chromophore contribute to the energy shift dimension for each of the three S_1 minima along the SCCS dihedral torsion. Although the oscillator strength dimension is yet to be analysed, it has been observed that, for the s-trans minima, this coordinate has to be highly correlated to the energy shift dimension. On the flip side, for the two twisted conformations the energy shift dimension is not correlated to neither the oscillator strength nor to the CT of the transition and the brightness of the emission for this conformations is, thus, facilitated by a different combination of the vibrational motions of the system.

- Chapter V shows that both *in vacuum* and aqueous OLUFF-based simulations present a similar probability distribution around the SCCS dihedral angle. In contrast, inside the active site of the WT enzyme the movement is sterically constrained and, therefore, any of the OLU chromophores can populate the twisted conformations. Moreover, it has been shown that, in aqueous solution, the solvation around the heteroatoms of the chromophores is directly related to their negative atomic charge while, in the active site of the WT protein, the solvation pattern is different from that obtained in water. On the one hand, fewer solvent molecules can diffuse inside the cavity and, on the other hand, there is a competition between direct interactions with the protein and AMP, and with water molecules. However, this competition is balanced by bridging interactions through water molecules, that increase the chromophore solvation.

- The OLUFF is not suitable for the calculation of the OLU/luciferase

interaction energies by means of the classical 1A-MM-GBSA approach. However, it can be employed to qualitatively select the amino acids of the protein that need to be included in the QM region for the calculation and characterization of the interaction energies by means of the QM/MM-EDA approach. When doing so, it has been observed that the electrostatic term governs the interaction energy in the S_1 state for the four possible OLU emitters, specially for the phenolate-enolate chromophore. This dianionic OLU species is able to attract the lateral chains of two positively charged ARG amino acids in the active pocket, thus, increasing the electrostatic term of the interaction energy. Furthermore, non-electrostatic terms also contribute to the interaction energy since the active site of the protein presents amino acids, as PHE249, with π-conjugated systems.

- When the phenolate-keto OLU simulations in S_1 state inside the WT and the red-emitting SER286ASN mutant are compared in Chapter VIII, it has been shown that the chromophore inside the mutant is able to sample the two SCCS twisted conformations while, in the WT luciferase, it is confined in the s-trans region. The twisted conformations inside the mutant present an increased solvation pattern with respect to the s-trans conformation inside the pocket of both the WT and the mutant luciferases caused by the reduced direct hydrogen bonding with the enzyme and the increased water bridging with the the enzyme and the AMP cofactor. This is able to overcome the significative increase on the direct hydrogen bonding with the cofactor, that is also larger for the twisted conformations than for the s-trans. Moreover, the local protein environment surrounding the chromophore is slightly more polar in the mutant than in the WT luciferase, stabilising the presence of the water molecules inside the active pocket.

In summary, this PhD thesis has provided, by means of a dynamical approach, a detailed description of the OLU chromophore inside the active site of the luciferase enzyme in both the electronic ground and first excited state, characterising not only the ligand/protein interaction at different levels of theory but also analysing the role of the water molecules in the hydrogen bond pattern of the pocket. Moreover, the electronic properties of the emission process have been related to different DOF of the chromophore, resurfacing the usually disregarded TICT red-shift mechanism. Moreover, a FF that allows for the sampling of the OLU chromophore, the OLUFF, has been proposed, thus allowing other groups to employ the MD machinery to further investi-

gate either the WT complex or its mutants to provide meaningful data on the system.

As already mentioned in Chapter VI, future work should focus on characterising the nuclear coordinate that leads to the increase of the oscillator strength of the S_0/S_1 electronic transition. This coordinate could be employed in combination with the energy shift dimension to predict the emission energy and the brightness of the transition inside different luciferase enzymes without the need of performing additional QM calculations. Furthermore, as mentioned in Chapter VIII, it would also be interesting to study the oxidation mechanism of the luciferyl adenilate to generate the OLU species in the excited state inside the active site to verify the possibility of a SCCS twisted intermediate. Finally, in order to characterise the effect of changing the amino acids of the active site, the systematic study of ALA mutations should also be performed.

Chapter X

Conclusiones

Esta tesis doctoral se ha centrado en la caracterización computacional del sistema BL de las luciérnagas, que consiste en el complejo OLU/luciferasa, con el objetivo de mejorar la comprensión de su mecanismo de modulación del color de la luz emitida. Para ello, se ha empleado una amplia variedad de técnicas de modelización computacional. En concreto, se han utilizado simulaciones de MD clásica para muestrear el espacio conformacional del cromóforo en cuatro medios diferentes, a saber: (i) en vacío, (ii) en medio acuoso y en el sitio activo de (iii.a) la luciferasa WT y (iii.b) el mutante SER286ASN. Además, se han empleado cálculos QM y QM/MM para estudiar las propiedades electrónicas del sistema y las interacciones OLU/proteína. En primer lugar, en el Capítulo III se comparó el muestreo basado en GAFF con el basado en un potencial QM/MM, destacando la necesidad de un FF específico para el cromóforo OLU. Por ello, en el Capítulo IV se parametrizó el OLUFF para realizar el muestreo del sistema. Este nuevo FF, obtenido a partir de cálculos QM en vacío, se empleó para caracterizar dinámicamente el sitio activo de (i) la enzima WT para los cuatro tautómeros emisores posibles en los Capítulos V y VII y (ii) el mutante SER286ASN para la especie emisora de luz roja, el fenolato-ceto, en el Capítulo VIII. Además, también se empleó el muestreo en vacío para estudiar, en el Capítulo VI, el movimiento nuclear del cromóforo fenolato-ceto que lleva al desplazamiento hacia el rojo, destacando el papel del mecanismo TICT que normalmente se desestima. A partir de estos análisis, se pudieron extraer las siguientes conclusiones de la presente tesis:

- Al comparar los muestreos basados en el GAFF y en un potencial QM/MM para el tautómero fenolato-ceto de la OLU en el Capítulo III, se demostró que las trayectorias clásicas no son capaces de muestrear adecuadamente el rango del ángulo de torsión SCCS. Sin embargo, este movimiento presenta un papel relevante para el desplazamiento al rojo de los espectros de absorción y emisión. Por lo tanto, para el éxito de esta tesis, era crucial desarrollar un conjunto de parámetros de torsión que pudieran muestrear con precisión el espacio conformacional del cromóforo. Esta parametrización se realizó, siguiendo el procedimiento discutido en el Capítulo IV, utilizando cálculos QM como referencia energética para la obtención de la función periódica de torsión con varios términos. Como resultado, se presentó y validó el campo de fuerzas OLUFF para los cuatro emisores posibles del cromóforo en los estados electrónicos S_0 y S_1. La validación, que consistió en la comparación de resultados QM con funciones de distribución de probabilidad del ángulo diedro, estructuras minimizadas, espectros IR y el análisis de los PC calculados a partir de simulaciones MD basadas en diferentes FFs; demostró que el OLUFF fun-

ciona mejor que GAFF y que puede utilizarse para investigar en mayor profundidad el complejo OLU/luciferasa y desentrañar la naturaleza de las interacciones entre la enzima y el sustrato.

- El modelo estático simplificado del fenolato-ceto OLU en vacío presentado en el Capítulo III, que presenta tres mínimos diferenciados del estado electrónico excitado S_1 a 90, 180 (s-trans) y 270º, mostró que el desplazamiento al rojo de la transición electrónica S_0/S_1 puede lograrse mediante la torsión del ángulo diedro SCCS. Al mismo tiempo, también mostró que esta torsión conduce al aumento del carácter CT y a la disminución de la fuerza osciladora de la transición electrónica. No obstante, esta correlación no se encontró al analizar ni las simulaciones clásicas ni las QM/MM, ya que el carácter CT en la conformación s-trans no está fuertemente correlacionado con el desplazamiento al rojo de la emisión, como se muestra en el Capítulo VI. Tras ajustar las proyecciones de los desplazamientos MW a la energía de emisión, se llevó a cabo una investigación más detallada de los cambios conformacionales que conducen a los desplazamientos de la energía de emisión. Se demostró que solo las tensiones en el plano dentro de cada uno de los dos fragmentos del cromóforo fenolato-ceto contribuyen a la dimensión del desplazamiento de energía para cada uno de los tres mínimos del S_1 a lo largo de la torsión SCCS. Aunque aún no se ha analizado la dimensión de la fuerza del oscilador, se observó que, para los mínimos s-trans, esta coordenada tiene que estar altamente correlacionada con la dimensión del desplazamiento de energía. Por otro lado, para las otras dos conformaciones, la dimensión del desplazamiento de energía no está correlacionada ni con la fuerza del oscilador ni con el carácter CT de la transición y, por lo tanto, el brillo de la emisión para estas conformaciones se ve facilitado por una combinación diferente de los movimientos vibracionales del sistema.

- En el Capítulo V se mostró que tanto las simulaciones en vacío como las acuosas presentan una distribución de probabilidad similar en torno al ángulo diedro SCCS. Por el contrario, dentro del sitio activo de la enzima WT, el movimiento está estéricamente restringido, por lo que ninguno de los tautómeros de la OLU puede ocupar una conformación diferente a la s-trans. Además, se demostró que, en solución acuosa, la solvatación alrededor de los heteroátomos de los cromóforos está directamente relacionada con su carga atómica negativa, mientras que, en el sitio activo de la proteína WT, el patrón de solvatación es diferente al obtenido en agua. Por un lado, menos moléculas de disolvente pueden difundir dentro de la cavidad y, por otro, existe una competencia entre

las interacciones directas con la proteína y el AMP, y con las moléculas de agua. Sin embargo, esta competencia se equilibra al considerar las moléculas de agua que hacen de puente entre los crómóforos y la proteína o el cofactor, aumentando la solvatación de la OLU.

- El OLUFF no es adecuado para el cálculo de la energía de interacción OLU/luciferasa mediante el tratamiento clásico 1A-MM-GBSA. Sin embargo, puede emplearse para seleccionar cualitativamente los aminoácidos de la proteína que deben incluirse en la región QM para el cálculo y la caracterización de las energías de interacción con el método QM/MM-EDA. Al hacerlo, se observó que el término electrostático gobierna la energía de interacción en el estado S_1 para los cuatro emisores posibles de la OLU, especialmente para el cromóforo fenolato-enolato. Esta especie dianiónica es capaz de atraer las cadenas laterales de dos argininas (con carga positiva) del sitio activo, aumentando así el término electrostático de la energía de interacción. Además, los términos no electrostáticos también contribuyen a la energía de interacción, ya que el sitio activo de la proteína presenta aminoácidos, como la PHE249, con sistemas π conjugados.

- Cuando se compararon las simulaciones del tautómero fenolato-ceto de la OLU en el estado S_1 en el sitio activo de la luciferasa WT y del mutante SER286ASN emisor de luz roja en el Capítulo VIII, se demostró que el cromóforo dentro del mutante es capaz de muestrear las tres conformaciones del diedro SCCS, mientras que, en la luciferasa WT, el muestreo está confinado en la región s-trans. Además, las conformaciones no planas dentro del mutante presentan una mayor solvatación con respecto a la conformación s-trans tanto dentro la luciferasa WT como dentro de la mutación, causado por la reducción de los enlaces de hidrógeno directos con la enzima y el aumento del número de moléculas de agua puente entre el cromóforo y la enzima o el cofactor AMP. Este incremento es capaz de superar el aumento significativo de los enlaces de hidrógeno directos con el cofactor, que también es mayor para las conformaciones torcidas que para la s-trans. Finalmente, el entorno proteico local que rodea al cromóforo también es ligeramente más polar en la mutante que en la luciferasa WT, estabilizando la presencia de las moléculas de agua dentro del sitio activo.

En resumen, esta tesis doctoral proporciona, mediante un enfoque dinámico, una descripción detallada del cromóforo OLU dentro del sitio activo de la enzima luciferasa, tanto en el estado fundamental como en el primer estado

electrónico excitado. No solo se caracterizó la interacción ligando/proteína a diferentes niveles de teoría, sino también analizó el papel de las moléculas de agua en el patrón de los enlaces de hidrógeno del sitio activo de la enzima. Además, las propiedades electrónicas del proceso de emisión fueron relacionadas con diferentes DOF del cromóforo, reviviendo el mecanismo de desplazamiento al rojo TICT. Además, se ha propuesto el OLUFF, un FF que permite el muestreo del cromóforo OLU, permitiendo a otros grupos emplear la MD e investigar más a fondo el complejo WT o sus mutantes con el fin de proporcionar datos significativos sobre el sistema.

Como ya se mencionó en el Capítulo VI, futuras investigaciones deberían centrarse en caracterizar la coordenada nuclear que conduce al aumento de la fuerza osciladora de la transición electrónica S_0/S_1. Esta coordenada podría emplearse, en combinación con la dimensión del desplazamiento energético, para predecir la energía de emisión y el brillo de la transición dentro de diferentes enzimas luciferasas sin necesidad de realizar cálculos QM adicionales. Además, como se menciona en el Capítulo VIII, también sería interesante estudiar el mecanismo de oxidación del adenilato luciferil a la especie OLU en el estado excitado dentro del sitio activo, con el fin de verificar la posibilidad de un intermediario con el dihedro SCCS distorsionado. Por último, con el fin de caracterizar el efecto del cambio de los aminoácidos del sitio activo, también debería realizarse el estudio sistemático de mutaciones por ALA.

Bibliography

(1) Kahlke, T.; Umbers, K. D. *Current Biology* **2016**, *26*, R313–R314.

(2) Kaskova, Z. M.; Tsarkova, A. S.; Yampolsky, I. V. *Chemical Society Reviews* **2016**, *45*, 6048–6077.

(3) Roda, A. In *Chemiluminescence and Bioluminescence*; The Royal Society of Chemistry: 2010.

(4) Boyle, R. *Philosophical Transactions of the Royal Society of London* **1668**, *2*, 581–600.

(5) Vacher, M.; Fdez. Galván, I.; Ding, B.-W., et al. *Chemical reviews* **2018**, *118*, 6927–6974.

(6) Malpighi, M., *Observationes anatomicaes in plantis and animalibus*, 1687.

(7) Harvey, E. N., *A history of luminescence from the earliest times until 1900*; American Philosophical Society: 1957; Vol. 44.

(8) Harvey, E. N. *Science* **1914**, *40*, 33–34.

(9) Harvey, E. N.; Stevens, K. P. *The Journal of General Physiology* **1928**, *12*, 269.

(10) White, E. H.; McCapra, F.; Field, G. F. *Journal of the American Chemical Society* **1963**, *85*, 337–343.

(11) McElroy, W.; Hastings, J.; Sonnenfeld, V., et al. *Science* **1953**, *118*, 385–386.

(12) McElroy, W. D.; Seliger, H. H.; White, E. H. *Photochemistry and Photobiology* **1969**, *10*, 153–170.

(13) Green, A. A.; McElroy, W. D. *Biochimica et biophysica acta* **1956**, *20*, 170–176.

(14) Tsuji, F. I.; Sowinski, R. *Journal of cellular and comparative physiology* **1961**, *58*, 125–129.

(15) Shimomura, O.; Johnson, F. H.; Saiga, Y. *Journal of cellular and comparative physiology* **1962**, *59*, 223–239.

(16) Evans, J. F.; McCRACKEN, S.; Miyamoto, C. M., et al. *Journal of Bacteriology* **1983**, *153*, 543–545.

(17) Prasher, D. C.; Eckenrode, V. K.; Ward, W. W., et al. *Gene* **1992**, *111*, 229–233.

(18) Seliger, H.; McElroy, W.; White, E., et al. *Proceedings of the National Academy of Sciences* **1961**, *47*, 1129–1134.

(19) Hastings, J. W.; Gibson, Q. H. *Journal of Biological Chemistry* **1963**, *238*, 2537–2554.

(20) Hart, R. C.; Stempel, K. E.; Boyer, P. D., et al. *Biochemical and Biophysical Research Communications* **1978**, *81*, 980–986.

(21) Hart, R. C.; Matthews, J. C.; Hori, K., et al. *Biochemistry* **1979**, *18*, 2204–2210.

(22) Kopecky, K. R.; Mumford, C. *Canadian Journal of Chemistry* **1969**, *47*, 709–711.

(23) McElroy, W. D. *Proceedings of the National Academy of Sciences* **1947**, *33*, 342–345.

(24) White, E. H.; Wörther, H.; Seliger, H. H., et al. *Journal of the American Chemical Society* **1966**, *88*, 2015–2019.

(25) White, E. H.; Branchini, B. R. *Journal of the American Chemical Society* **1975**, *97*, 1243–1245.

(26) Cohn, D. H.; Ogden, R. C.; Abelson, J. N., et al. *Proceedings of the National Academy of Sciences* **1983**, *80*, 120–123.

(27) Baldwin, T. O.; Berends, T.; Bunch, T. A., et al. *Biochemistry* **1984**, *23*, 3663–3667.

(28) Devine, J. H.; Kutuzova, G. D.; Green, V. A., et al. *Biochimica et Biophysica Acta (BBA)-Gene Structure and Expression* **1993**, *1173*, 121–132.

(29) Inouye, S.; Noguchi, M.; Sakaki, Y., et al. *Proceedings of the National Academy of Sciences* **1985**, *82*, 3154–3158.

(30) Branchini, B. R.; Southworth, T. L.; DeAngelis, J. P., et al. *Comparative Biochemistry and Physiology Part B: Biochemistry and Molecular Biology* **2006**, *145*, 159–167.

(31) Fleiss, A.; Sarkisyan, K. S. *Current genetics* **2019**, *65*, 877–882.

(32) Inoue, S.; Kakoi, H.; Goto, T. *Journal of the Chemical Society, Chemical Communications* **1976**, 1056–1057.

(33) Hamdan, F. F.; Audet, M.; Garneau, P., et al. *SLAS Discovery* **2005**, *10*, 463–475.

(34) Prinz, A.; Diskar, M.; Herberg, F. W. *ChemBioChem* **2006**, *7*, 1007–1012.

(35) Chung, E.; Yamashita, H.; Au, P., et al. *PloS one* **2009**, *4*, e8316.

(36) Saito, K.; Chang, Y.; Horikawa, K., et al. *Nature communications* **2012**, *3*, 1262.

(37) Shimomura, O.; Goto, T.; Hirata, Y. *Bulletin of the Chemical Society of Japan* **1957**, *30*, 929–933.

(38) Abraham, E. H.; Salikhova, A. Y.; Hug, E. B. *Drug development research* **2003**, *59*, 152–160.

(39) Noguchi, T.; Ikeda, M.; Ohmiya, Y., et al. *PLoS One* **2012**, *7*, e37093.

(40) Yamada, Y.; Nishide, S.-y.; Nakajima, Y., et al. *Analytical Biochemistry* **2013**, *439*, 80–87.

(41) Miesenböck, G.; Rothman, J. E. *Proceedings of the National Academy of Sciences* **1997**, *94*, 3402–3407.

(42) Wu, C.; Kawasaki, K.; Ogawa, Y., et al. *Analytical Chemistry* **2007**, *79*, 1634–1638.

(43) Wu, C.; Irie, S.; Yamamoto, S., et al. *Luminescence* **2009**, *24*, 131–133.

(44) Daubner, S. C.; Astorga, A. M.; Leisman, G. B., et al. *Proceedings of the National Academy of Sciences* **1987**, *84*, 8912–8916.

(45) Björkman, K. M.; Karl, D. M. *Journal of microbiological methods* **2001**, *47*, 159–167.

(46) Hastings, J. W.; Sweeney, B. M. *Journal of Cellular and Comparative Physiology* **1957**, *49*, 209–225.

(47) Bode, V. C.; Hastings, J. W. *Archives of Biochemistry and Biophysics* **1963**, *103*, 488–499.

(48) Dunlap, J. C.; Hastings, J. *Biochemistry* **1981**, *20*, 983–989.

(49) Nakamura, H.; Kishi, Y.; Shimomura, O., et al. *Journal of the American Chemical Society* **1989**, *111*, 7607–7611.

(50) Shimomura, O.; Johnson, F. H. *Biochemistry* **1967**, *6*, 2293–2306.

(51) Shimomura, O. *Journal of bioluminescence and chemiluminescence* **1995**, *10*, 91–101.

(52) Wilson, T.; Hastings, J. W. *Annual review of cell and developmental biology* **1998**, *14*, 197–230.

(53) Purtov, K. V.; Petushkov, V. N.; Baranov, M. S., et al. *Angewandte Chemie* **2015**, *127*, 8242–8246.

(54) Stevani, C. V.; Oliveira, A. G.; Mendes, L. F., et al. *Photochemistry and Photobiology* **2013**, *89*, 1318–1326.

(55) Kaskova, Z. M.; Dörr, F. A.; Petushkov, V. N., et al. *Science advances* **2017**, *3*, e1602847.

(56) Kotlobay, A. A.; Sarkisyan, K. S.; Mokrushina, Y. A., et al. *Proceedings of the National Academy of Sciences* **2018**, *115*, 12728–12732.

(57) Shimomura, O.; Johnson, F. H. *Biochemistry* **1968**, *7*, 1734–1738.

(58) Shimomura, O.; Johnson, F. H. *Biochemistry* **1968**, *7*, 2574–2580.

(59) Rodionova, N. S.; Rota, E.; Tsarkova, A. S., et al. *Photochemistry and Photobiology* **2017**, *93*, 416–428.

(60) Rudie, N. G.; Mulkerrin, M. G.; Wampler, J. E. *Biochemistry* **1981**, *20*, 344–350.

(61) Mulkerrin, M. G.; Wampler, J. E. In *Methods in Enzymology*; Elsevier: 1978; Vol. 57, pp 375–381.

(62) Wampler, J. E.; Mulkerrin, M. G.; Rich Jr, E. S. *Clinical chemistry* **1979**, *25*, 1628–1634.

(63) Petushkov, V.; Rodionova, N. In *Doklady Biochemistry & Biophysics*, 2005; Vol. 401.

(64) Petushkov, V. N.; Dubinnyi, M. A.; Tsarkova, A. S., et al. *Angewandte Chemie* **2014**, *126*, 5672–5674.

(65) Yue, L.; Liu, Y.-J.; Fang, W.-H. *Journal of the American Chemical Society* **2012**, *134*, 11632–11639.

(66) Koo, J.-Y.; Schuster, G. B. *Journal of the American Chemical Society* **1977**, *99*, 6107–6109.

(67) Schmidt, S. P.; Schuster, G. B. *Journal of the American Chemical Society* **1978**, *100*, 1966–1968.

(68) Schuster, G. B. *Accounts of Chemical Research* **1979**, *12*, 366–373.

(69) Schmidt, S. P.; Schuster, G. B. *Journal of the American Chemical Society* **1980**, *102*, 7100–7103.

(70) Adam, W.; Bronstein, I.; Trofimov, A. V., et al. *Journal of the American Chemical Society* **1999**, *121*, 958–961.

(71) Takano, Y.; Tsunesada, T.; Isobe, H., et al. *Bulletin of the Chemical Society of Japan* **1999**, *72*, 213–225.

(72) Isobe, H.; Takano, Y.; Okumura, M., et al. *Journal of the American Chemical Society* **2005**, *127*, 8667–8679.

(73) Catalani, L. H.; Wilson, T. *Journal of the American Chemical Society* **1989**, *111*, 2633–2639.

(74) McCapra, F. *Journal of Photochemistry and Photobiology A: Chemistry* **1990**, *51*, 21–28.

(75) Wilson, T. *Photochemistry and Photobiology* **1995**, *62*, 601–606.

(76) Wang, M.-Y.; Liu, Y.-J. *The Journal of Organic Chemistry* **2021**, *86*, 1874–1881.

(77) Ding, B.-W.; Naumov, P.; Liu, Y.-J. *Journal of Chemical Theory and Computation* **2015**, *11*, 591–599.

(78) Ding, B.-W.; Liu, Y.-J. *Journal of the American Chemical Society* **2017**, *139*, 1106–1119.

(79) Pinto da Silva, L.; Pereira, R. F.; Magalhaes, C. M., et al. *The Journal of Physical Chemistry B* **2017**, *121*, 7862–7871.

(80) Tang, Y.-Q.; Liu, Y.-J. *Journal of Photochemistry and Photobiology A: Chemistry* **2019**, *380*, 111870.

(81) Liu, F.; Liu, Y.; De Vico, L., et al. *Journal of the American Chemical Society* **2009**, *131*, 6181–6188.

(82) Lloyd, J. E. In *Encyclopedia of entomology*; Springer: 2008, pp 1429–1452.

(83) Ghiradella, H.; Schmidt, J. T. In *Encyclopedia of Entomology*; Springer: 2008, pp 1452–1463.

(84) Aprille, J. R.; Lagace, C. J.; Modica-Napolitano, J., et al. *Integrative and comparative biology* **2004**, *44*, 213–219.

(85) Tsai, Y.-L.; Li, C.-W.; Hong, T.-M., et al. *Physical Review Letters* **2014**, *113*, 258103.

(86) Keller, G.-A.; Gould, S.; Deluca, M., et al. *Proceedings of the National Academy of Sciences* **1987**, *84*, 3264–3268.

(87) Marques, S. M.; Esteves da Silva, J. C. *IUBMB life* **2009**, *61*, 6–17.

(88) Conti, E.; Franks, N. P.; Brick, P. *Structure* **1996**, *4*, 287–298.

(89) Gould, S. J.; Keller, G.-A.; Hosken, N., et al. *The Journal of cell biology* **1989**, *108*, 1657–1664.

(90) Tsarkova, A. S. *Frontiers in Ecology and Evolution* **2021**, *9*, 667829.

(91) Vongsangnak, W.; Chumnanpuen, P.; Sriboonlert, A. *PeerJ* **2016**, *4*, e2534.

(92) Zhang, R.; He, J.; Dong, Z., et al. *Scientific reports* **2020**, *10*, 15882.

(93) Antonenkov, V. D.; Hiltunen, J. K. *Biochimica et Biophysica Acta (BBA)-Molecular Basis of Disease* **2012**, *1822*, 1374–1386.

(94) Kanie, S.; Nishikawa, T.; Ojika, M., et al. *Scientific Reports* **2016**, *6*, 24794.

(95) Kanie, S.; Nakai, R.; Ojika, M., et al. *Bioorganic Chemistry* **2018**, *80*, 223–229.

(96) Niwa, K.; Nakamura, M.; Ohmiya, Y. *FEBS letters* **2006**, *580*, 5283–5287.

(97) Maeda, J.; Kato, D.-i.; Okuda, M., et al. *Biochimica et Biophysica Acta (BBA)-General Subjects* **2017**, *1861*, 2112–2118.

(98) Strehler, B. L.; Press, G. D.; Raychaudhuri, A. *Annals of the Entomological Society of America* **1967**, *60*, 81–91.

(99) Zannoni, V.; Lomtevas, N.; Goldfinger, S. *Biochimica et Biophysica Acta (BBA)-General Subjects* **1969**, *177*, 94–105.

(100) Moran, G. R. *Archives of Biochemistry and Biophysics* **2005**, *433*, 117–128.

(101) Taylor, A. M.; Kammath, V.; Bleakley, A. *Medical Hypotheses* **2016**, *91*, 77–80.

(102) Merilainen, G.; Poikela, V.; Kursula, P., et al. *Biochemistry* **2009**, *48*, 11011–11025.

(103) Seedorf, U.; Brysch, P.; Engel, T., et al. *Journal of Biological Chemistry* **1994**, *269*, 21277–21283.

(104) Fallon, T. R.; Li, F.-S.; Vicent, M. A., et al. *Biochemistry* **2016**, *55*, 3341–3344.

(105) Ando, Y.; Niwa, K.; Yamada, N., et al. *Nature Photonics* **2008**, *2*, 44–47.

(106) Rhodes, W.; McElroy, W. D. *Journal of biological chemistry* **1958**, *233*, 1528–1537.

(107) Dukhovich, A.; Sillero, A.; Sillero, M. A. G. *FEBS letters* **1996**, *395*, 188–190.

(108) Fraga, H.; Esteves da Silva, J. C.; Fontes, R. *ChemBioChem* **2004**, *5*, 110–115.

(109) Branchini, B. R.; Magyar, R. A.; Murtiashaw, M. H., et al. *Biochemistry* **2001**, *40*, 2410–2418.

(110) Viviani, V. R. *Cellular and Molecular Life Sciences CMLS* **2002**, *59*, 1833–1850.

(111) Seliger, H.; McElroy, W. *Science* **1962**, *138*, 683–685.

(112) Branchini, B. R.; Behney, C. E.; Southworth, T. L., et al. *Journal of the American Chemical Society* **2015**, *137*, 7592–7595.

(113) DeLuca, M.; McElroy, W. D. *Biochemistry* **1974**, *13*, 921–925.

(114) Shimomura, O.; Goto, T.; Johnson, F. H. *Proceedings of the National Academy of Sciences* **1977**, *74*, 2799–2802.

(115) Orlova, G.; Goddard, J. D.; Brovko, L. Y. *Journal of the American Chemical Society* **2003**, *125*, 6962–6971.

(116) Naumov, P.; Ozawa, Y.; Ohkubo, K., et al. *Journal of the American Chemical Society* **2009**, *131*, 11590–11605.

(117) Maltsev, O. V.; Nath, N. K.; Naumov, P., et al. *Angewandte Chemie* **2014**, *126*, 866–869.

(118) Song, C.-i.; Rhee, Y. M. *Journal of the American Chemical Society* **2011**, *133*, 12040–12049.

(119) Al-Handawi, M. B.; Polavaram, S.; Kurlevskaya, A., et al. *Chemical reviews* **2022**, *122*, 13207–13234.

(120) Ghose, A.; Rebarz, M.; Maltsev, O. V., et al. *The Journal of Physical Chemistry B* **2015**, *119*, 2638–2649.

(121) Fraga, H.; Fernandes, D.; Novotny, J., et al. *ChemBioChem* **2006**, *7*, 929–935.

(122) Branchini, B. R.; Rosenberg, J. C.; Fontaine, D. M., et al. *Journal of the American Chemical Society* **2011**, *133*, 11088–11091.

(123) Sundlov, J. A.; Fontaine, D. M.; Southworth, T. L., et al. *Biochemistry* **2012**, *51*, 6493–6495.

(124) Nakatsu, T.; Ichiyama, S.; Hiratake, J., et al. *Nature* **2006**, *440*, 372–376.

(125) Gomi, K.; Kajiyama, N. *Journal of Biological Chemistry* **2001**, *276*, 36508–36513.

(126) Gomi, K.; Hirokawa, K.; Kajiyama, N. *Gene* **2002**, *294*, 157–166.

(127) Day, J. C.; Tisi, L. C.; Bailey, M. J. *Luminescence: The journal of biological and chemical luminescence* **2004**, *19*, 8–20.

(128) Emamzadeh, R.; Hosseinkhani, S.; Hemati, R., et al. *Enzyme and Microbial Technology* **2010**, *47*, 159–165.

(129) Inouye, S. *Cellular and molecular life sciences* **2010**, *67*, 387–404.

(130) Hemmati, R.; Hosseinkhani, S.; Sajedi, R. H., et al. *Photochemistry and Photobiology* **2015**, *91*, 828–836.

(131) Carrasco-Lopez, C.; Lui, N. M.; Schramm, S., et al. *Nature Reviews Chemistry* **2021**, *5*, 4–20.

(132) Seliger, H. H.; McElroy, W. D. *Archives of biochemistry and biophysics* **1960**, *88*, 136–141.

(133) Suzuki, N.; Sato, M.; Nishikawa, K., et al. *Tetrahedron Letters* **1969**, *10*, 4683–4684.

(134) Suzuki, N.; Goto, T. *Tetrahedron Letters* **1971**, *12*, 2021–2024.

(135) Suzuki, N.; Goto, T. *Agricultural and Biological Chemistry* **1972**, *36*, 2213–2221.

(136) Hopkins, T. A.; Seliger, H. H.; White, E. H., et al. *Journal of the American Chemical Society* **1967**, *89*, 7148–7150.

(137) Kajiyama, N.; Nakano, E. *Protein Engineering, Design and Selection* **1991**, *4*, 691–693.

(138) Branchini, B. R.; Murtiashaw, M. H.; Magyar, R. A., et al. *Journal of the American Chemical Society* **2002**, *124*, 2112–2113.

(139) McCapra, F. *Bioluminescence and Chemiluminescence: Fundamental and applied aspects* **1994**.

(140) McCapra, F. *Bioluminescence and Chemiluminescence* **1996**.

(141) Franks, N.; Jenkins, A.; Conti, E., et al. *Biophysical journal* **1998**, *75*, 2205–2211.

(142) Hirano, T.; Hasumi, Y.; Ohtsuka, K., et al. *Journal of the American Chemical Society* **2009**, *131*, 2385–2396.

(143) Naumov, P.; Kochunnoonny, M. *Journal of the American Chemical Society* **2010**, *132*, 11566–11579.

(144) Auld, D. S.; Lovell, S.; Thorne, N., et al. *Proceedings of the National Academy of Sciences* **2010**, *107*, 4878–4883.

(145) Thorne, N.; Inglese, J.; Auld, D. S. *Chemistry & biology* **2010**, *17*, 646–657.

(146) Cruz, P. G.; Auld, D. S.; Schultz, P. J., et al. *Chemistry & biology* **2011**, *18*, 1442–1452.

(147) Kheirabadi, M.; Sharafian, Z.; Naderi-Manesh, H., et al. *Biochimica et Biophysica Acta (BBA)-Proteins and Proteomics* **2013**, *1834*, 2729–2735.

(148) Wu, W.; Su, J.; Tang, C., et al. *Analytical Chemistry* **2017**, *89*, 4808–4816.

(149) Branchini, B. R.; Southworth, T. L.; Fontaine, D. M., et al. *Photochemistry and Photobiology* **2017**, *93*, 479–485.

(150) Zhang, H.; Su, J.; Lin, Y., et al. *Analytical Chemistry* **2017**, *89*, 6099–6105.

(151) Carrasco-López, C.; Ferreira, J. C.; Lui, N. M., et al. *Life science alliance* **2018**, *1*.

(152) Branchini, B. R.; Fontaine, D. M.; Southworth, T. L., et al. *Biochemistry* **2019**, *58*, 4293–4303.

(153) Stowe, C. L.; Burley, T. A.; Allan, H., et al. *Elife* **2019**, *8*, e45801.

(154) Yang, T.; Goddard, J. D. *The Journal of Physical Chemistry A* **2007**, *111*, 4489–4497.

(155) Nakatani, N.; Hasegawa, J.-y.; Nakatsuji, H. *Journal of the American Chemical Society* **2007**, *129*, 8756–8765.

(156) Branchini, B. R.; Southworth, T. L.; Murtiashaw, M. H., et al. *Biochemistry* **2004**, *43*, 7255–7262.

(157) Ando, Y.; Akiyama, H. *Japanese journal of applied physics* **2010**, *49*, 117002.

(158) Hiyama, M.; Akiyama, H.; Yamada, K., et al. *Photochemistry and Photobiology* **2013**, *89*, 571–578.

(159) Wang, Y.; Akiyama, H.; Terakado, K., et al. *Scientific reports* **2013**, *3*, 2490.

(160) Wang, Y.; Hayamizu, Y.; Akiyama, H. *The Journal of Physical Chemistry B* **2014**, *118*, 2070–2076.

(161) Mochizuki, T.; Wang, Y.; Hiyama, M., et al. *Applied Physics Letters* **2014**, *104*.

(162) Ugarova, N.; Maloshenok, L.; Uporov, I., et al. *Biochemistry (Moscow)* **2005**, *70*, 1262–1267.

(163) Gandelman, O.; Brovko, L. Y.; Ugarova, N., et al. *Journal of Photochemistry and Photobiology B: Biology* **1993**, *19*, 187–191.

(164) Gandelman, O.; Brovko, L. Y.; Chikishev, A. Y., et al. *Journal of Photochemistry and Photobiology B: Biology* **1994**, *22*, 203–209.

(165) Sandalova, T.; Ugarova, N. *Biochemistry. Biokhimiia* **1999**, *64*, 962–967.

(166) Ugarova, N.; Brovko, L. Y. *Russian chemical bulletin* **2001**, *50*, 1752–1761.

(167) Ugarova, N.; Brovko, L. Y. *Luminescence: the journal of biological and chemical luminescence* **2002**, *17*, 321–330.

(168) Hirano, T.; Nagai, H.; Matsuhashi, T., et al. *Photochemical & Photobiological Sciences* **2012**, *11*, 1281–1284.

(169) Evans, M. S.; Chaurette, J. P.; Adams Jr, S. T., et al. *Nature methods* **2014**, *11*, 393–395.

(170) Conley, N. R.; Dragulescu-Andrasi, A.; Rao, J., et al. *Angewandte Chemie (International ed. in English)* **2012**, *51*, 3350.

(171) Branching, B. R.; Hayward, M. M.; Bamford, S., et al. *Photochemistry and Photobiology* **1989**, *49*, 689–695.

(172) Branchini, B. R. **2000**.

(173) Whitehead, T. P.; Kricka, L. J.; Carter, T., et al. *Clinical chemistry* **1979**, *25*, 1531–1546.

(174) Wannlund, J.; Azari, J.; Levine, L., et al. *Biochemical and biophysical research communications* **1980**, *96*, 440–446.

(175) Wood, K.; De Wet, J.; Dewji, N., et al. *Biochemical and biophysical research communications* **1984**, *124*, 592–596.

(176) Ow, D. W.; Wood, K. V.; DeLuca, M., et al. *Science* **1986**, *234*, 856–859.

(177) De Wet, J. R.; Wood, K. V.; DeLuca, M., et al. *Molecular and cellular biology* **1987**, *7*, 725–737.

(178) Roda, A., *Chemiluminescence and bioluminescence: past, present and future*; Royal Society of Chemistry: 2011.

(179) Wibom, R.; Hultman, E. *American Journal of Physiology-Endocrinology and Metabolism* **1990**, *259*, E204–E209.

(180) Lanza, I. R.; Nair, K. S. *Methods in enzymology* **2009**, *457*, 349–372.

(181) Short, K. R.; Bigelow, M. L.; Kahl, J., et al. *Proceedings of the National Academy of Sciences* **2005**, *102*, 5618–5623.

(182) Huang, Y.-S.; Chen, Y.-C.; Chen, M.-L., et al. *American Journal of Infection Control* **2015**, *43*, 882–886.

(183) Chan, M.-C.; Lin, T.-Y.; Chiu, Y.-H., et al. *Journal of Hospital Infection* **2015**, *90*, 344–346.

(184) Österberg, E.; Hallander, H.; Kallner, A., et al. *European Journal of Clinical Microbiology and Infectious Diseases* **1991**, *10*, 70–73.

(185) Dong, T.; Zhao, X. *Analytical Chemistry* **2015**, *87*, 2410–2418.

(186) Troy, T.; Jekic-McMullen, D.; Sambucetti, L., et al. *Molecular imaging* **2004**, *3*, 15353500200403196.

(187) Awais, M.; Ozawa, T. *Molecular Biosystems* **2011**, *7*, 1376–1387.

(188) Azad, T.; Tashakor, A.; Hosseinkhani, S. *Analytical and bioanalytical chemistry* **2014**, *406*, 5541–5560.

(189) Deluca, M. *Advances in enzymology and related areas of molecular biology* **1976**, *44*, 37–68.

(190) Hida, N.; Awais, M.; Takeuchi, M., et al. *PLoS One* **2009**, *4*, e5868.

(191) Ozawa, T.; Kaihara, A.; Sato, M., et al. *Analytical Chemistry* **2001**, *73*, 2516–2521.

(192) Paulmurugan, R.; Umezawa, Y.; Gambhir, S. *Proceedings of the National Academy of Sciences* **2002**, *99*, 15608–15613.

(193) Luker, K. E.; Smith, M. C.; Luker, G. D., et al. *Proceedings of the National Academy of Sciences* **2004**, *101*, 12288–12293.

(194) Baird, G. S.; Zacharias, D. A.; Tsien, R. Y. *Proceedings of the National Academy of Sciences* **1999**, *96*, 11241–11246.

(195) Wigdal, S. S.; Anderson, J. L.; Vidugiris, G. J., et al. *Current chemical genomics* **2008**, *2*, 16.

(196) Meighen, E. A.; Dunlap, P. V. *Advances in microbial physiology* **1993**, *34*, 1–67.

(197) Kanno, A.; Yamanaka, Y.; Hirano, H., et al. *Angewandte Chemie International Edition* **2007**, *46*, 7595–7599.

(198) Kanno, A.; Umezawa, Y.; Ozawa, T. In *Bioluminescence: Methods and Protocols*; Springer: 2009, pp 105–114.

(199) Ray, P.; Gambhir, S. S. In *Reporter Genes: A Practical Guide*; Springer: 2007, pp 131–144.

(200) Brogan, J.; Li, F.; Li, W., et al. *Radiation research* **2012**, *177*, 508–513.

(201) Tromberg, B. J.; Shah, N.; Lanning, R., et al. *Neoplasia* **2000**, *2*, 26–40.

(202) Contag, C. H.; Bachmann, M. H. *Annual review of biomedical engineering* **2002**, *4*, 235–260.

(203) De Broglie, L. Recherches sur la théorie des quanta, Ph.D. Thesis, Migration-université en cours d'affectation, 1924.

(204) Schrödinger, E. *Annalen der Physik* **1926**, *79*, 734–756.

(205) Schrödinger, E. *Annalen der Physik* **1926**, *386*, 109–139.

(206) Schrödinger, E. *Naturwissenschaften* **1926**, *14*, 664–666.

(207) Schrödinger, E. *Physical Review* **1926**, *28*, 1049.

(208) Born, M. *Zeitschrift f r Physik* **1926**, *37*, 863–867.

(209) Levine, I. N.; Busch, D. H.; Shull, H., *Quantum chemistry*; Pearson Prentice Hall Upper Saddle River, NJ: 2009; Vol. 6.

(210) Szabo, A.; Ostlund, N. S., *Modern quantum chemistry: introduction to advanced electronic structure theory*; Courier Corporation: 1996.

(211) Cramer, C. J., *Essentials of computational chemistry: theories and models*; John Wiley & Sons: 2013.

(212) Born, M.; Oppenheimer, R. *Annalen der Physik* **1927**, *84*, 457–484.

(213) Jensen, F., *Introduction to computational chemistry*; John wiley & sons: 2017.

(214) Agostini, F.; Curchod, B. F. *Wiley interdisciplinary reviews: computational molecular science* **2019**, *9*, e1417.

(215) Persico, M.; Granucci, G., *Photochemistry*; Springer: 2018.

(216) Pauli, W. *Naturwiss* **1924**, *12*, 741.

(217) Pauli, W.; Belinfante, F. J. *Physica* **1940**, *7*, 177–192.

(218) Pauli, W. *Physical Review* **1940**, *58*, 716.

(219) Roothaan, C. C. J. *Reviews of modern physics* **1951**, *23*, 69.

(220) Hall, G. G. *Proceedings of the Royal Society of London. Series A. Mathematical and Physical Sciences* **1951**, *205*, 541–552.

(221) Pople, J. A.; Nesbet, R. K. *The Journal of Chemical Physics* **1954**, *22*, 571–572.

(222) Wilkinson, J. *Press, London* **1965**.

(223) Juan, A.; Beltrán, J. *Castellón de la plata, Universidad Jaime I* **2000**.

(224) Hohenberg, P.; Kohn, W. *Physical Review* **1964**, *136*, B864.

(225) Levy, M. *Proceedings of the National Academy of Sciences* **1979**, *76*, 6062–6065.

(226) Kohn, W.; Sham, L. J. *Physical Review* **1965**, *140*, A1133.

(227) Vosko, S.; Wilk, L.; Nusair, M. *Canadian Journal of Physics* **1980**, *58*, 1200.

(228) Lee, C.; Yang, W.; Parr, R. G. *Physical Review B* **1988**, *37*, 785.

(229) Miehlich, B.; Savin, A.; Stoll, H., et al. *Chemical Physics Letters* **1989**, *157*, 200–206.

(230) Harris, J. *Physical Review A* **1984**, *29*, 1648.

(231) Becke, A. D. *The Journal of Chemical Physics* **1993**, *98*, 5648–5652.

(232) Stephens, P. J.; Devlin, F. J.; Chabalowski, C. F., et al. *The Journal of Physical Chemistry* **1994**, *98*, 11623–11627.

(233) Yanai, T.; Tew, D. P.; Handy, N. C. *Chemical Physics Letters* **2004**, *393*, 51–57.

(234) Talman, J. D.; Shadwick, W. F. *Physical Review A* **1976**, *14*, 36.

(235) Bonetti, A. F.; Engel, E.; Schmid, R., et al. *Physical Review Letters* **2001**, *86*, 2241.

(236) Bartlett, R. J.; Lotrich, V. F.; Schweigert, I. V. *The Journal of Chemical Physics* **2005**, *123*, 062205.

(237) Møller, C.; Plesset, M. S. *Physical Review* **1934**, *46*, 618.

(238) Runge, E.; Gross, E. K. *Physical Review Letters* **1984**, *52*, 997.

(239) Gross, E. K. U.; Dobson, J. F.; Petersilka, M. In *Density Functional Theory II: Relativistic and Time Dependent Extensions*, Nalewajski, R. F., Ed.; Springer Berlin Heidelberg: Berlin, Heidelberg, 1996, pp 81–172.

(240) González, L.; Lindh, R., *Quantum chemistry and dynamics of excited states: methods and applications*; John Wiley & Sons: 2020.

(241) Marques, M. A.; Maitra, N. T.; Nogueira, F. M., et al., *Fundamentals of time-dependent density functional theory*; Springer: 2012; Vol. 837.

(242) Casida, M. E. In *Recent Advances In Density Functional Methods: (Part I)*; World Scientific: 1995, pp 155–192.

(243) Hirata, S.; Head-Gordon, M. *Chemical Physics Letters* **1999**, *314*, 291–299.

(244) Cauchy, A. *Comptes Rendus de l'Académie des Sciences* **1847**, *25*, 536–538.

(245) Fletcher, R.; Reeves, C. M. *The computer journal* **1964**, *7*, 149–154.

(246) Császár, P.; Pulay, P. *Journal of Molecular Structure* **1984**, *114*, 31–34.

(247) Li, X.; Frisch, M. J. *Journal of Chemical Theory and Computation* **2006**, *2*, 835–839.

(248) Schlegel, H. B. *Journal of computational chemistry* **1982**, *3*, 214–218.

(249) Plasser, F. *The Journal of Chemical Physics* **2020**, *152*.

(250) Mai, S.; Plasser, F.; Dorn, J., et al. *Coordination Chemistry Reviews* **2018**, *361*, 74–97.

(251) Voityuk, A. A. *The Journal of Chemical Physics* **2014**, *140*.

(252) Grotendorst, J.; Marx, D.; Muramatsu, A., *Quantum Simulations of Complex Many-Body Systems: From Theory to Algorithms - Lecture Notes*; NIC-Directors: 2002; Vol. 10.

(253) Verlet, L. *Physical Review* **1967**, *159*, 98.

(254) Helgaker, T.; Uggerud, E.; Jensen, H. J. A. *Chemical Physics Letters* **1990**, *173*, 145–150.

(255) Vázquez, S. A.; Otero, X. L.; Martinez-Nunez, E. *Molecules* **2018**, *23*, 3156.

(256) Martínez-Núñez, E.; Barnes, G. L.; Glowacki, D. R., et al. *Journal of computational chemistry* **2021**, *42*, 2036–2048.

(257) Diamantis, P.; Tavernelli, I.; Rothlisberger, U. *Journal of Chemical Theory and Computation* **2020**, *16*, 6690–6701.

(258) Cárdenas, G.; Marquetand, P.; Mai, S., et al. *Catalysts* **2021**, *11*, 493.

(259) Monticelli, L.; Tieleman, D. P. In *Biomolecular Simulation Methods and Protocols*; Springer: 2013, pp 197–213.

(260) Case, D.; Aktulga, H.; Belfon, K., et al. **2020**.

(261) Wang, J.; Wolf, R. M.; Caldwell, J. W., et al. *Journal of computational chemistry* **2004**, *25*, 1157–1174.

(262) Tian, C.; Kasavajhala, K.; Belfon, K. A., et al. *Journal of Chemical Theory and Computation* **2019**, *16*, 528–552.

(263) Jorgensen, W.; Chandrasekhar, J.; Madura, J., et al. *The Journal of Chemical Physics* **1983**, *79*, 926–935.

(264) Mateo-delaFuente, H.; Nogueira, J. J. *Photochemistry and Photobiology Sciences* **2024**, 1–15.

(265) Seminario, J. M. *Int. J. Quantum Chem.* **1996**, *60*, 1271–1277.

(266) Chandler, D., *Introduction to Modern Statistical Mechanics*; Oxford University Press, Inc.: 1987, p 288.

(267) Sun, X.; Lin, T.; Gezelter, J. D. *The Journal of Chemical Physics* **2008**, *128*, 1352.

(268) Davidchack, R. L.; Handel, R.; Tretyakov, M. V. *The Journal of Chemical Physics* **2009**, *130*, 234101.

(269) Frenkel, D.; Smit, B., *Understanding molecular simulation : from algorithms to applications*; Academic Press: 1996; Vol. 50.

(270) Åqvist, J.; Wennerström, P.; Nervall, M., et al. *Chemical Physics Letters* **2004**, *384*, 288–294.

(271) Berendsen, H. J. C.; Postma, J. P. M.; van Gunsteren, W. F., et al. *The Journal of Chemical Physics* **1984**, *81*, 3684.

(272) Darden, T.; York, D.; Pedersen, L. *The Journal of Chemical Physics* **1993**, *98*, 10089–10092.

(273) Frenkel, D.; Smit, B., *Understanding molecular simulation : from algorithms to applications*; Academic Pressf: 1996; Vol. 50.

(274) Ryckaert, J.-P.; Ciccotti, G.; Berendsen, H. J. *Journal of computational physics* **1977**, *23*, 327–341.

(275) Andersen, H. C. *J. Comput. Phys.* **1983**, *52*, 24–34.

(276) Miyamoto, S.; Kollman, P. *Journal of computational chemistry* **1992**, *13*, cited By 4794, 952–962.

(277) Metropolis, N.; Ulam, S. *Journal of the American Statistical Association* **1949**, *44*, 335–341.

(278) Boulard, B.; Kieffer, J.; Phifer, C., et al. *Journal of non-crystalline solids* **1992**, *140*, 350–358.

(279) Bornhauser, P.; Bougeard, D. *The Journal of Physical Chemistry B* **2001**, *105*, 36–41.

(280) Amadei, A.; Linssen, A. B.; Berendsen, H. J. *Proteins: Structure, Function, and Bioinformatics* **1993**, *17*, 412–425.

(281) Hayward, S.; Kitao, A.; Gō, N. *Protein Science* **1994**, *3*, 936–943.

(282) Senn, H. M.; Thiel, W. *Angewandte Chemie International Edition* **2009**, *48*, 1198–1229.

(283) Warshel, A.; Karplus, M. *Journal of the American Chemical Society* **1972**, *94*, 5612–5625.

(284) Warshel, A.; Levitt, M. *Journal of Molecular Biology* **1976**, *103*, 227–249.

(285) Field, M. J.; Bash, P. A.; Karplus, M. *Journal of computational chemistry* **1990**, *11*, 700–733.

(286) Liu, M.; Wang, Y.; Chen, Y., et al. *Israel journal of chemistry* **2014**, *54*, 1250–1263.

(287) Srinivasan, J.; Miller, J.; Kollman, P. A., et al. *Journal of Biomolecular Structure and Dynamics* **1998**, *16*, 671–682.

(288) Srinivasan, J.; Cheatham, T. E.; Cieplak, P., et al. *Journal of the American Chemical Society* **1998**, *120*, 9401–9409.

(289) Kollman, P. A.; Massova, I.; Reyes, C., et al. *Accounts of chemical research* **2000**, *33*, 889–897.

(290) Wang, E.; Sun, H.; Wang, J., et al. *Chemical reviews* **2019**, *119*, 9478–9508.

(291) Lee, M. S.; Salsbury Jr, F. R.; Brooks III, C. L. *The Journal of Chemical Physics* **2002**, *116*, 10606–10614.

(292) Im, W.; Lee, M. S.; Brooks III, C. L. *Journal of computational chemistry* **2003**, *24*, 1691–1702.

(293) Still, W. C.; Tempczyk, A.; Hawley, R. C., et al. *Journal of the American Chemical Society* **1990**, *112*, 6127–6129.

(294) Sigalov, G.; Scheffel, P.; Onufriev, A. *The Journal of Chemical Physics* **2005**, *122*.

(295) Onufriev, A.; Bashford, D.; Case, D. A. *Proteins: Structure, Function, and Bioinformatics* **2004**, *55*, 383–394.

(296) Onufriev, A.; Bashford, D.; Case, D. A. *The Journal of Physical Chemistry B* **2000**, *104*, 3712–3720.

(297) Sanner, M. F.; Olson, A. J.; Spehner, J.-C. *Biopolymers* **1996**, *38*, 305–320.

(298) Morokuma, K. *The Journal of Chemical Physics* **1971**, *55*, 1236–1244.

(299) Phipps, M. J.; Fox, T.; Tautermann, C. S., et al. *Chemical society reviews* **2015**, *44*, 3177–3211.

(300) Zhao, L.; von Hopffgarten, M.; Andrada, D. M., et al. *Wiley Interdisciplinary Reviews: Computational Molecular Science* **2018**, *8*, e1345.

(301) Jeziorski, B.; Moszynski, R.; Szalewicz, K. *Chemical Reviews* **1994**, *94*, 1887–1930.

(302) Mandado, M.; Hermida-Ramón, J. M. *Journal of Chemical Theory and Computation* **2011**, *7*, 633–641.

(303) Ramos-Berdullas, N.; Pérez-Juste, I.; Van Alsenoy, C., et al. *Physical Chemistry Chemical Physics* **2015**, *17*, 575–587.

(304) Boys, S. F.; Bernardi, F. *Molecular physics* **1970**, *19*, 553–566.

(305) Simon, S.; Duran, M.; Dannenberg, J. *The Journal of Chemical Physics* **1996**, *105*, 11024–11031.

(306) Van Duijneveldt, F. B.; van Duijneveldt-van de Rijdt, J. G.; van Lenthe, J. H. *Chemical Reviews* **1994**, *94*, 1873–1885.

(307) Johansson, A.; Kollman, P.; Rothenberg, S. *Theoretica chimica acta* **1973**, *29*, 167–172.

(308) Daudey, J.; Claverie, P.; Malrieu, J. *International Journal of Quantum Chemistry* **1974**, *8*, 1–15.

(309) Kaplan, I. G. et al., *Intermolecular interactions: physical picture, computational methods and model potentials*; Wiley Online Library: 2006; Vol. 367.

(310) Cárdenas, G.; Pérez-Barcia, Á.; Mandado, M., et al. *Physical Chemistry Chemical Physics* **2021**, *23*, 20533–20540.

(311) García-Iriepa, C.; Zemmouche, M.; Ponce-Vargas, M., et al. *Physical Chemistry Chemical Physics* **2019**, *21*, 4613–4623.

(312) Branchini, B. R.; Magyar, R. A.; Murtiashaw, M. H., et al. *Biochemistry* **1999**, *38*, 13223–13230.

(313) Krishnan, R.; Binkley, J. S.; Seeger, R., et al. *The Journal of Chemical Physics* **1980**, *72*, 650–654.

(314) Francl, M. M.; Pietro, W. J.; Hehre, W. J., et al. *The Journal of Chemical Physics* **1982**, *77*, 3654–3665.

(315) McLean, A.; Chandler, G. *The Journal of Chemical Physics* **1980**, *72*, 5639–5648.

(316) Lee, C.; Yang, W.; Parr, R. *Physical Review B: Condensed Matter and Materials Physics* **1988**, *37*, 785.

(317) Zhao, Y.; Truhlar, D. G. *Theoretical chemistry accounts* **2008**, *120*, 215–241.

(318) Adamo, C.; Barone, V. *The Journal of Chemical Physics* **1999**, *110*, 6158–6170.

(319) Chai, J.-D.; Head-Gordon, M. *Physical Chemistry Chemical Physics* **2008**, *10*, 6615–6620.

(320) Frisch, M. J.; Trucks, G. W.; Schlegel, H. B., et al. Gaussian 16 Revision C.01, Gaussian Inc. Wallingford CT, 2016.

(321) Da Silva, L. P.; Esteves da Silva, J. C. *Journal of Chemical Theory and Computation* **2011**, *7*, 809–817.

(322) Chen, S.-F.; Liu, Y.-J.; Navizet, I., et al. *Journal of Chemical Theory and Computation* **2011**, *7*, 798–803.

(323) Cheng, Y.-Y.; Liu, Y.-J. *Journal of Chemical Theory and Computation* **2015**, *11*, 5360–5370.

(324) Schrödinger, L.; DeLano, W. PyMOL, version 2.4.0, 2020.

(325) Singh, U. C.; Kollman, P. A. *Journal of computational chemistry* **1984**, *5*, 129–145.

(326) Berraud-Pache, R.; Navizet, I. *Physical Chemistry Chemical Physics* **2016**, *18*, 27460–27467.

(327) Garcia-Iriepa, C.; Gosset, P.; Berraud-Pache, R., et al. *Journal of Chemical Theory and Computation* **2018**, *14*, 2117–2126.

(328) Cerezo, J.; García-Iriepa, C.; Santoro, F., et al. *Physical Chemistry Chemical Physics* **2023**, *25*, 5007–5020.

(329) Cárdenas, G.; Lucia-Tamudo, J.; Mateo-delaFuente, H., et al. *Journal of computational chemistry* **2023**, *44*, 516–533.

(330) Weingart, O.; Nenov, A.; Altoè, P., et al. *Journal of molecular modeling* **2018**, *24*, 271.

(331) Avagliano, D.; Bonfanti, M.; Nenov, A., et al. *Journal of computational chemistry* **2022**, *43*, 1641–1655.

(332) Clark, T.; Chandrasekhar, J.; Spitznagel, G. W., et al. *Journal of computational chemistry* **1983**, *4*, 294–301.

(333) Spitznagel, G. W.; Clark, T.; von Ragué Schleyer, P., et al. *Journal of computational chemistry* **1987**, *8*, 1109–1116.

(334) Avagliano, D.; Bonfanti, M.; Garavelli, M., et al. *Journal of Chemical Theory and Computation* **2021**, *17*, 4639–4647.

(335) Miller III, B. R.; McGee Jr, T. D.; Swails, J. M., et al. *Journal of Chemical Theory and Computation* **2012**, *8*, 3314–3321.

(336) Palmisano, V. F.; Gómez-Rodellar, C.; Pollak, H., et al. *Physical Chemistry Chemical Physics* **2021**, *23*, 3552–3564.

(337) Hartigan, J. A., *Clustering algorithms*; John Wiley & Sons, Inc.: 1975.

(338) Peach, M. J.; Benfield, P.; Helgaker, T., et al. *The Journal of Chemical Physics* **2008**, *128*.

(339) Song, C.-I.; Rhee, Y. M. *International Journal of Quantum Chemistry* **2011**, *111*, 4091–4105.

(340) Mateo-delaFuente, H.; Avagliano, D.; Garavelli, M., et al. *Physical Chemistry Chemical Physics* **2023**, *25*, 16507–16519.

(341) Frisch, M.; Clemente, F., et al. *Scalmani, V. Barone, B. Mennucci, GA Petersson, H. Nakatsuji, M. Caricato, X. Li, HP Hratchian, AF Izmaylov, J. Bloino, G. Zhe* **2009**, 20–44.

(342) Francl, M. M.; Pietro, W. J.; Hehre, W. J., et al. *The Journal of Chemical Physics* **1982**, *77*, 3654–3665.

(343) Krishnan, R.; Binkley, J. S.; Seeger, R., et al. *The Journal of Chemical Physics* **1980**, *72*, 650–654.

(344) McLean, A. D.; Chandler, G. S. *The Journal of Chemical Physics* **1980**, *72*, 5639–5648.

(345) Betz, R. M.; Walker, R. C. *Journal of computational chemistry* **2015**, *36*, 79–87.

(346) Neese, F. et al. *Rev.: Comput. Mol. Sci* **2012**, *2*, 73–78.

(347) Neese, F.; Wennmohs, F.; Becker, U., et al. *The Journal of Chemical Physics* **2020**, *152*, DOI: 10.1063/5.0004608.

(348) Ditchfield, R.; Hehre, W. J.; Pople, J. A. *The Journal of Chemical Physics* **1971**, *54*, 724–728.

(349) Gordon, M. S.; Binkley, J. S.; Pople, J. A., et al. *Journal of the American Chemical Society* **1982**, *104*, 2797–2803.

(350) Hariharan, P. C.; Pople, J. A. *Theoretica chimica acta* **1973**, *28*, 213–222.

(351) Hehre, W. J.; Ditchfield, R.; Pople, J. A. *The Journal of Chemical Physics* **1972**, *56*, 2257–2261.

(352) Weigend, F. *Physical Chemistry Chemical Physics* **2006**, *8*, 1057–1065.

(353) Neese, F.; Wennmohs, F.; Hansen, A., et al. *The Journal of Chemical Physics* **2009**, *356*, 98–109.

(354) Roe, D. R.; Cheatham III, T. E. *Journal of Chemical Theory and Computation* **2013**, *9*, 3084–3095.

(355) Patel, A. M.; Henley, A.; Parkes, M. A., et al. *Physical Chemistry Chemical Physics* **2020**, *22*, 19022–19032.

(356) De Almeida Barbosa, N. M.; Gosset, P.; Réal, E., et al. *Physical Chemistry Chemical Physics* **2020**, *22*, 21731–21740.

(357) Zemmouche, M.; García-Iriepa, C.; Navizet, I. *Physical Chemistry Chemical Physics* **2020**, *22*, 82–91.

(358) Case, D. A.; Aktulga, H. M.; Belfon, K., et al., *Amber 2022*; University of California, San Francisco: 2022.

(359) Salomon-Ferrer, R.; Gotz, A. W.; Poole, D., et al. *Journal of Chemical Theory and Computation* **2013**, *9*, 3878–3888.

(360) Rustaman, R.; Rahmawan, R. R.; Maksum, I. P. *Turkish Computational and Theoretical Chemistry* **2023**, *7*, 58–69.

(361) Le Grand, S.; Götz, A. W.; Walker, R. C. *Computer Physics Communications* **2013**, *184*, 374–380.

(362) Martín-Fernández, C.; Montero-Campillo, M. M.; Alkorta, I. *The Journal of Physical Chemistry Letters* **2024**, *15*, 4105–4110.

(363) Tagami, A.; Ishibashi, N.; Kato, D.-i., et al. *Chemical Physics Letters* **2009**, *472*, 118–123.

(364) Šrut, A.; Lear, B. J.; Krewald, V. *Chemical science* **2023**, *14*, 9213–9225.

(365) Neese, F. *Wiley Interdisciplinary Reviews: Computational Molecular Science* **2012**, *2*, 73–78.

(366) Mateo-delaFuente, H.; Nogueira, J. J. Structural Analysis of the Oxyluciferin Emitting Species in Aqueous and Luciferase Media, 2025.

(367) Ramos-Berdullas, N.; Pérez-Juste, I.; Van Alsenoy, C., et al. *Physical Chemistry Chemical Physics* **2015**, *17*, 575–587.

(368) Mandado, M. In *Computational Methods for the Analysis of Non-Covalent Interactions*, Martín-Fernández, C., Macgregor, S. A., Eds.; Springer Nature Switzerland: Cham, 2025, pp 71–108.

(369) Clark, T.; Chandrasekhar, J.; Spitznagel, G. W., et al. *Journal of computational chemistry* **1983**, *4*, 294–301.

(370) Spitznagel, G. W.; Clark, T.; Schleyer, P. v. R., et al. *Journal of computational chemistry* **1987**, *8*, 1109–1116.

(371) Frisch, M. J.; Pople, J. A.; Binkley, J. S. *The Journal of Chemical Physics* **1984**, *80*, 3265–3269.

(372) Grimme, S.; Ehrlich, S.; Goerigk, L. *Journal of computational chemistry* **2011**, *32*, 1456–1465.

(373) Grimme, S.; Antony, J.; Ehrlich, S., et al. *The Journal of Chemical Physics* **2010**, *132*.

(374) Waszkowycz, B.; Hillier, I. H.; Gensmantel, N., et al. *Journal of the Chemical Society, Perkin Transactions 2* **1991**, 2025–2032.

(375) Lin, H.; Truhlar, D. G. *Theoretical Chemistry Accounts* **2007**, *117*, 185–199.

(376) Mandado, M.; Ramos-Berdullas, N. In *International Conference on Computational Science and Its Applications*, 2024, pp 375–386.

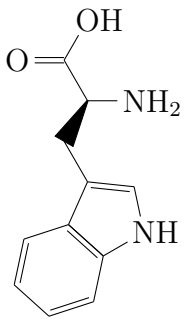
(377) Cárdenas, G.; Pérez-Barcia, Á.; Mandado, M., et al. *Physical Chemistry Chemical Physics* **2021**, *23*, 20533–20540.

(378) Kato, D.-i.; Kubo, T.; Maenaka, M., et al. *Journal of Molecular Catalysis B: Enzymatic* **2013**, *87*, 18–23.

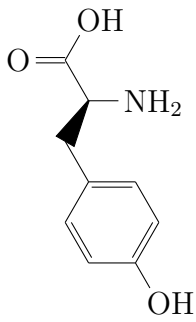
(379) Barbatti, M.; Sen, K. *International Journal of Quantum Chemistry* **2016**, *116*, 762–771.

Appendix A

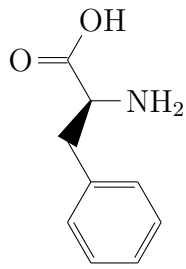
Structure and Abbreviations of Amino Acids



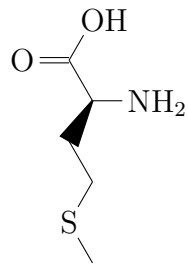
Tryptophan (TRP)



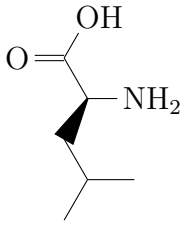
Tyrosine (TYR)



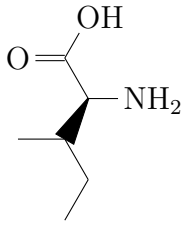
Phenylalanine (PHE)



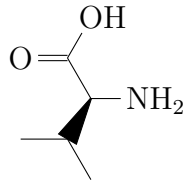
Methionine (MET)



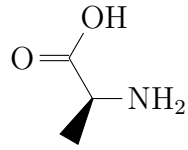
Leucine (LEU)



Isoleucine (ILE)

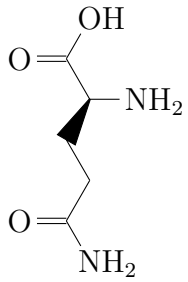


Valine (VAL)

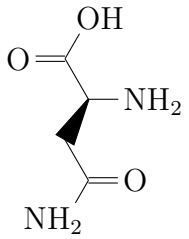


Alanine (ALA)

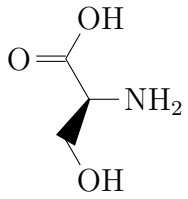
Figure A.1: Amino acids with hydrophobic side chains.



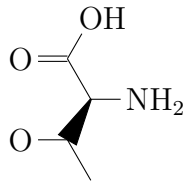
Glutamine (GLN)



Asparagine (ASN)

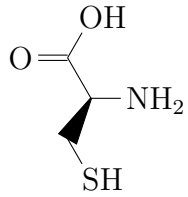


Serine (SER)

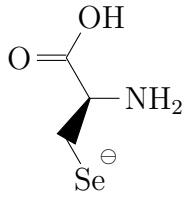


Threonine (THR)

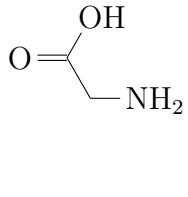
Figure A.2: Amino acids with polar uncharged side chains.



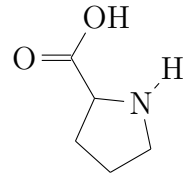
Cysteine (CYS)



Selenocysteine (SEC)

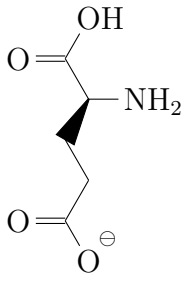


Glycine (GLY)

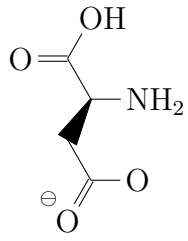


Proline (PRO)

Figure A.3: Special amino acids.

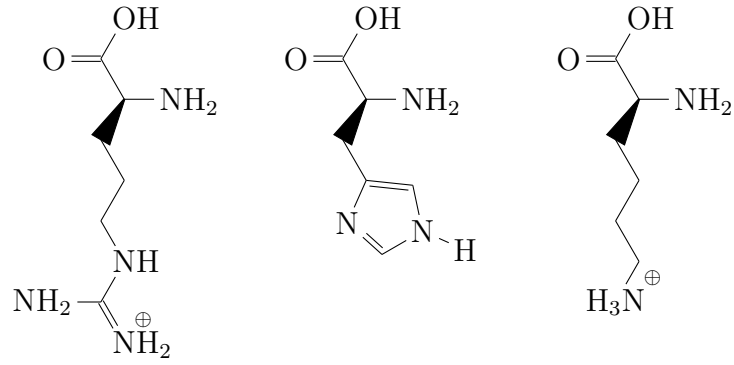


Glutamic Acid (GLU)



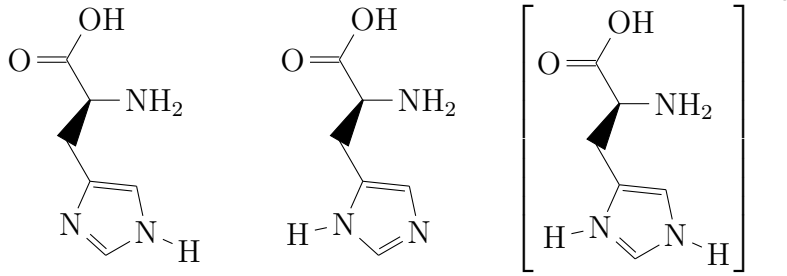
Aspartic Acid (ASP)

Figure A.4: Amino acids with negatively charged side chains.



Arginine (ARG) Histidine (HIS) Lysine (LYS)

Figure A.5: Amino acids with positively charged side chains.



HIE HID HIP

Figure A.6: Possible protonation states of histidine.

Appendix B

Electronic Properties of the Phenolate-Keto Oxyluciferin inside the Wild Type Luciferase

B.1 Oxyluciferin Parameters for the Classical Molecular Dynamics Simulations

B.1.1 Atom Types and Charges

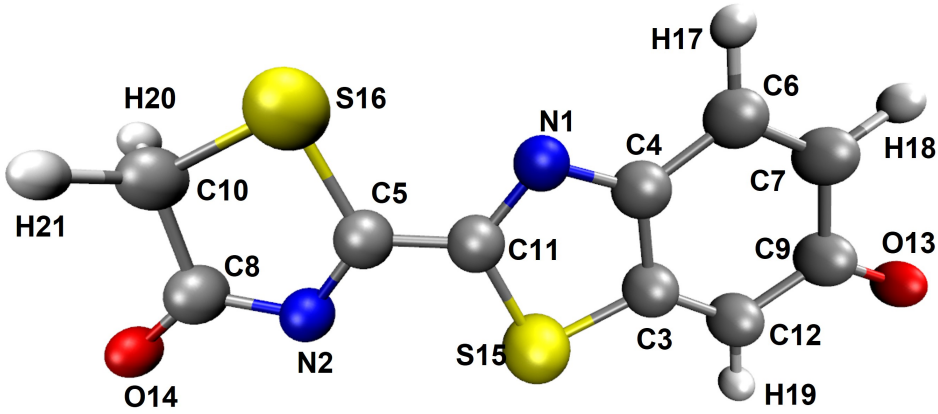


Figure B.1: Structure of the OLU with labelled atoms.

Atom	Atom type	Charge	
		S_0	S_1
N1	NC	-0.318749	-0.507990
N2	NE	-0.323092	-0.522480
C3	CA	-0.187395	0.138902
C4	CA	0.194161	0.238767
C5	CF	-0.004360	0.171311
C6	CA	-0.178360	-0.235721
C7	CA	-0.301687	-0.432107
C8	C	0.335751	0.644779
C9	CA	0.281274	0.600080
C10	C3	-0.682846	0.032659
C11	CD	-0.007420	0.258466
C12	CA	-0.347720	-0.405953
O13	O	-0.445425	-0.618213
O14	O	-0.355803	-0.608688
S15	SS	0.207947	-0.084461
S16	SS	0.141643	-0.254811
H17	HA	0.173894	0.162641
H18	HA	0.162581	0.157964
H19	HA	0.166778	0.197981
H20	H1	0.244164	0.033211
H21	H1	0.244665	0.033664

Table B.1: Atom types and charges of the OLU chromophore as represented in Figure B.1.

B.1.2 Stretching

Taking into account the Hamiltonian description of the Amber FF, then the stretching term is described by:

$$E^{stretching} = \sum_{bonds} k_b(r - r_0)^2 \tag{B.1}$$

where k_b and r_0 represent the stretching force constant in kcal/mol and the bond length in Å, respectively.

Stretching	S_0		S_1	
	k_b	r_0	k_b	r_0
NC-CA	377.701	1.340	236.501	1.380
NC-CD	416.221	1.319	324.665	1.323
NE-CF	458.539	1.309	406.144	1.320
NE-C	294.500	1.379	313.284	1.367
CA-CA	256.038	1.439	277.658	1.412
CA-CA	451.744	1.370	357.402	1.384
CA-SS	158.716	1.756	95.295	1.754
CA-CA	341.153	1.420	342.363	1.414
CF-CD	334.443	1.412	252.306	1.425
CF-SS	119.431	1.786	120.216	1.782
CA-CA	495.168	1.359	444.964	1.374
CA-HA	376.657	1.084	373.389	1.085
CA-CA	241.339	1.465	226.137	1.449
CA-HA	375.522	1.085	377.921	1.084
C -C3	140.708	1.540	131.071	1.544
C -O	746.684	1.218	697.688	1.228
CA-CA	301.545	1.443	256.391	1.457
CA-O	585.278	1.255	553.729	1.257
C3-SS	141.318	1.819	135.249	1.823
C3-H1	359.823	1.090	355.135	1.091
C3-H1	363.312	1.088	357.178	1.090
CD-SS	126.329	1.784	115.780	1.770
CA-HA	378.966	1.083	385.174	1.082

Table B.2: Stretching parameters for the OLU chromophore. Force constants appear in kcal/mol and bond lengths in Å.

B.1.3 Bending

In the case of the bending, the Hamiltonian term is described by:

$$E^{bending} = \sum_{angles} k_\theta(\theta - \theta_0)^2 \tag{B.2}$$

where k_θ and θ_0 represent the bending force constant in kcal/mol/radian^2 and the equilibrium angle in degrees, respectively.

Bending	S_0		S_1	
	k_θ	θ_0	k_θ	θ_0
NC-CA-CA	127.744	116.106	135.009	115.517
NC-CA-CA	136.731	125.480	160.500	125.852
NC-CD-CF	90.143	123.975	94.941	125.092
NC-CD-SS	87.079	114.374	70.479	115.698
NE-CF-CD	111.777	123.523	111.352	123.935
NE-CF-SS	98.118	117.523	89.791	117.373
NE-C -C3	102.894	113.937	105.649	113.807
NE-C -O	107.507	124.563	112.087	125.332
CA-CA-CA	168.666	118.414	161.108	118.632
CA-CA-CA	139.873	119.977	144.296	119.773
CA-CA-HA	57.373	122.115	57.161	122.396
CA-SS-CD	142.809	88.849	148.032	88.348
CA-NC-CD	125.552	112.442	110.362	110.685
CA-CA-CA	148.626	122.435	131.358	122.426
CA-CA-SS	135.996	108.229	102.784	109.753
CA-CA-CA	150.852	119.751	149.373	120.570
CA-CA-HA	62.076	118.682	61.577	118.572
CF-NE-C	134.402	113.139	142.658	113.465
CF-CD-SS	111.565	121.651	83.053	119.210
CF-SS-C3	120.223	88.975	121.190	88.847

Table B.3: Bending parameters for the OLU chromophore. Force constants appear in kcal/mol/radian^2 and equilibrium angles in degrees. Part 1.

Bending	S_0		S_1	
	k_θ	θ_0	k_θ	θ_0
CA-CA-CA	168.734	123.061	160.130	122.228
CA-CA-HA	63.359	120.856	62.997	120.875
CA-CA-HA	64.218	121.567	64.158	120.858
CA-CA-CA	217.531	116.362	217.730	116.369
CA-CA-O	124.901	121.167	122.433	122.962
C -C3-SS	118.433	106.426	118.662	106.507
C -C3-H1	66.861	109.886	66.983	110.297
C -C3-H1	66.273	109.991	66.502	110.372
CA-CA-HA	63.480	116.083	60.755	116.897
CA-CA-HA	60.999	117.908	62.987	117.830
C3-C -O	92.922	121.500	91.674	120.861
CD-CF-SS	117.039	118.955	111.526	118.692
CA-CA-SS	145.301	129.335	121.419	127.821
CA-CA-O	123.460	122.471	114.548	120.669
SS-C3-H1	48.855	110.727	46.200	110.534
SS-C3-H1	48.100	110.794	45.677	110.544
H1-C3-H1	34.684	108.994	33.747	108.587

Table B.4: Bending parameters for the OLU chromophore. Force constants appear in kcal/mol/radian^2 and equilibrium angles in degrees. Part 2.

B.1.4 Torsion

Torsion movements, both proper and improper are included in the Hamiltonian term by:

$$E^{dihedral} = \sum_{dihedrals} V_n(1 + cos(n\phi - \gamma)) \tag{B.3}$$

where V_n represents the torsion barrier term (in kcal/mol); γ, the phase (degrees) and n, the periodicity. Additionally, Amber20 employs a divider term that splits the torsion term into individual contributions for each pair of atoms involved in the torsion.

Dihedral	S_0			
	Divider	V_n	γ	n
CA-CA-CA-NC	4	14.500	180.000	2.000
HA-CA-CA-NC	4	14.500	180.000	2.000
NC-CD-SS-CA	2	2.200	180.000	2.000
NC-CD-CF-NE	4	4.000	180.000	2.000
SS-CD-CF-NE	4	4.000	180.000	2.000
NE-CF-SS-C3	2	2.200	180.000	2.000
NE-C -C3-SS	6	0.000	180.000	2.000
NE-C -C3-H1	6	0.000	180.000	2.000
CA-CA-CA-CA	4	14.500	180.000	2.000
CA-CA-CA-HA	4	14.500	180.000	2.000
CF-CD-NC-CA	2	9.500	180.000	2.000
SS-CD-NC-CA	2	9.500	180.000	2.000
CA-CA-SS-CD	2	0.800	180.000	2.000
C3-C -NE-CF	2	0.400	180.000	2.000
O -C -NE-CF	2	0.400	180.000	2.000
CF-CD-SS-CA	2	2.200	180.000	2.000
CA-CA-CA-O	4	14.500	180.000	2.000
CD-CF-NE-C	2	1.600	180.000	2.000
SS-CF-NE-C	2	1.600	180.000	2.000
C -C3-SS-CF	3	1.000	0.000	3.000
CA-CA-NC-CD	2	9.600	180.000	2.000
CD-CF-SS-C3	1	1.100	180.000	-2.000
CD-CF-SS-C3	1	0.700	180.000	3.000
HA-CA-CA-O	4	14.500	180.000	2.000
O -C -C3-SS	6	0.000	180.000	2.000
O -C -C3-H1	1	0.800	0.000	-1.000
O -C -C3-H1	1	0.000	0.000	-2.000
O -C -C3-H1	1	0.080	180.000	3.000
NC-CA-CA-SS	4	14.500	180.000	2.000
CA-CA-CA-SS	4	14.500	180.000	2.000
HA-CA-CA-SS	4	14.500	180.000	2.000
NC-CD-CF-SS	4	4.000	180.000	2.000
SS-CD-CF-SS	4	4.000	180.000	2.000
HA-CA-CA-HA	4	14.500	180.000	2.000
H1-C3-SS-CF	3	1.000	0.000	3.000

Table B.5: Dihedral parameters for the OLU chromophore in the S_0. Torsional barriers and dividers appear in kcal/mol and phases in degrees.

Dihedral	S_1			
	Divider	V_n	γ	n
CA-CA-CA-NC	4	14.500	180.000	2.000
HA-CA-CA-NC	4	14.500	180.000	2.000
NC-CD-SS-CA	2	2.200	180.000	2.000
NC-CD-CF-NE	4	4.000	180.000	2.000
SS-CD-CF-NE	4	4.000	180.000	2.000
NE-CF-SS-C3	2	2.200	180.000	2.000
NE-C -C3-SS	6	0.000	180.000	2.000
NE-C -C3-H1	6	0.000	180.000	2.000
CA-CA-CA-CA	4	14.500	180.000	2.000
CA-CA-CA-HA	4	14.500	180.000	2.000
CF-CD-NC-CA	2	9.500	180.000	2.000
SS-CD-NC-CA	2	9.500	180.000	2.000
CA-CA-SS-CD	2	0.800	180.000	2.000
C3-C -NE-CF	2	0.400	180.000	2.000
O -C -NE-CF	2	0.400	180.000	2.000
CF-CD-SS-CA	2	2.200	180.000	2.000
CA-CA-CA-O	4	14.500	180.000	2.000
CD-CF-NE-C	2	1.600	180.000	2.000
SS-CF-NE-C	2	1.600	180.000	2.000
C -C3-SS-CF	3	1.000	0.000	3.000
CA-CA-NC-CD	2	9.600	180.000	2.000
CD-CF-SS-C3	1	1.100	180.000	-2.000
CD-CF-SS-C3	1	0.700	180.000	3.000
HA-CA-CA-O	4	14.500	180.000	2.000
O -C -C3-SS	6	0.000	180.000	2.000
O -C -C3-H1	1	0.800	0.000	-1.000
O -C -C3-H1	1	0.000	0.000	-2.000
O -C -C3-H1	1	0.080	180.000	3.000
NC-CA-CA-SS	4	14.500	180.000	2.000
CA-CA-CA-SS	4	14.500	180.000	2.000
HA-CA-CA-SS	4	14.500	180.000	2.000
NC-CD-CF-SS	4	4.000	180.000	2.000
SS-CD-CF-SS	4	4.000	180.000	2.000
HA-CA-CA-HA	4	14.500	180.000	2.000
H1-C3-SS-CF	3	1.000	0.000	3.000

Table B.6: Dihedral parameters for the OLU chromophore in the S_1. Torsional barriers and dividers appear in kcal/mol and phases in degrees.

Dihedral	S_0			S_1		
	V_n	γ	n	V_n	γ	n
CA-CA-CA-SS	1.1	180.0	2.0	1.1	180.0	2.0
CA-CA-CA-NC	1.1	180.0	2.0	1.1	180.0	2.0
CD-NE-CF-SS	1.1	180.0	2.0	1.1	180.0	2.0
CA-CA-CA-HA	1.1	180.0	2.0	1.1	180.0	2.0
C3-NE-C -O	10.5	180.0	2.0	10.5	180.0	2.0
CA-CA-CA-O	1.1	180.0	2.0	1.1	180.0	2.0
CF-NC-CD-SS	1.1	180.0	2.0	1.1	180.0	2.0

Table B.7: Improper torsional parameters for the OLU chromophore. Torsional barriers and dividers appear in kcal/mol. The torsional barrier height is NOT divided by the divider.

B.1.5 Non-Bonding Parameters

Finally, for each pair of atoms the non bonding interaction is given by:

$$V_{ij} = \epsilon \left[(R_{min}/r_{ij})^{12} - 2 (R_{min}/r_{ij})^{6} \right] \tag{B.4}$$

where ϵ represents the 6-12 potential well depth in kcal/mol and R_{min}, the van der Waals radii in Å.

Atom type	S_0		S_1	
	$R_{min}/2$	ϵ	$R_{min}/2$	ϵ
NC	1.8240	0.1700	1.8240	0.1700
NE	1.8240	0.1700	1.8240	0.1700
CA	1.9080	0.0860	1.9080	0.0860
CF	1.9080	0.0860	1.9080	0.0860
C	1.9080	0.0860	1.9080	0.0860
C3	1.9080	0.1094	1.9080	0.1094
CD	1.9080	0.0860	1.9080	0.0860
O	1.6612	0.2100	1.6612	0.2100
SS	2.0000	0.2500	2.0000	0.2500
HA	1.4590	0.0150	1.4590	0.0150
H1	1.3870	0.0157	1.3870	0.0157

Table B.8: Non-bonding parameters for the OLU chromophore van der Waals radii appear in Å and 6-12 potential well depth in kcal/mol.

B.2 Variation of the Oscillator Strength and Charge Transfer for different Density Functionals

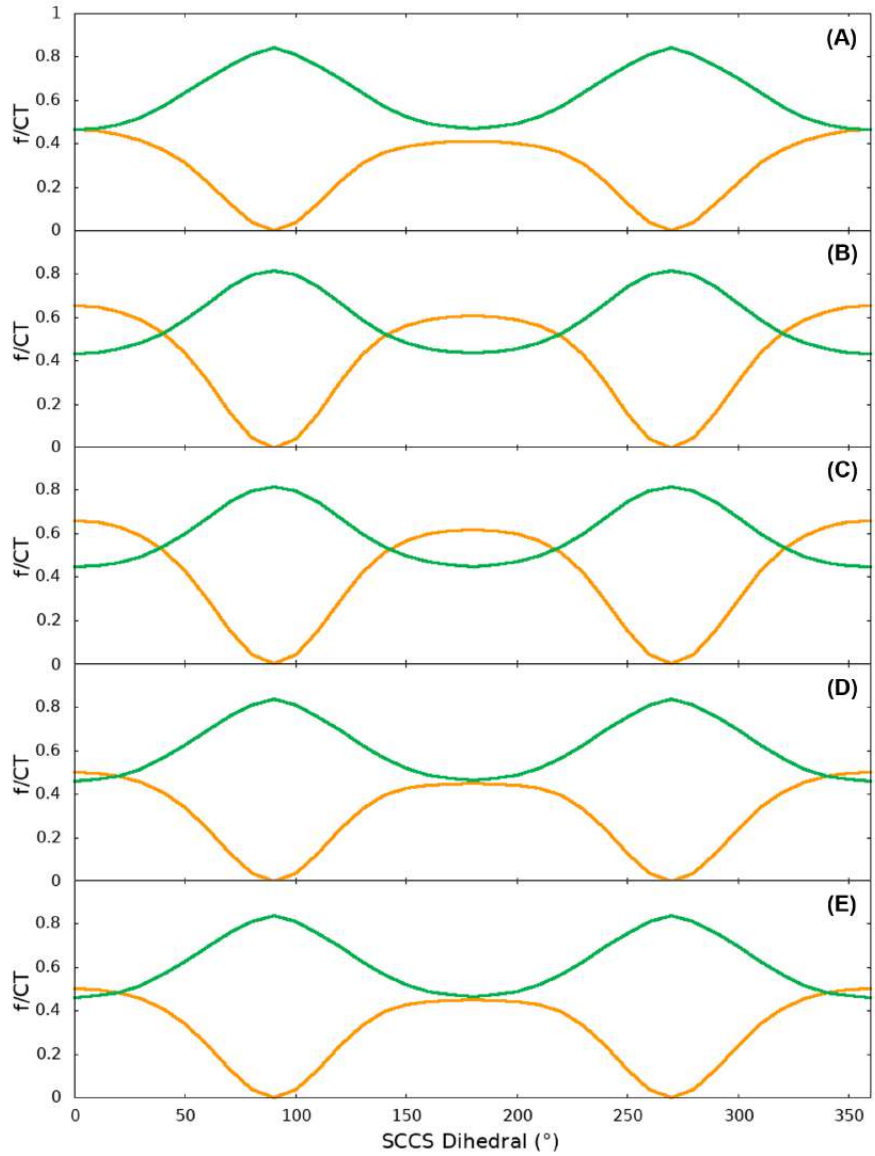


Figure B.2: Variation of the oscillator strength (orange) and CT number (green) for the S_0 - S_1 transition along the dihedral angle calculated with different density functionals: (A) B3LYP, (B) CAM-B3LYP, (C) M06-2X, (D) PBE0 and (E) ωB97XD.

B.3 Absorption and Emission Spectra from the Classical and QM/MM Trajectories and Angle Contributions.

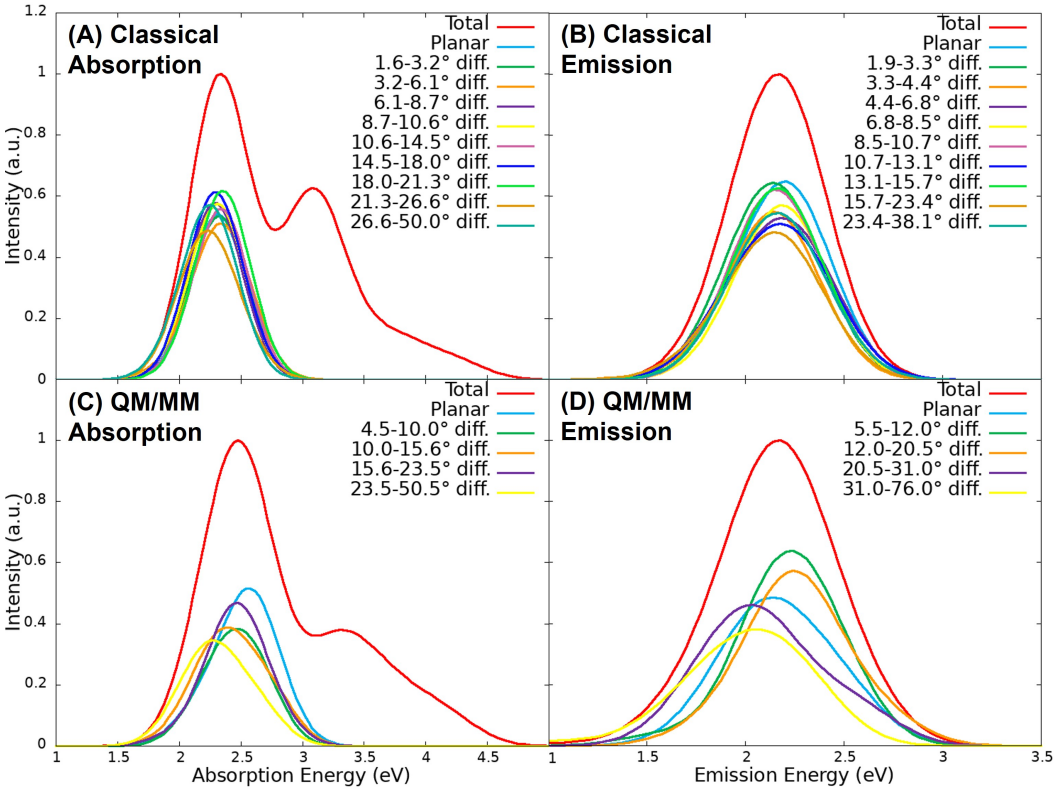


Figure B.3: (A,C) Absorption and (B,D) emission spectra computed from the classical (A,B) and QM/MM (C,D) trajectories, together with their respective contributions from the S_0 to the S_1 band depending on the SCCS dihedral angle.

B.4 Probability Distribution of the SCCS Dihedral Angles for the 500 ns Trajectory

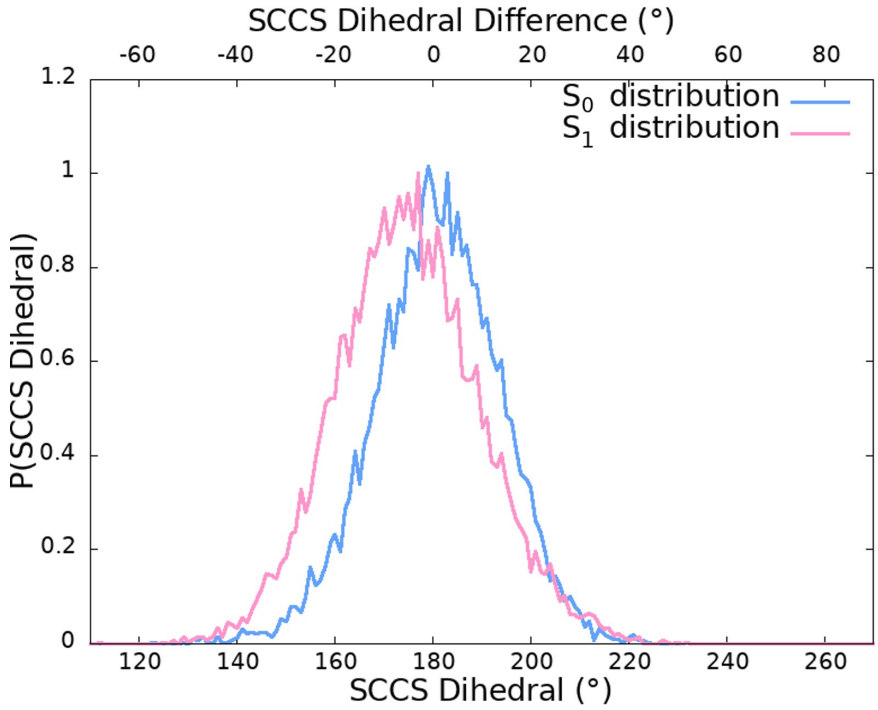


Figure B.4: Probability distributions of the SCCS dihedral angles for the S_0 (blue) and the S_1 (pink) trajectories evolved at classical level for 500 ns.

B.5 Environment of the Two Clusters with the Smallest and Largest CT Number Values from the QM/MM Trajectory of the S_0 State.

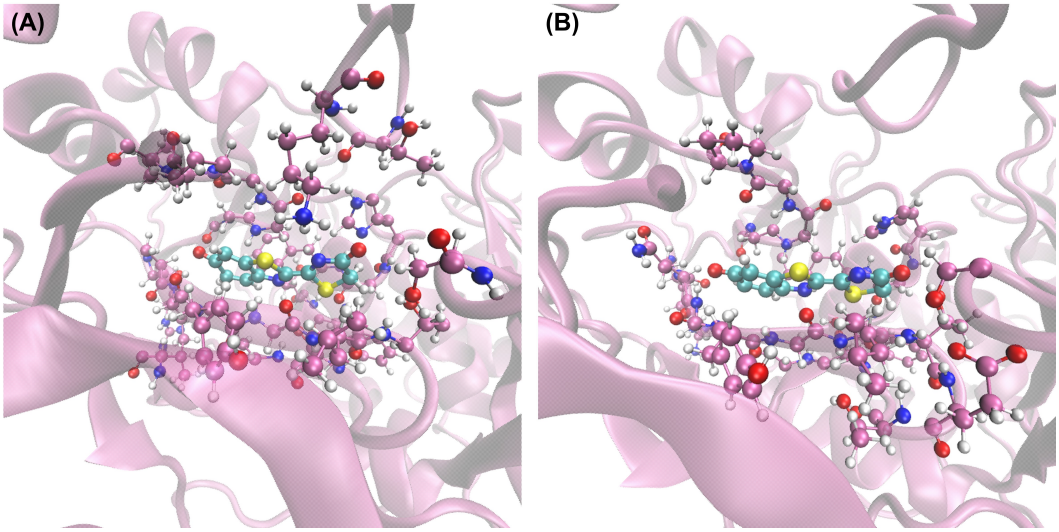


Figure B.5: Representative structure with the protein environment of the two clusters with the 20 snapshots with smallest (A) and largest (B) CT number values extracted from the QM/MM trajectory of the S_0 state.

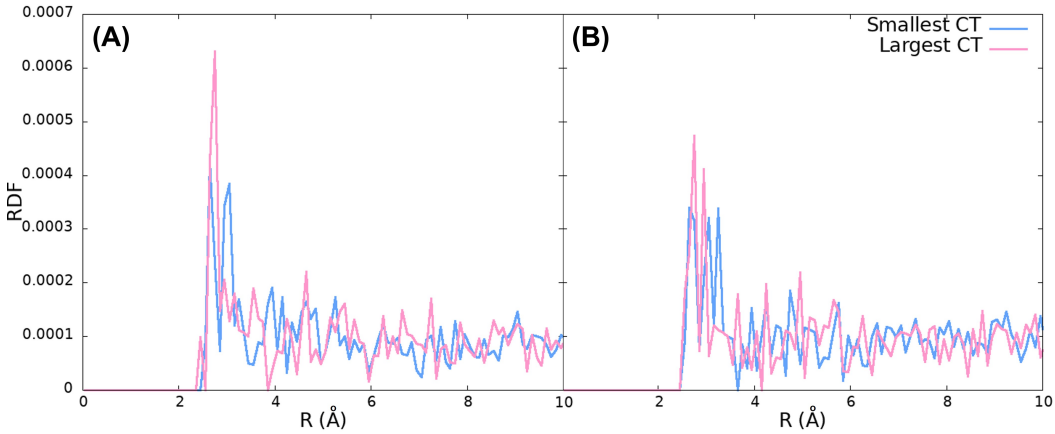


Figure B.6: RDF between (A) the O atom of the tz fragment and the O atom of water and (B) the O atom of the btz fragment and the O atom of water for the two clusters with the 20 snapshots with smallest (blue) and largest (pink) CT number values extracted from the QM/MM trajectory of the S_0 state.

Cluster	<CT>	H_2O		
		O^{tz}	O^{btz}	Total
Small	0.45±0.06	0.1±0.3	0.0±0.0	0.1±0.3
Large	0.66±0.06	0.2±0.4	0.3±0.4	0.5±0.6

Table B.9: Number of hydrogen bonds between the O atom of the tz and btz fragments of the OLU and the surrounding water molecules.

Cluster	<CT>	Protein		
		O^{tz}	O^{btz}	Total
Small	0.45±0.06	0.0±0.0	0.6±1.5	0.6±1.5
Large	0.66±0.06	0.2±0.9	0.4±1.2	0.6±1.5

Table B.10: Number of hydrogen bonds between the O atom of the tz and btz fragments of the OLU and the surrounding protein.

Cluster	<CT>	Protein + H_2O		
		O^{tz}	O^{btz}	Total
Small	0.45±0.06	0.1±0.3	0.6±1.5	0.7±1.5
Large	0.66±0.06	0.4±0.9	0.7±1.4	1.1±1.6

Table B.11: Number of hydrogen bonds between the O atom of the tz and btz fragments of the OLU and its environment.

B.6 Total Binding Free Energy per Residue of the Classical S_1 Trajectories Without and With AMP

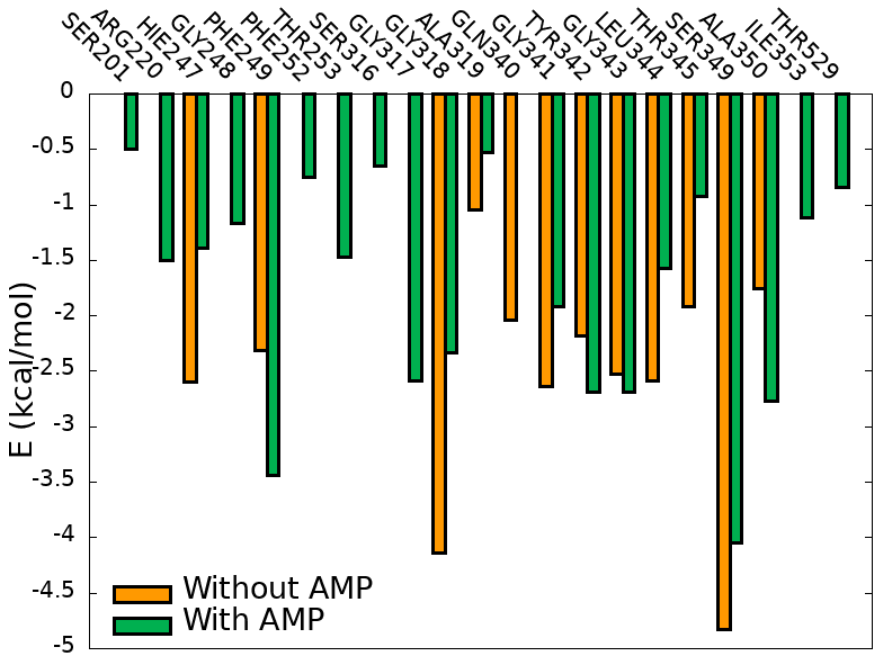


Figure B.7: Total binding free energy per residue of the classical S_1 trajectories without (orange) and with AMP (green) for the residues that are closer for at least 1% of the simulation time.

B.7 Representation of the Three Positions Found for the Analysed Snapshots of the S_0 QM/MM Trajectory

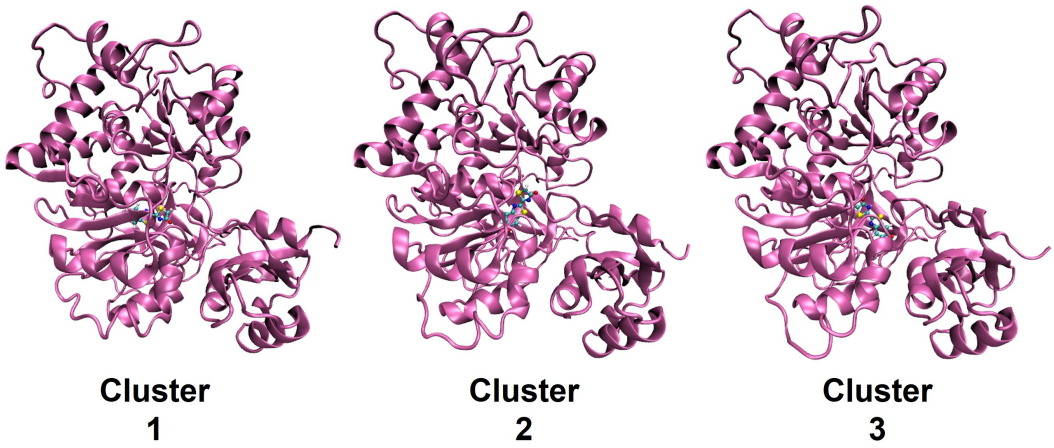


Figure B.8: Representation of the three clusters of the analysed snapshots of the S_0 QM/MM trajectory

B.8 Charge Transfer Number Variation with Respect to the Distance of the Closest Amino-acids of the Protein

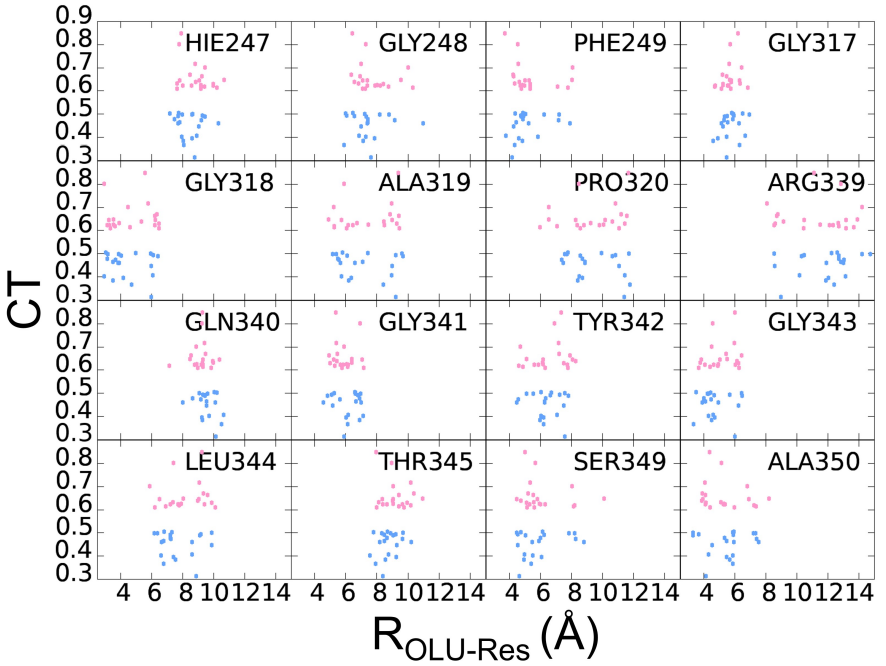


Figure B.9: CT number variation with the distance between the centres of mass of the chromophore and different amino-acids for the clusters of the 20 geometries with smallest (blue) and largest (pink) CT number values

Appendix C

OLUFF: a Novel Set of Ground and Excited State FF Parameters of the Emitting Oxyluciferin Species

C.1 System Preparation and Molecular Dynamics simulations

In order to prepare the structure of the chromophore for the classical MD, the *tleap* input shown in Listing C.1 was employed. In this input, the first four lines correspond to the loading of the parameters and initial structure while the rest, to the creation of a PDB file, a parameters file and a restart file to perform the initial minimisation of the system.

Listing C.1: Example of the input file to generate the parameters and initial structure to perform the clasical MD simulations. In this example, *leaprc.custom* contains the information about the atom types of the system and *OLU.mol2*, the structure and charges of the OLU chromophore.

```
# Load new parameters
source leaprc.custom

# Load mol2 file
SYS=loadmol2 OLU.mol2

# Create pdb file
```

```
savepdb SYS final_sistema.pdb
# Create params and initial structure for minimization
saveamberparm SYS final_system.prmtop final_system.rst7

quit
```

Consecutively, the initial structure was minimised using the input shown in Listing C.2, in which non-periodic boundary conditions are used and the non-bonded cutoff is define so all the interactions are considered.

Listing C.2: Example of the input file to the initial minimisation of the structure to perform the clasical MD simulations.

```
Minimization input file in vacuum
 &cntrl
    ! Minimization options
    imin=1,          ! Turn on minimization
    ntx=1,           ! Read position but not velocities
    irest=0,         ! Do not restart the simulation
    maxcyc=10000,    ! Max number of minimization cycles
    ncyc=5000,       ! Method switch from steepest descent
                       to conjugate gradient after NCYC
    ntb=0,           ! Non-periodic boundary conditions

    ! Potential energy function options
    cut=999.0,       ! Non-bonded cutoff in Angstroms
                     !    - calculate all

    !Restraint options
    nmropt=1,        !Restraints

    ! Control how often information is
    ! printed to the output file
    ntpr=100,        ! Print energies every 100 steps
    ntwx=0,          ! No trajectory file
 /
&wt type = 'END' /
```

Then, the minimised structure was heated by means of the input in Listing C.3 considering an initial interval of 500 ps in which the system was gradually heated followed by 500 ps used to stabilise it.

Listing C.3: Example of the input file for the heatting of the minimised structure.

```
Heat
 &cntrl
   imin=0,          ! Turn off minimization
   ntx=1,           ! Read position but not vels
   irest=0,         ! Do not restart the simulation

   nstlim=500000,   ! Perform MD for 500,000 steps
   dt=0.002,        ! 2 fs time step
   ntf=2, ntc=2,    ! Shake hydrogen atoms
   tempi=0.0,       ! Initial temperature
   temp0=300.0,     ! Target temperature

   ! Control how often information is
   ! printed to the output file
   ntpr=100,        ! E. info. printed every ntpr steps
   ntwx=100,        ! Coords in mdcrd every ntwx steps

   ! Potential energy function options
   cut=999.0,       ! Non-bonded cutoff in Angstroms

   ! Ensemble control
   ntb=0,           ! Non-periodic boundary conditions
   ntp=0,           ! No pressure scaling (Default)
   ntt=3,           ! Langevin dynamics
   gamma_ln=2.0,    ! Collision frequency for the
                    ! Langevin dynamics in ps-1

   ! Wrap coordinates when printing them
   ! to the same unit cell
   nmropt=1,
   ! Random seed to use to generate vel distribution
   ig=-1,
 /

! Change conditions during simulation
! Increase T from 0 to 300K during steps 0 to 250000
&wt type='TEMP0', istep1=0, istep2=250000,
                          value1=0.0, value2=300.0 /
```

```
! Keep T cte for the rest of the simulation
&wt type='TEMP0', istep1=250001, istep2=500000,
                        value1=300.0, value2=300.0 /
&wt type='END' /
```

Finally, as shown in the input in Listing C.4, the system was allowed to evolve for 1 μs.

Listing C.4: Example of the input file to the production of the MD simulation.

```
A NPT simulation for common production-level simulations
 &cntrl
    imin=0,             ! No minimization
    irest=1,            ! This IS a restart of an old MD
    ntx=5,              ! So our inpcrd file has vels

    ! Temperature control
    ntt=3,              ! Langevin dynamics
    gamma_ln=1.0,       ! Friction coefficient (ps^-1)
    temp0=303.15,       ! Target temperature
    ig=-1,              ! Seed for velocities

    ! Potential energy control
    cut=999.0,          ! Non-bonded cutoff, in Angstroms

    ! MD settings
    nstlim=500000000, ! 1 micros total
    dt=0.002,           ! time step (ps)

    ! SHAKE
    ntc=2,              ! Constrain bonds containing H
    ntf=2,              ! Don't calculate forces of bonds
                        !                    containing H

    ntb=0,              ! Non-periodic

    ! Control how often information is printed
    ntpr=1000,          ! Print energies every 1000 steps
    ntwx=25000,         ! Coords every ntwx steps to traj
    ntwr=10000,         ! Print a rst file every 10K steps
    ntwv=1000,          ! Print vels every 1000 steps
    ntxo=2,             ! Write NetCDF format
```

```
    ioutfm=0,           ! Formatted ASCII trajectory

  ! Wrap coordinates when printing them
  ! to the same unit cell
    nmropt=1,

    ! Constant pressure control.
    !barostat=1,        ! Berendsen barostat
    !ntp=1,             ! isotropic
    !pres0=1.0,         ! Target external pressure (bar)

 &ewald
    vdwmeth = 0,
 /
&wt type = 'END' /
```

C.2 CAM-B3LYP Basis Set Comparison

Figure C.1, that represents the comparison of the rigid potential energy curves of the SCCS dihedral angles of the OLU phenolate-keto in the ground state at CAM-B3LYP/6-311G(2d,p)CAM-B3LYP/6-311G(2d,p) and CAM-B3LYP/6-31G* levels of theory, shows that they are very similar in shape. Thus, a quantum-mechanic molecular dynamics simulation at CAM-B3LYP/6-31G* level of theory should obtain a similar P(SCCS) distribution than a simulation at CAM-B3LYP/6-311G(2d,p), which is much more computationally demanding.

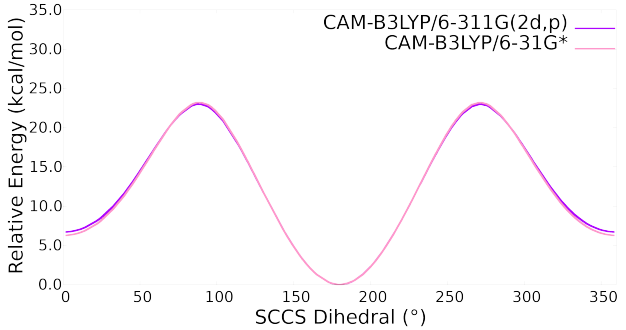


Figure C.1: Comparison of the rigid PESs of the SCCS dihedral angles of the OLU phenolate-keto in the S_0 at CAM-B3LYP/6-311G(2d,p) and CAM-B3LYP/6-31G* levels of theory.

C.3 Nature of the S_1 Transition

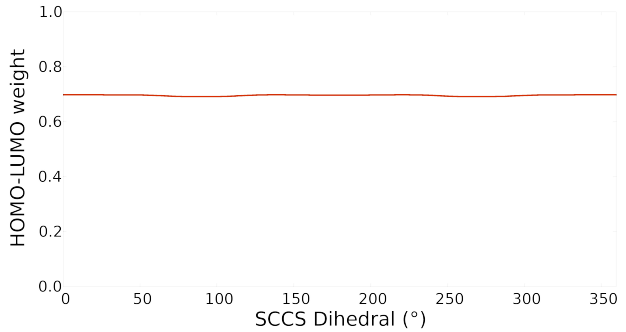
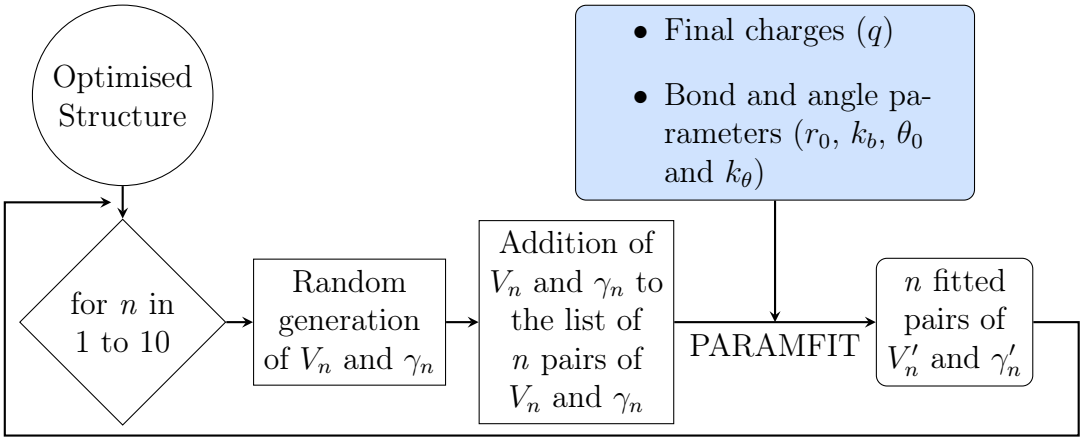


Figure C.2: Participation of the HOMO-LUMO transition on the S_1 state of the OLU phenolate-keto.

C.4 Convergence of the Fitting Process

Additionally to the parametrization, the convergence of the fitted parameters with the number of terms in Eq. 2 after the fitting process has also been studied for the OLU phenolate-keto S_1. For this, instead of generating randomly all the parameter for all the fittings, for each of the five studied cases the random generation of the initial guesses was only performed for each n the first time that term was added to the expression as depicted in Scheme C.1. For instance, for the fist fitting, in which there is only one term in the summation, the initial guesses V_1 and γ_1 are generated randomly. However, in the second one only V_2 and γ_2 are randomly generated while the initial guesses V_1 and γ_1 are the same than in the previous fitting. In the same way, when three terms are included in the expression, only V_3 and γ_3 are obtained randomly and while the initial guesses V_2 and γ_2 and V_1 and γ_1 are the same than in the previous case and so on and so forth. As it can be observed in Scheme C.3, the fitted V_n' presents small oscillations within 2 kcal/mol for $n = 2 - 10$ whereas for $n = 1$, the fitted V_1' is always much smaller than for the rest of the n values. For γ_n', the oscillations of all the parameters are even smaller since 0 and 360^{o} represent the same angle, with the exception of the fitted γ_7' that presents strong oscillations in all the cases. From this results, it can be observed that regardless of the number of terms in the dihedral energy expression the studied parameters converge to a very similar value.



Scheme C.1: Convergence process.

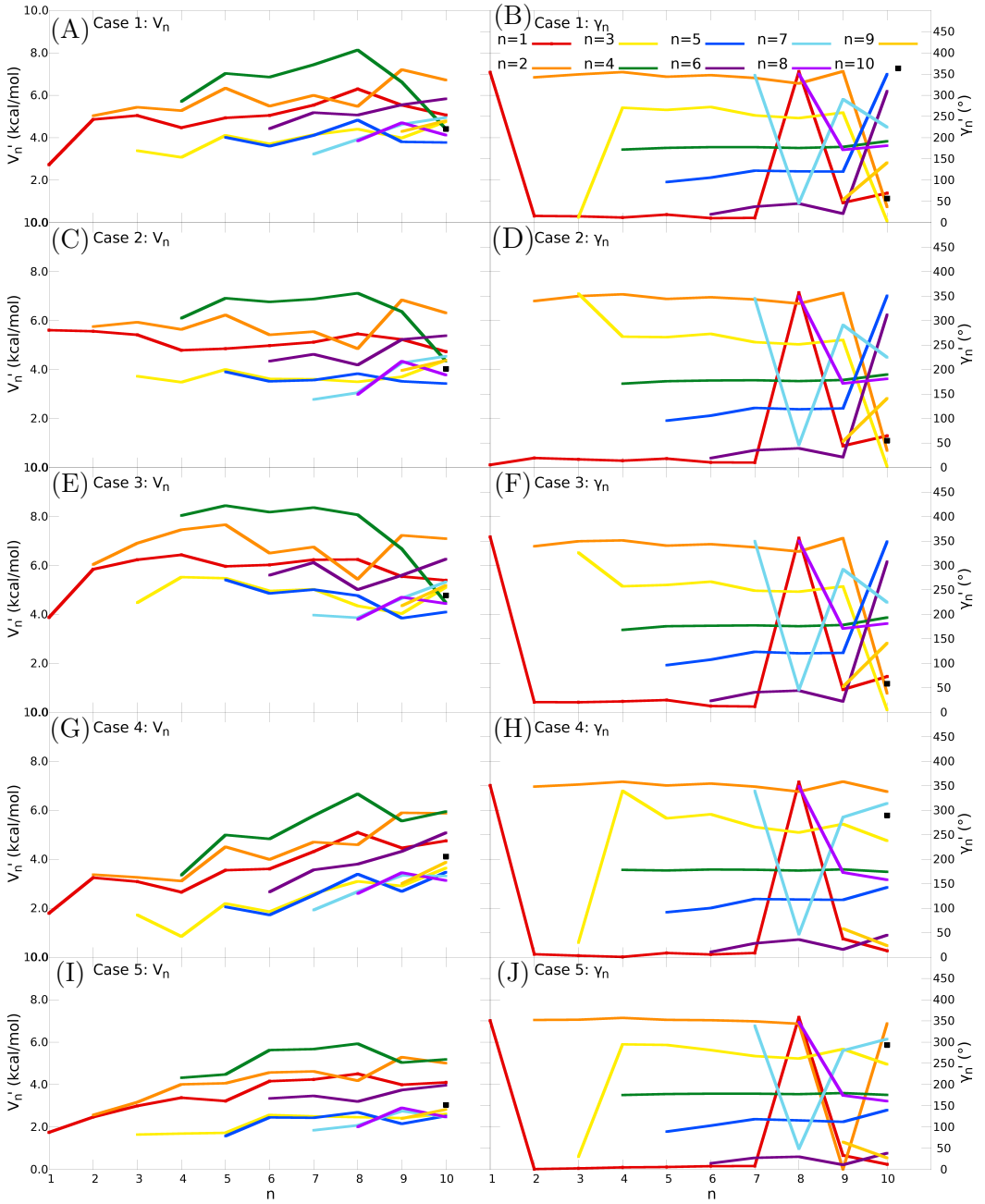


Figure C.3: Convergence of the fitted V_n' (A, C, E, G, I) and γ_n' (B, D, F, H, J) terms with respect to the number of terms in the fitted expression for the SCCS dihedral angle for five different cases of the fitting of the OLU phenolate-keto S_1.

C.5 Fitting Process

Both C.4 and C.5 Figures represent the final fitted torsional barriers (Figure C.4) and torsional phases (C.5) for all the studied chromophores in the ground and first excited states for the final selected n values. As it can be observed, most of the fittings tend to oscillate around the same fitted V_n' and γ_n' values for each of the n terms which means that they tend to converge to the same value. However, as it can be observed in Figure C.6, the fitted value of the torsional barriers is slightly biased by the initial guess since higher values of these initial guesses usually lead to higher fitted barriers. On the flip side, Figure C.7 shows how the fitted phases converge to the same value regardless the initial guess. Thus, there is no bias for the fitting of the phases.

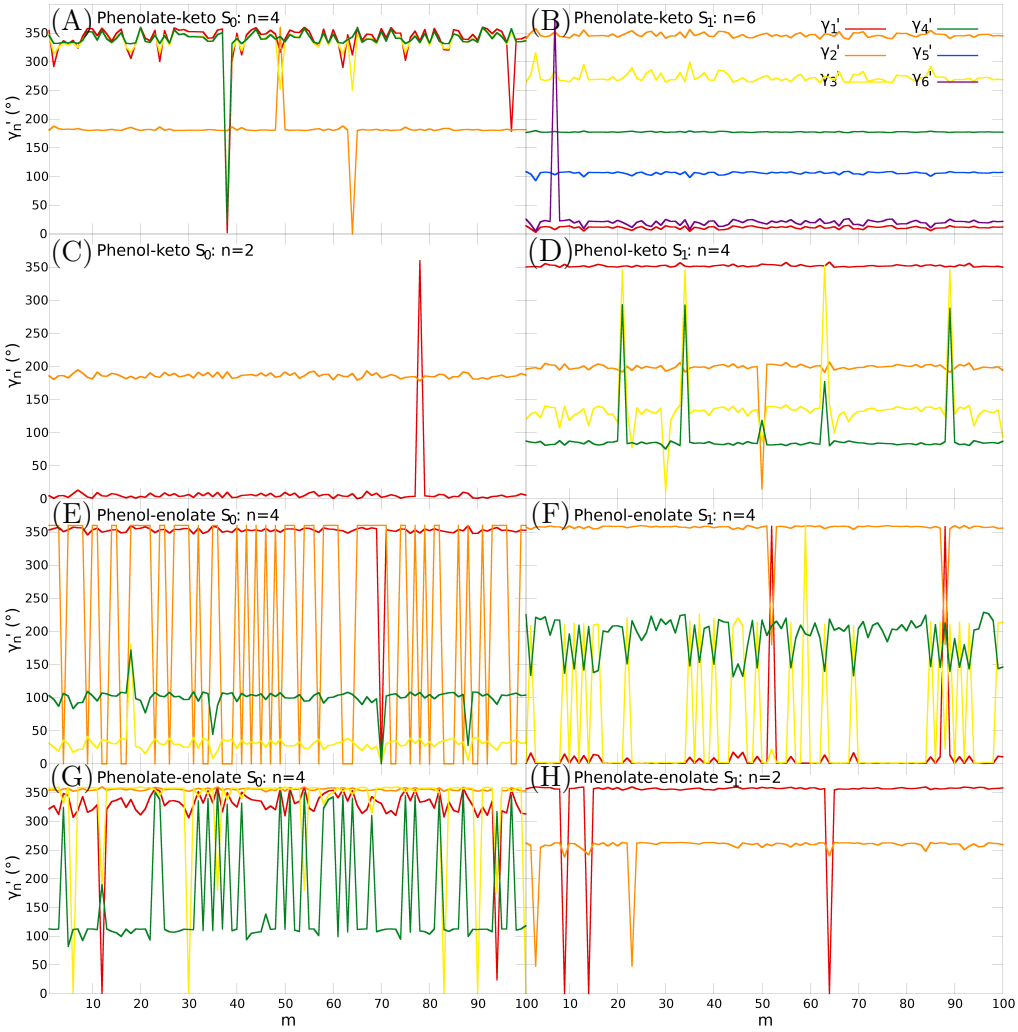


Figure C.5: Fitted torsional phases of the n SCCS dihedral angle parameters for each of the 100 m fittings for the (A, C, E, G) S_0 and (B, D, F, H) S_1 states of the OLU (A, B) phenolate-keto, (C, D) phenol-keto, (E, F) phenol-enolate and (G, H) phenolate-enolate.

Figure C.4: Fitted torsional barriers of the n SCCS dihedral angle parameters for each of the 100 m fittings for the (A, C, E, G) S_0 and (B, D, F, H) S_1 states of the OLU (A, B) phenolate-keto, (C, D) phenol-keto, (E, F) phenol-enolate and (G, H) phenolate-enolate.

Figure C.6: Fitted torsional barriers of the n SCCS dihedral angle parameters with respect to their respective initial guesses for the (A, C, E, G) S_0 and (B, D, F, H) S_1 states of the OLU (A, B) phenolate-keto, (C, D) phenol-keto, (E, F) phenol-enolate and (G, H) phenolate-enolate.

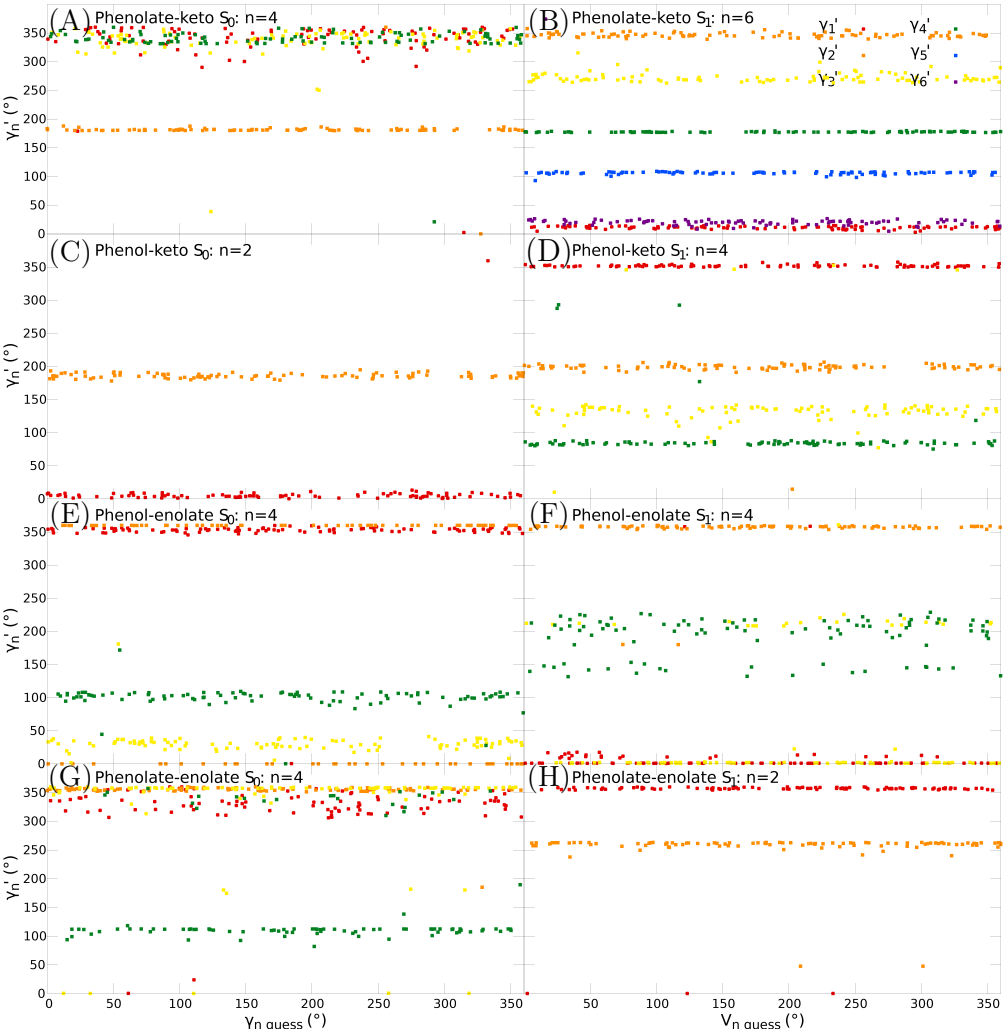


Figure C.7: Fitted torsional phases of the n SCCS dihedral angle parameters with respect to their respective initial guesses for the (A, C, E, G) S_0 and (B, D, F, H) S_1 states of the OLU (A, B) phenolate-keto, (C, D) phenol-keto, (E, F) phenol-enolate and (G, H) phenolate-enolate.

C.6 Periodicity of the Torsional Functions Outside of the Fitted Domain

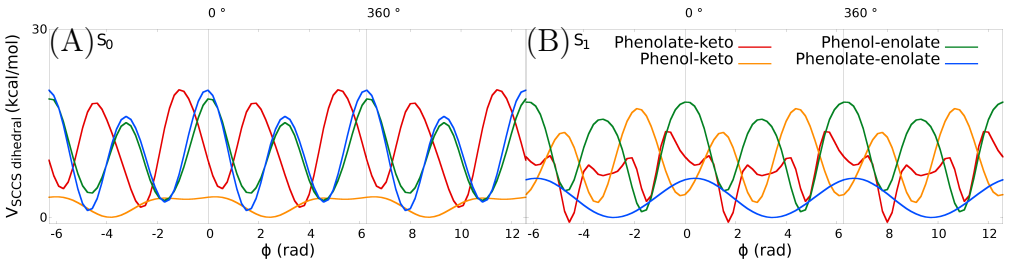


Figure C.8: Periodicity of the torsional functions of the (A) S_0 and (B) S_1 OLU tautomers outside of the fitted domain.

C.7 Evolution of the SCCS Dihedral Angle

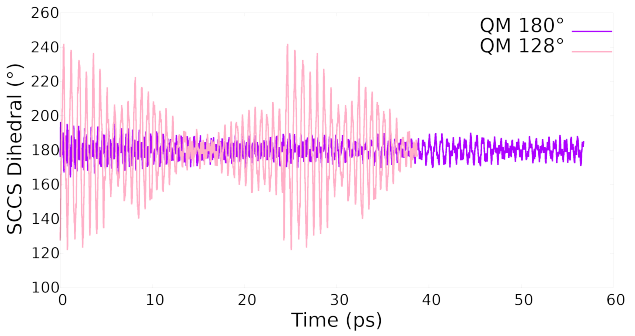


Figure C.9: Evolution of the SCCS dihedral angle of the OLU phenol-enolate in the S_0 along a QM MD simulation starting from two different dihedral angle values.

C.8 Infrared Spectra

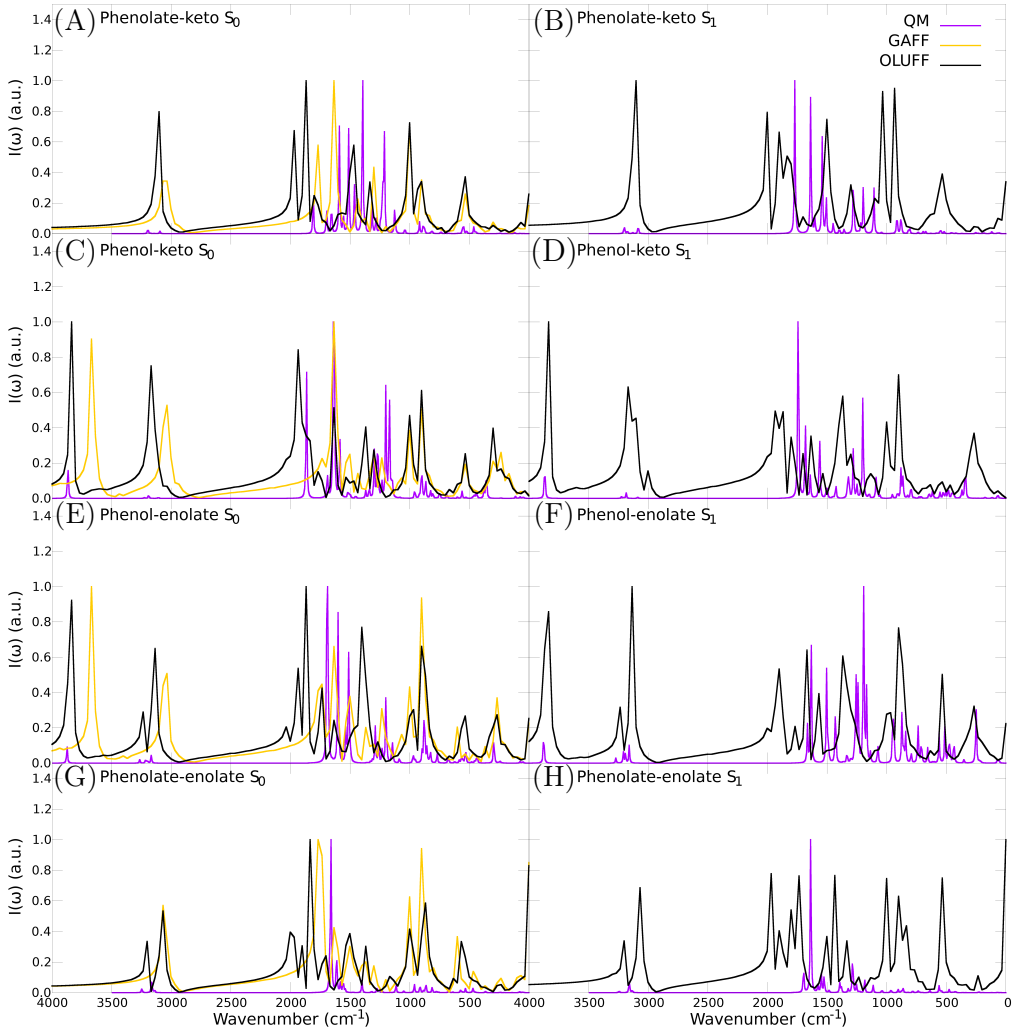


Figure C.10: Infrared spectra of the OLU (A, B) phenol-keto, (C, D) phenol-enolate and (E, F) phenolate-enolate of the (A, C, E) S_0 and (B, D, F) S_1 states calculated from the QM geometry optimization and the MD simulations using the GAFF and the parametrised FF.

Appendix D

Structure of the Oxyluciferin Emitting Species in Aqueous and Luciferase Media

D.1 Root Mean Square Displacement of the Protein

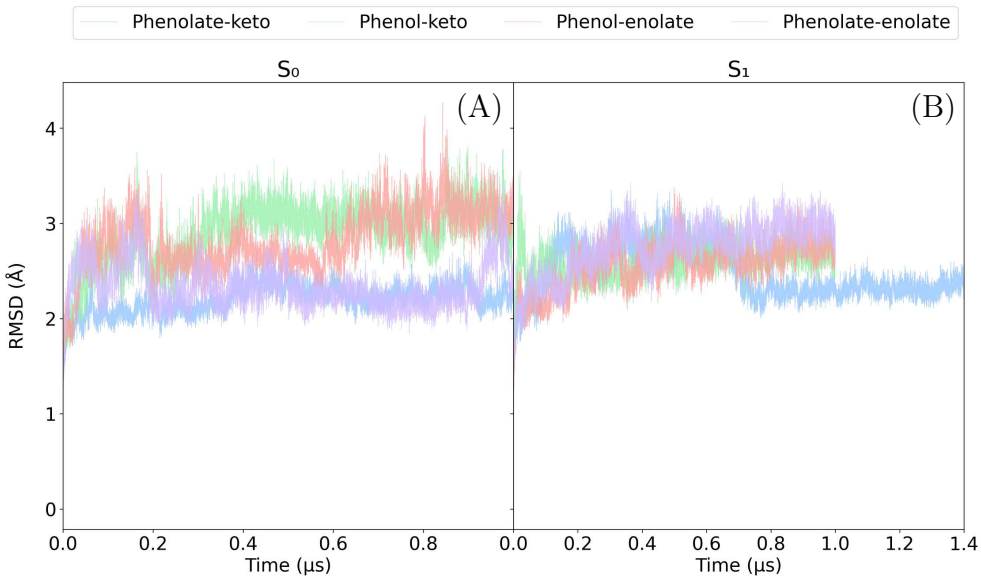


Figure D.1: RMSD of the protein for the MD trajectories with the four OLU chromophores in the (A) S_0 and (B) S_1 states.

D.2 Root Mean Square Displacement of the OLU Chromophores in Aqueous Media

As it can be observed in Figure D.2, even though the chromophores are aligned before the RMSD measurement, they still present large oscillations. In the case of the ground state the phenolate-enolate OLU presents the largest oscillations and the phenol-keto, the lowest while both phenolate-keto and phenol-enolate present similar RMSD values. S_1 aqueous simulations present a different trend in RMSD. In this case, both enolate forms present a similar RMSD value, that is lower than those of S_0 simulations. However, the keto forms present larger oscillations in the excited state, specially the monoanion, for which the large oscillations arise from the population of the three possible minima within the only torsional degree of freedom that is truly able to rotate, the SCCS dihedral angle.

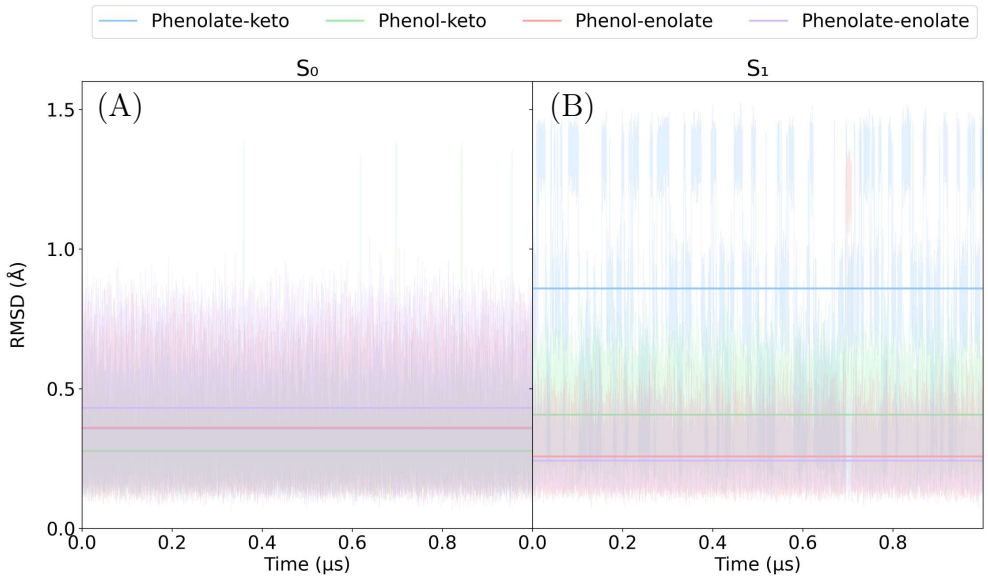


Figure D.2: RMSD of each of the OLU chromophores in the (A) S_0 and (B) S_1 states in aqueous media.

D.3 Change in the SCCS Dihedral of the Phenol-Keto OLU in the S_0

As it can be observed in Figure D.3, at the beginning of the simulation the hydrogen bond acceptor of SER349 is only close the the N atom of the btz moiety of the OLU, the O of the btz moiety is close to SER316 and LYS531 and the O of the tz moiety is relatively far from both GLY318 and GLY343. However at around 61 ns of simulation (Figures D.4(A,B)), the distances between the N atoms and the O of the tz moiety with SER349 (Figure D.3(C)) and between the O atom of the btz moiety and GLY318 and GLY343 (Figure D.3(F)) start to decrease and, at the same time, the chromophore moves away from SER316 and LYS531 (Figure D.3(G)). This movement, which comprehends the simulation time between the first 2 vertical lines, takes 7 ns and it is accompanied by the separation of the chromophore away from AMP (Figure D.3(H)). After this, the system reorganizes. The N atoms and the O atom of the tz moiety of the OLU move away form SER349, remaining close to one another, in contrast to the initial situation (Figure D.3(C)). At around 125 ns of simulation, the distance with SER316 starts to increase, while it remains similar for LYS531 induced by the positioning of the hydrogen bond donor of SER349 at an equidistant position from the N atoms and the O atom of the tz moiety of the OLU (Figures D.3(C,G)). This conformational change allows an extra water molecule to enter inside the pocket bridging the SER349 and AMP at 170 ns, between the next two vertical lines (Figure D.3(B) and Figure D.4(C)). Consecutively, two additional water molecules enter the cavity (Figure D.4(D)), tilting the chromophore and generating a more polar environment in the side of the SER349, thus giving an extra push for the conformational change of the OLU from the s-trans to the s-cis at 210 ns (Figure D.3(A)). Once this happens, one of the water on the side of the SER349 leaves and the chromophore begins to form a hydrogen bond between the SER349 via the O atom of the tz moiety (Figure D.3(C) and Figure D.4(E)), while the remaining water molecules bridges the hydrogen bond donor of SER349 with the two N atoms of the OLU. At the same time, the chromophore moves away from GLY318 and GLY343 (Figure D.3(F)). When the water molecule that bridged the O of the tz moiety with SER349 and AMP leaves (Figure D.3(F)), the hydrogen bond between the OLU and SER349 also disappears (Figure D.3(C)) and the phosphate group of the AMP approaches again the chromophore (Figure D.3(H)), thus making it impossible to return to the s-trans conformation. Moreover, the btz moiety moves closer to SER316 although the distance to LYS531 remains similar leading to a different chemical environment than for

the s-cis conformation (Figure D.3(G)).

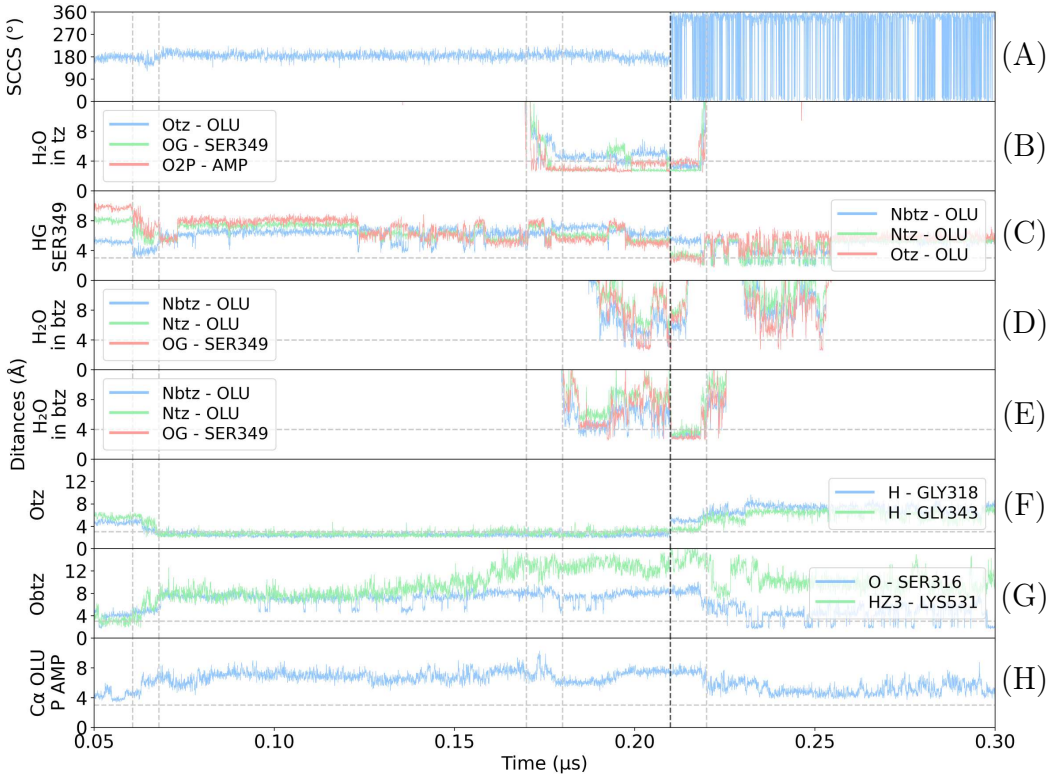
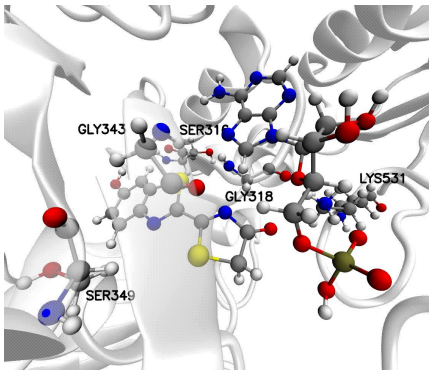
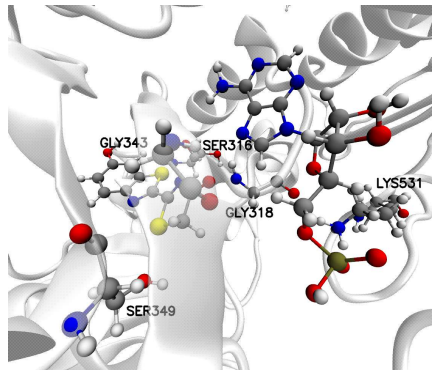


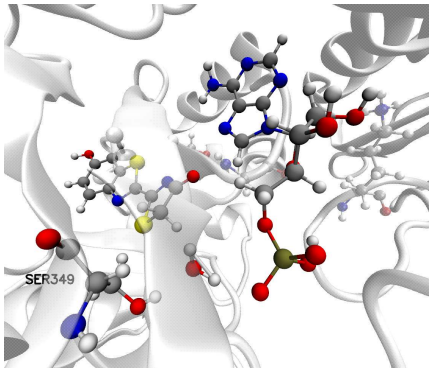
Figure D.3: (A) SCCS dihedral angle of the phenol-keto OLU during the first 300 ns of the MD simulation in the S_0 state. (B-H) Different distances along the simulation. (B) Distances from the O atom of one water molecule to the O atoms of OLU, SER349 and AMP. (C) Distances from the hydrogen bond (HB) donor of SER349 to different HB acceptors of the chromophore. (D,E) Distances from the O atom of two different water molecules to the N atoms of the OLU and the HB acceptors of SER349. (F) Distances from the O atom of of the tz moiety to the lateral chain of GLY318 and GLY343. (G) Distances from the O atom of of the btz moiety to different atoms of SER316 and LYS531 labelled as in the FF. (H) Distances from the C_α of the tz moiety of the OLU to the P atom of the AMP. Vertical-doted lines stress the relevant regions along the simulation and the darkest one represents the conformational change. Horizontal-doted lines represent the HB threshold of 3 Å. Taking into account that for some distances there is a H-O bond in between the O-O distance, this line is shifted to 3.98 Å when required.



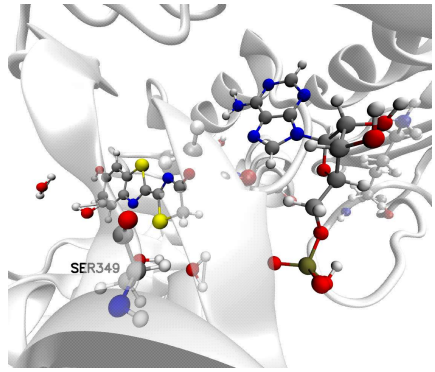
(A) 61 ns



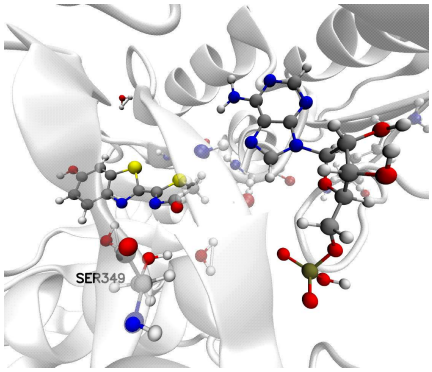
(B) 68 ns



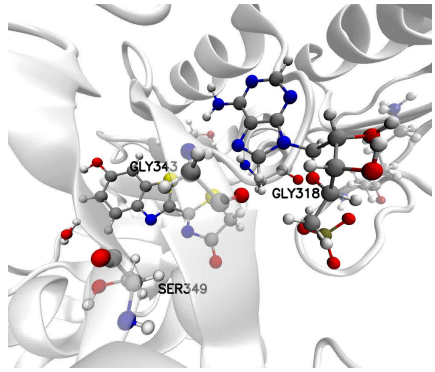
(C) 180 ns



(D) 202.5 ns



(E) 215 ns



(F) 220 ns

Figure D.4: Representative snapshots of the phenol-keto OLU MD simulation in the S_0 during the first 300 ns.

Appendix E

Characterization of the Nuclear Coordinate for the Phenolate-Keto Oxyluciferin Red-Shift Emission from Quantum Mechanical Calculations

E.1 Convergence of the Sampling of the Principal Components

The convergence of the sampling of the PC with respect to the number of frames has been studied, for each of the explored conformations, by projecting the MW PC on the MW displacements along the trajectory and representing their probability distributions. Hence, each of the vertical lines in Figures E.1-E.21 represent a histogram of the projection for different frames and the convergence is reached when the vertical lines are the same.

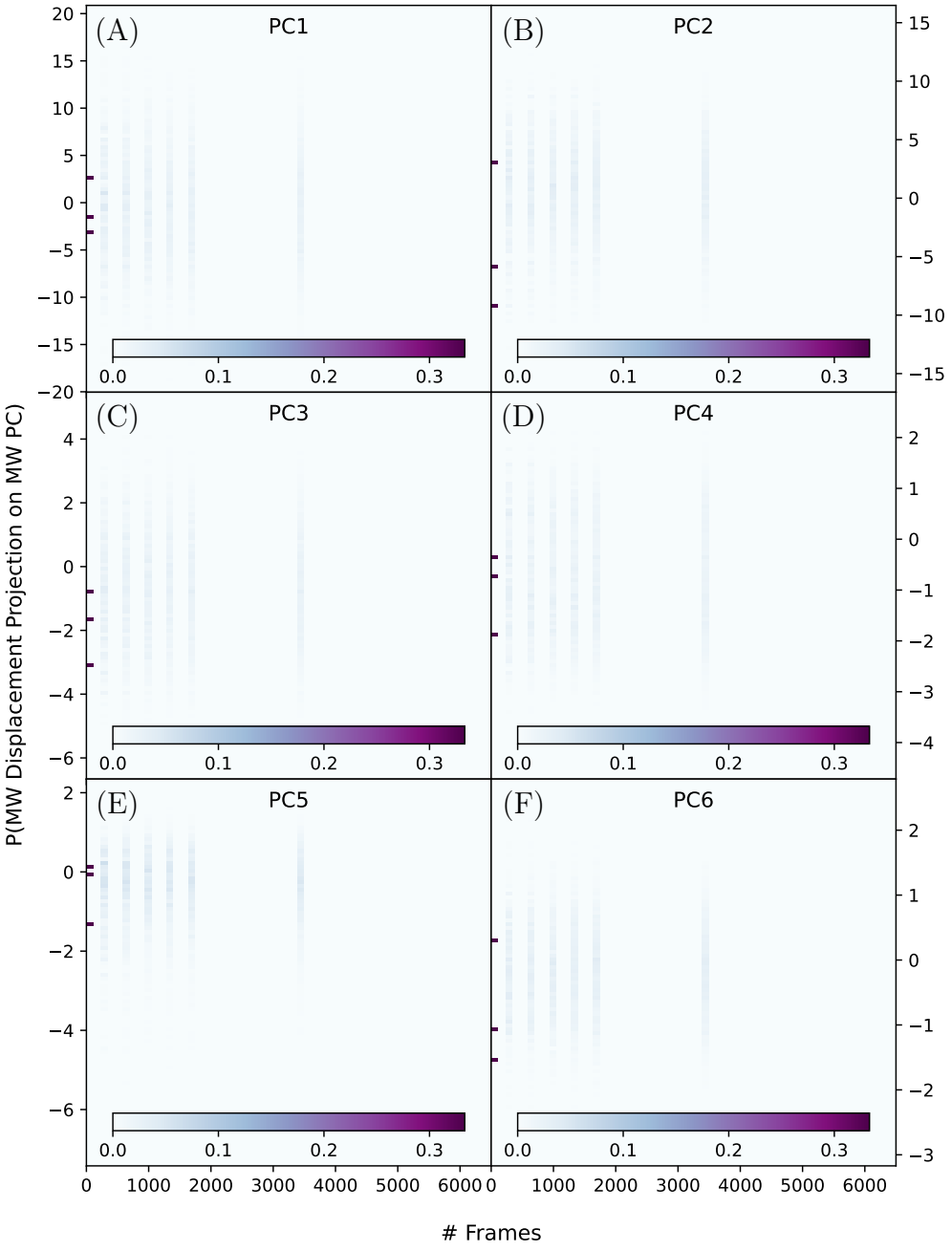


Figure E.1: Convergence of the sampling of PCs in the 90º conformation. The colour bar represents the probability distribution of the projection. Part 1 of 7.

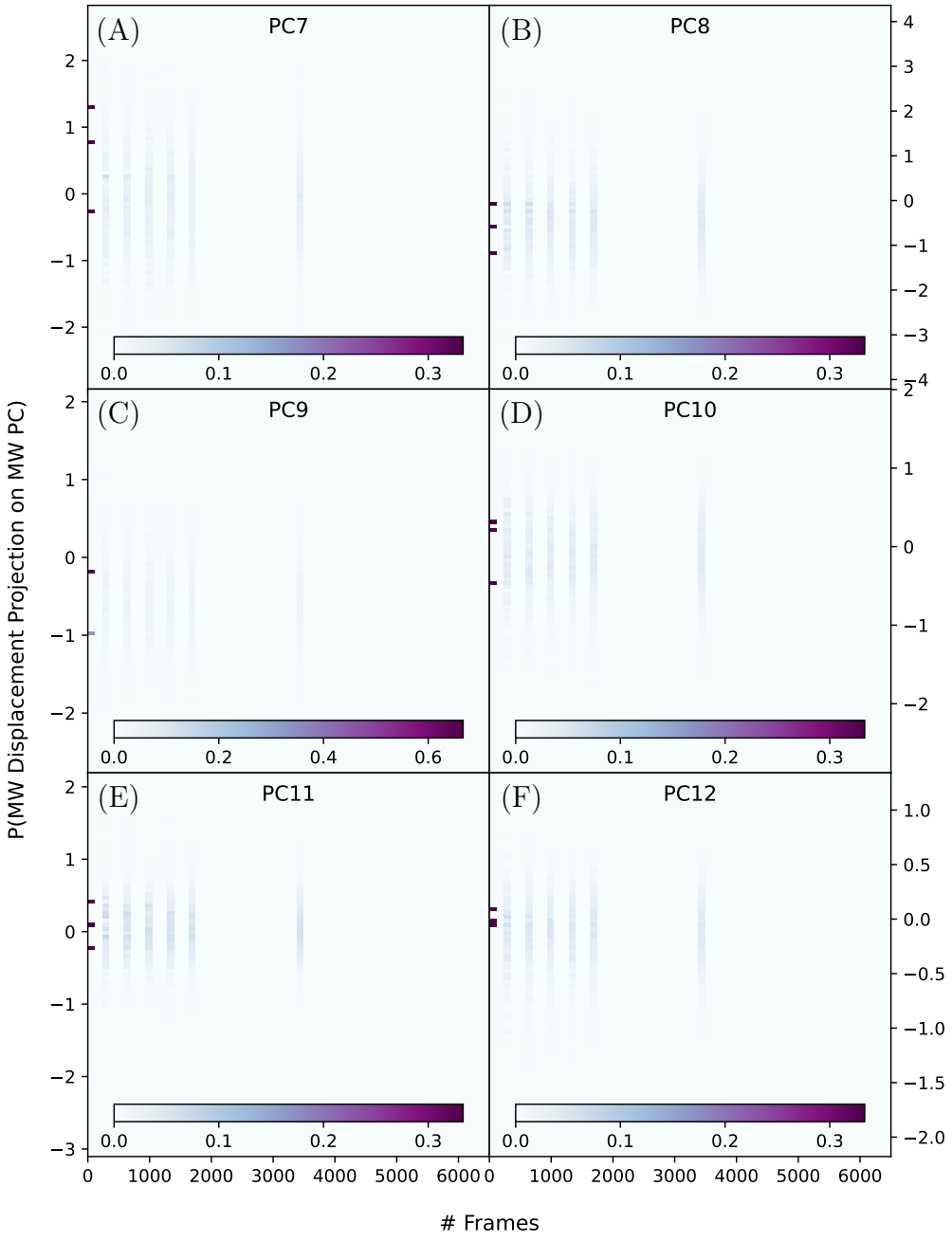


Figure E.2: Convergence of the sampling of PCs in the 90º conformation. The colour bar represents the probability distribution of the projection. Part 2 of 7.

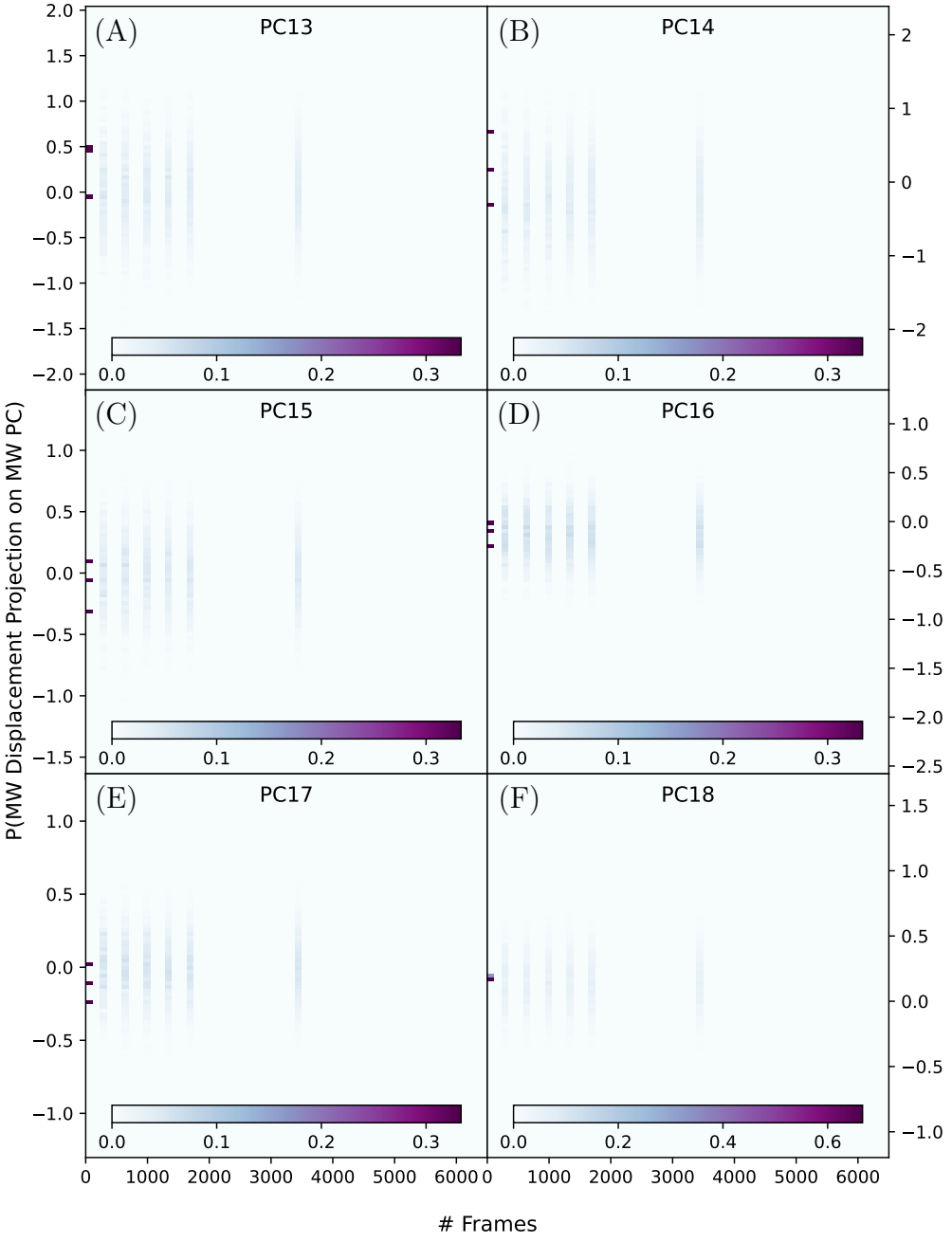


Figure E.3: Convergence of the sampling of PCs in the 90º conformation. The colour bar represents the probability distribution of the projection. Part 3 of 7.

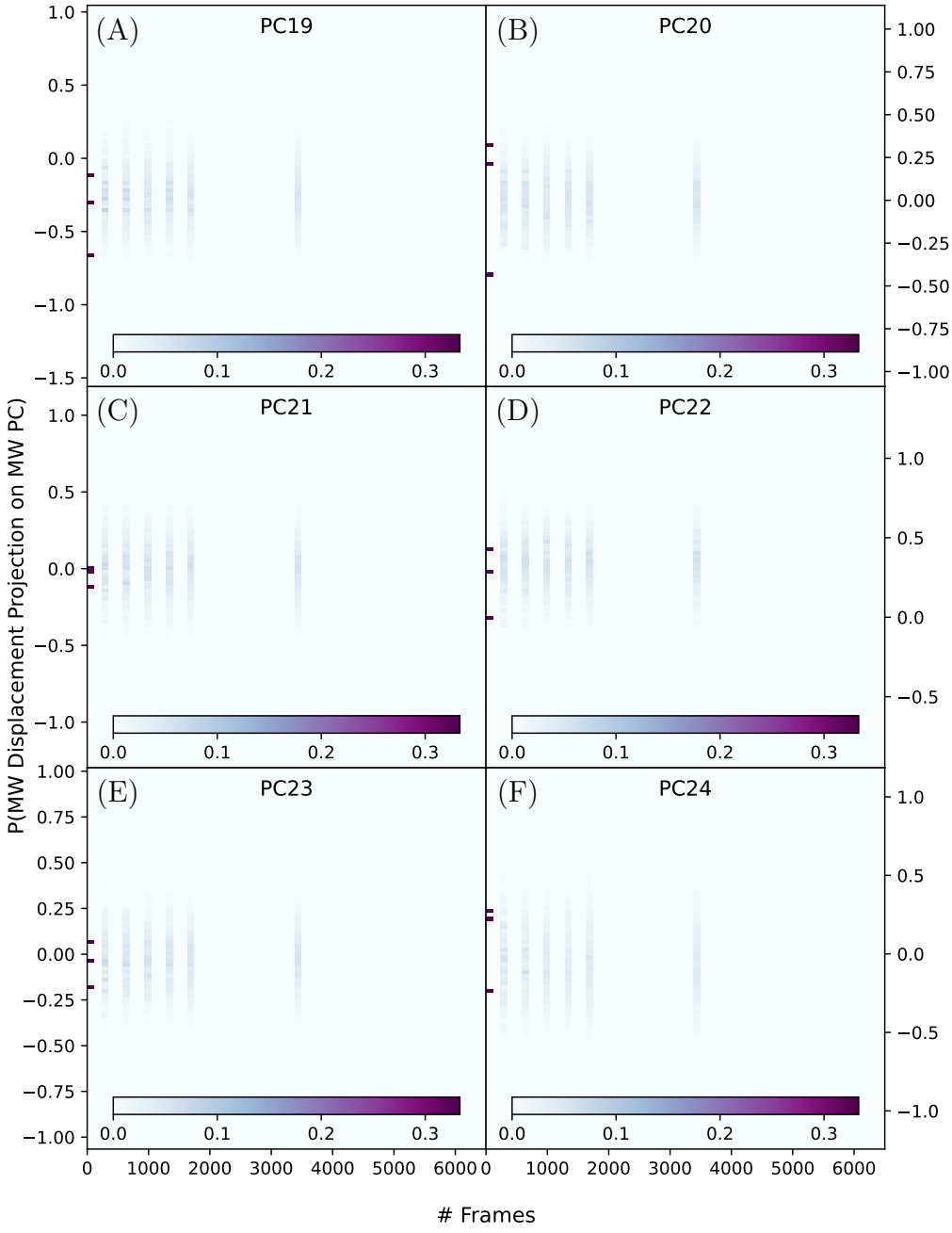


Figure E.4: Convergence of the sampling of PCs in the 90º conformation. The colour bar represents the probability distribution of the projection. Part 4 of 7.

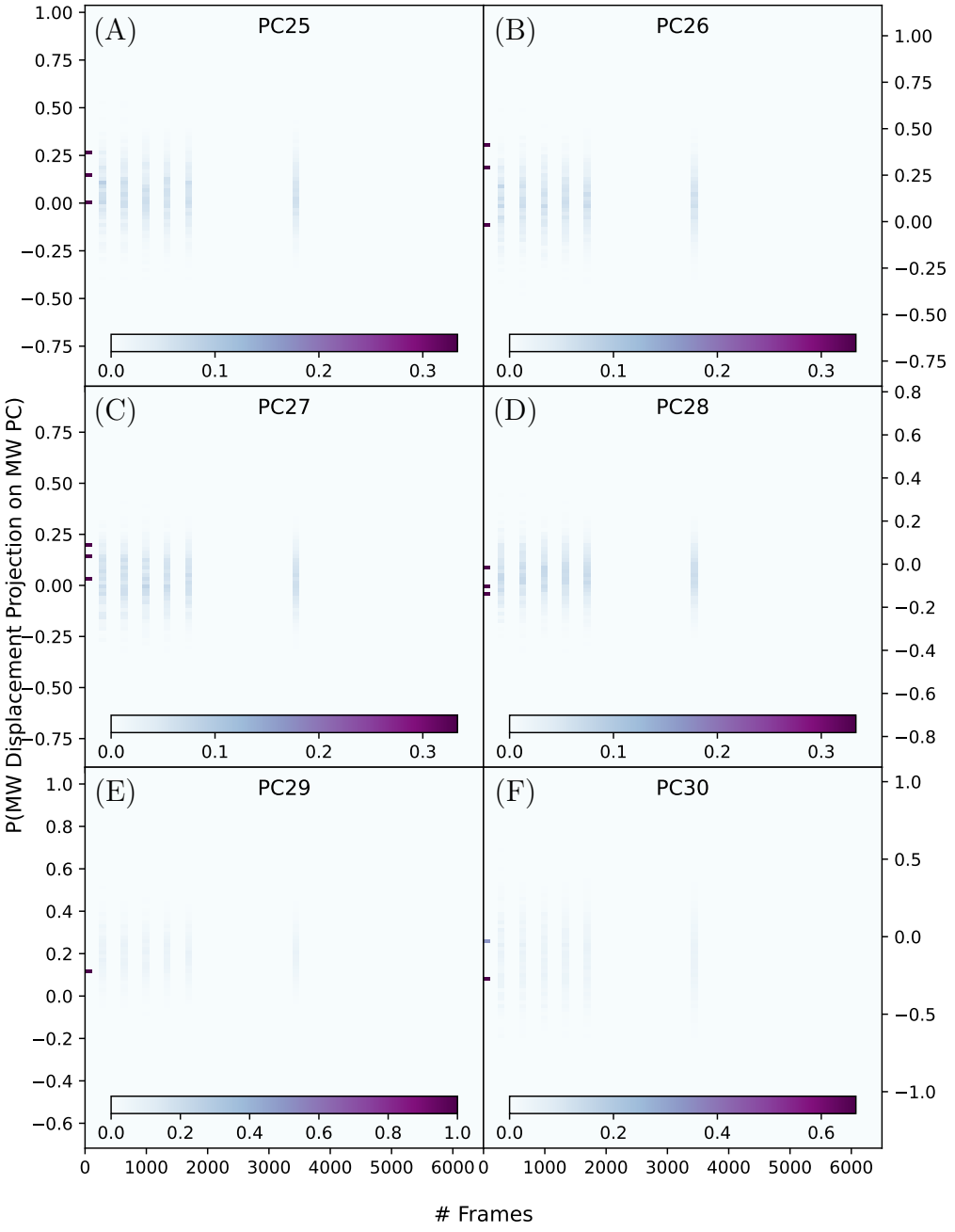


Figure E.5: Convergence of the sampling of PCs in the 90º conformation. The colour bar represents the probability distribution of the projection. Part 5 of 7.

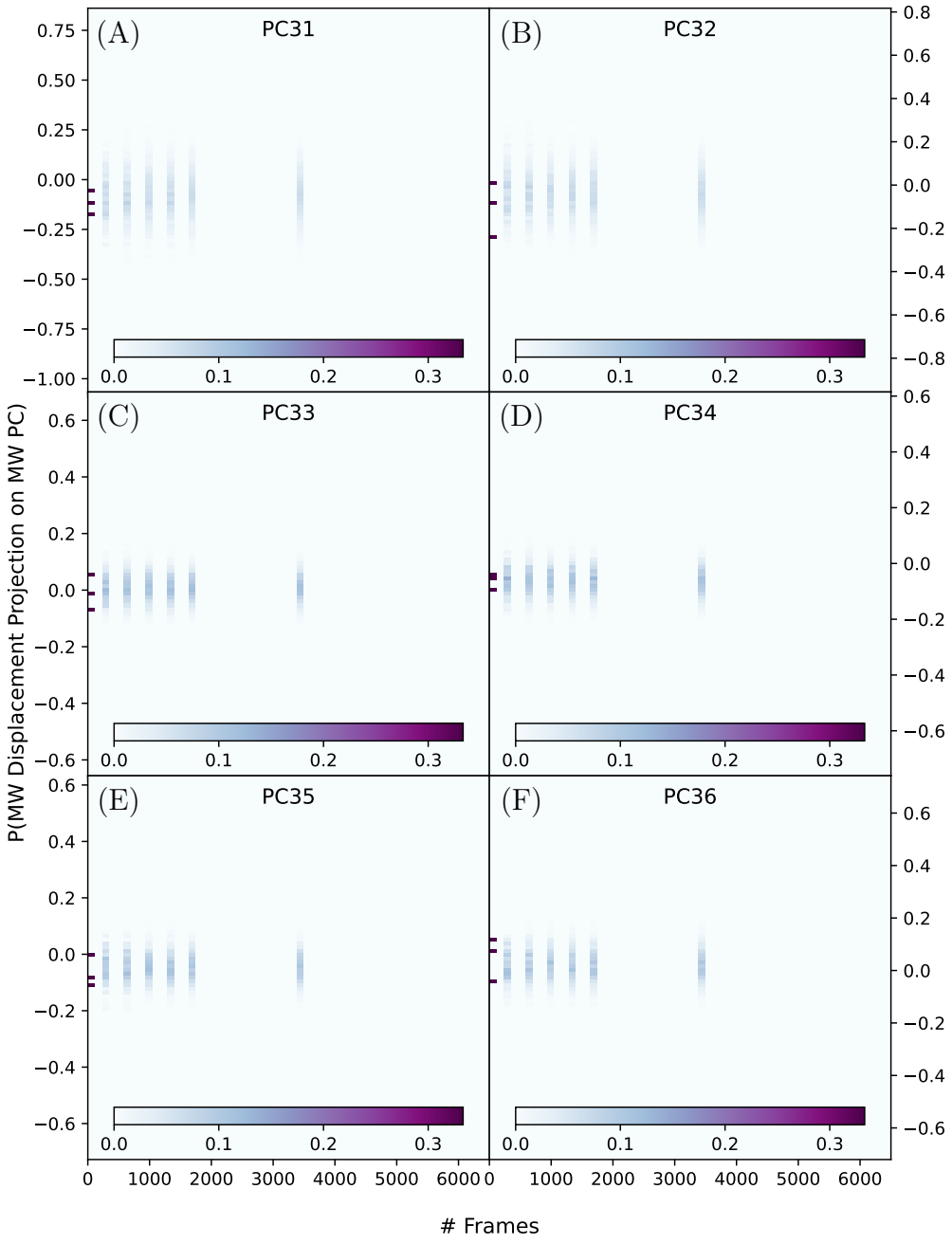


Figure E.6: Convergence of the sampling of PCs in the 90º conformation. The colour bar represents the probability distribution of the projection. Part 6 of 7.

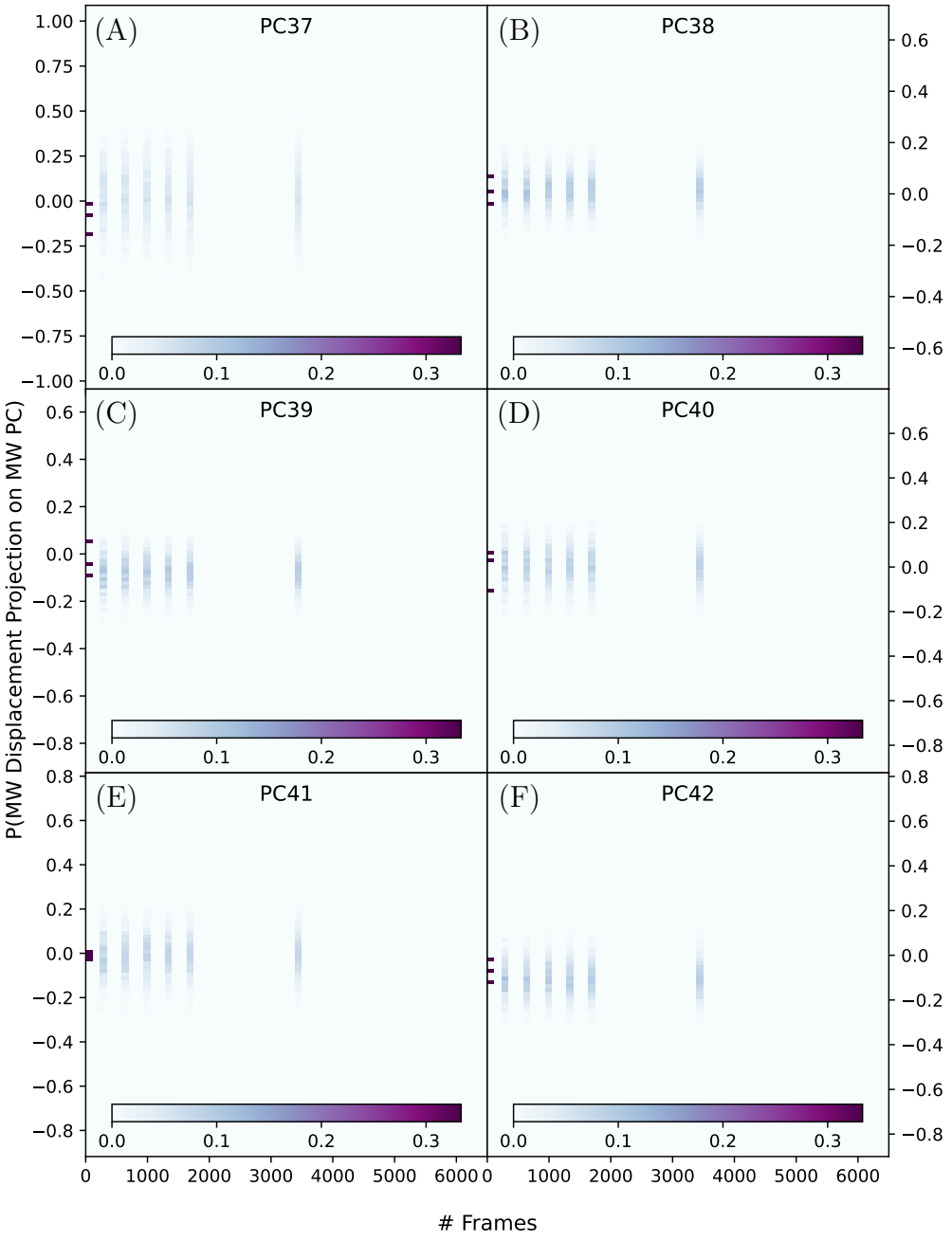


Figure E.7: Convergence of the sampling of PCs in the 90º conformation. The colour bar represents the probability distribution of the projection. Part 7 of 7.

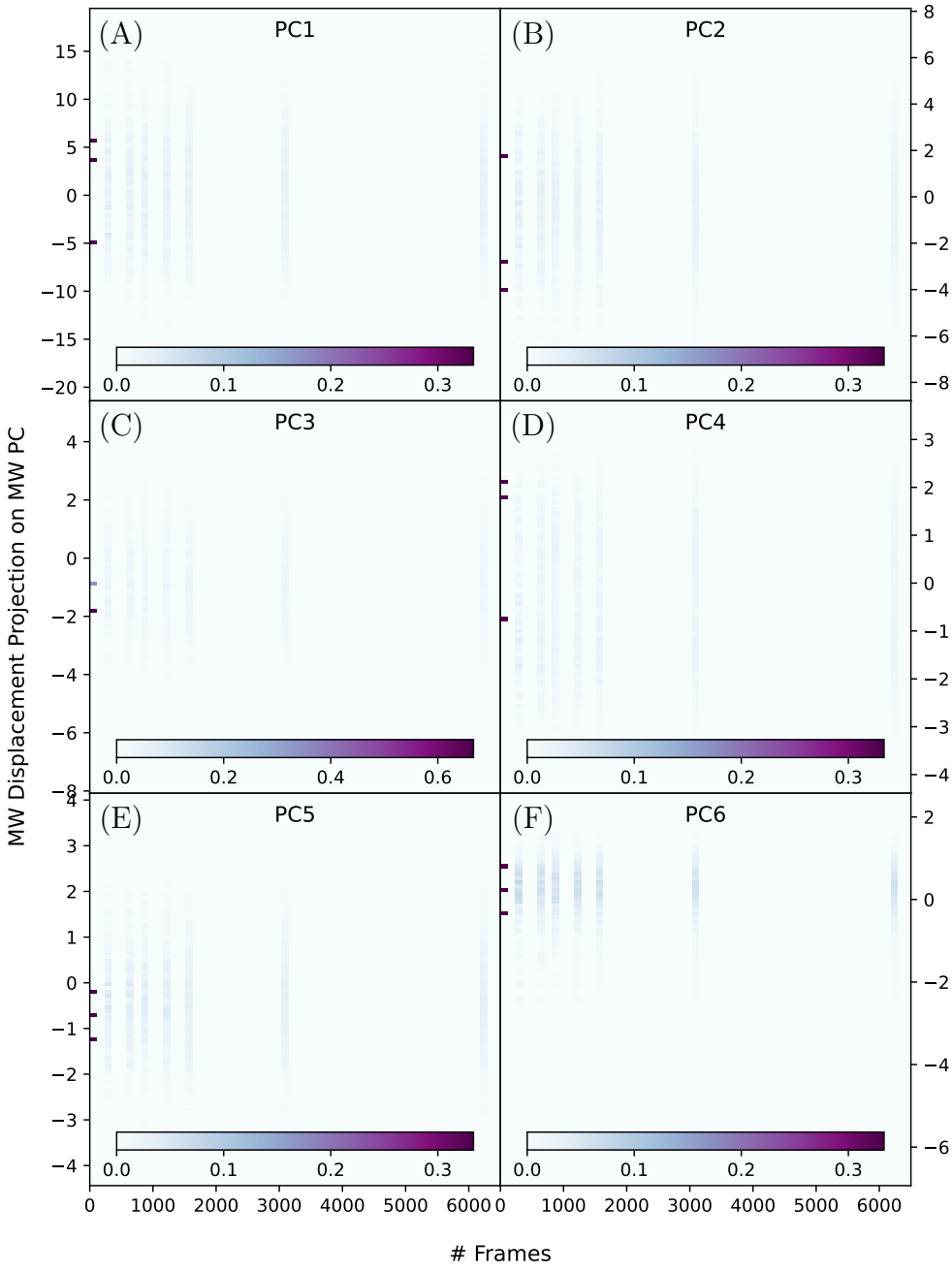


Figure E.8: Convergence of the sampling of PCs in the s-trans conformation. The colour bar represents the probability distribution of the projection. Part 1 of 7.

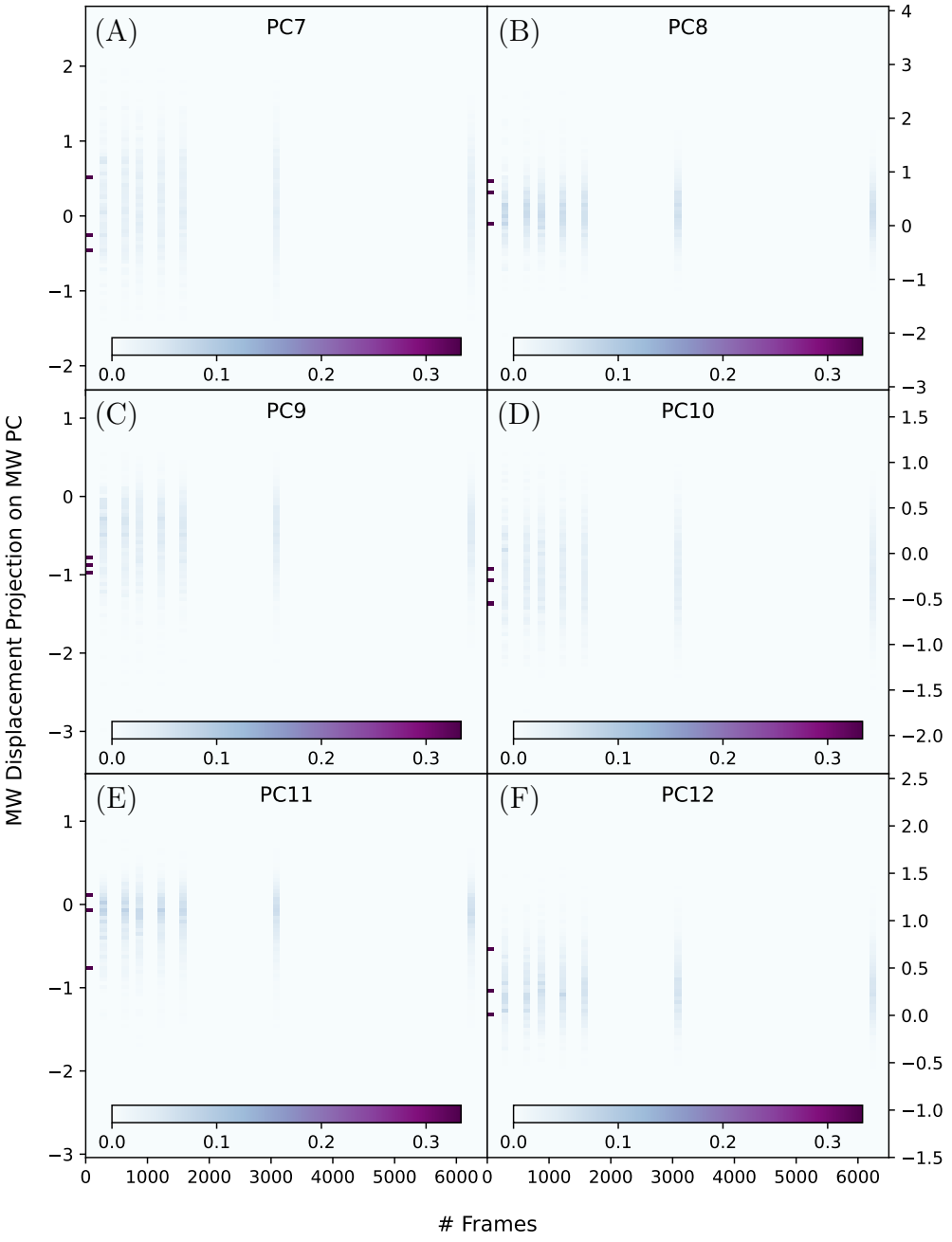


Figure E.9: Convergence of the sampling of PCs in the s-trans conformation. The colour bar represents the probability distribution of the projection. Part 2 of 7.

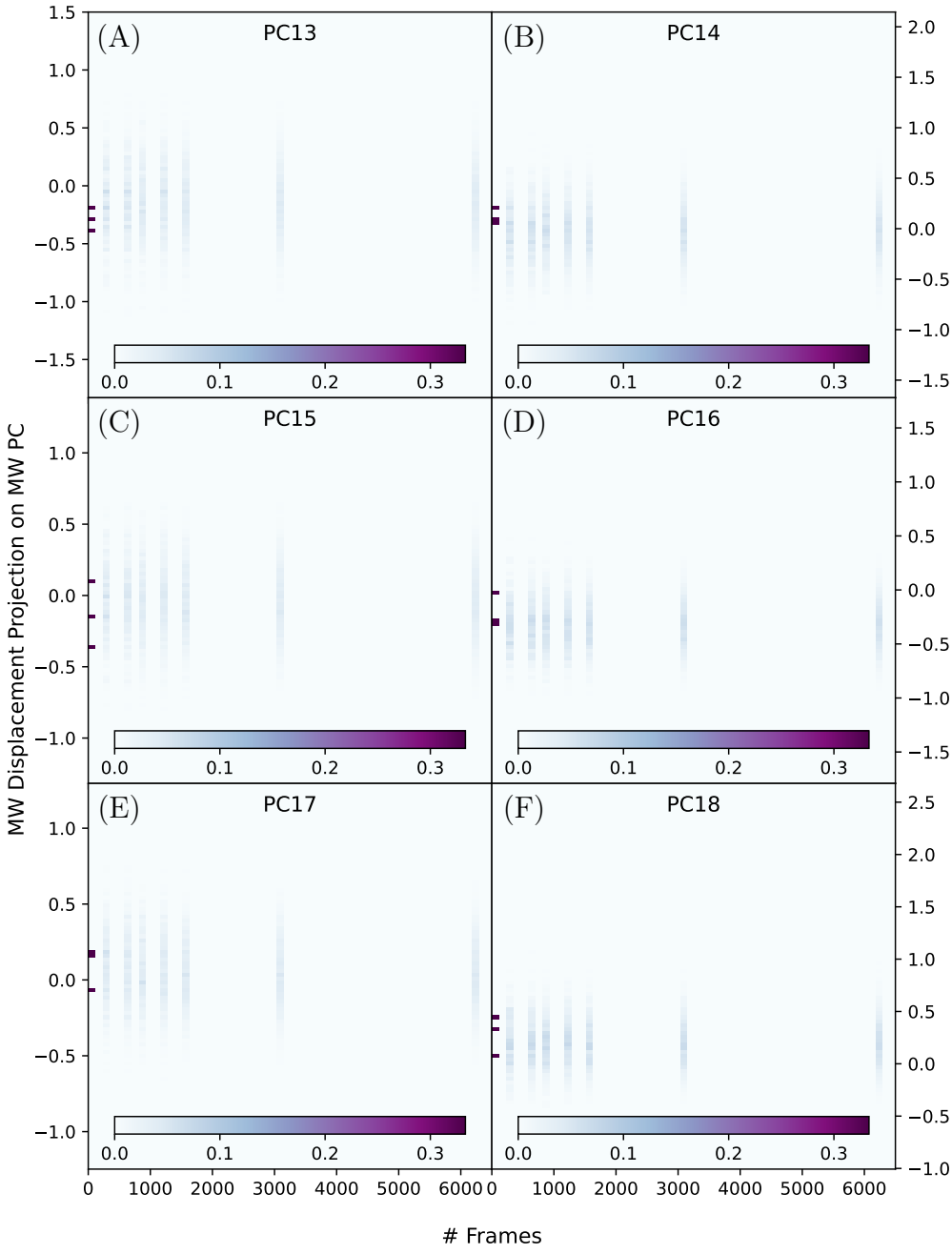


Figure E.10: Convergence of the sampling of PCs in the s-trans conformation. The colour bar represents the probability distribution of the projection. Part 3 of 7.

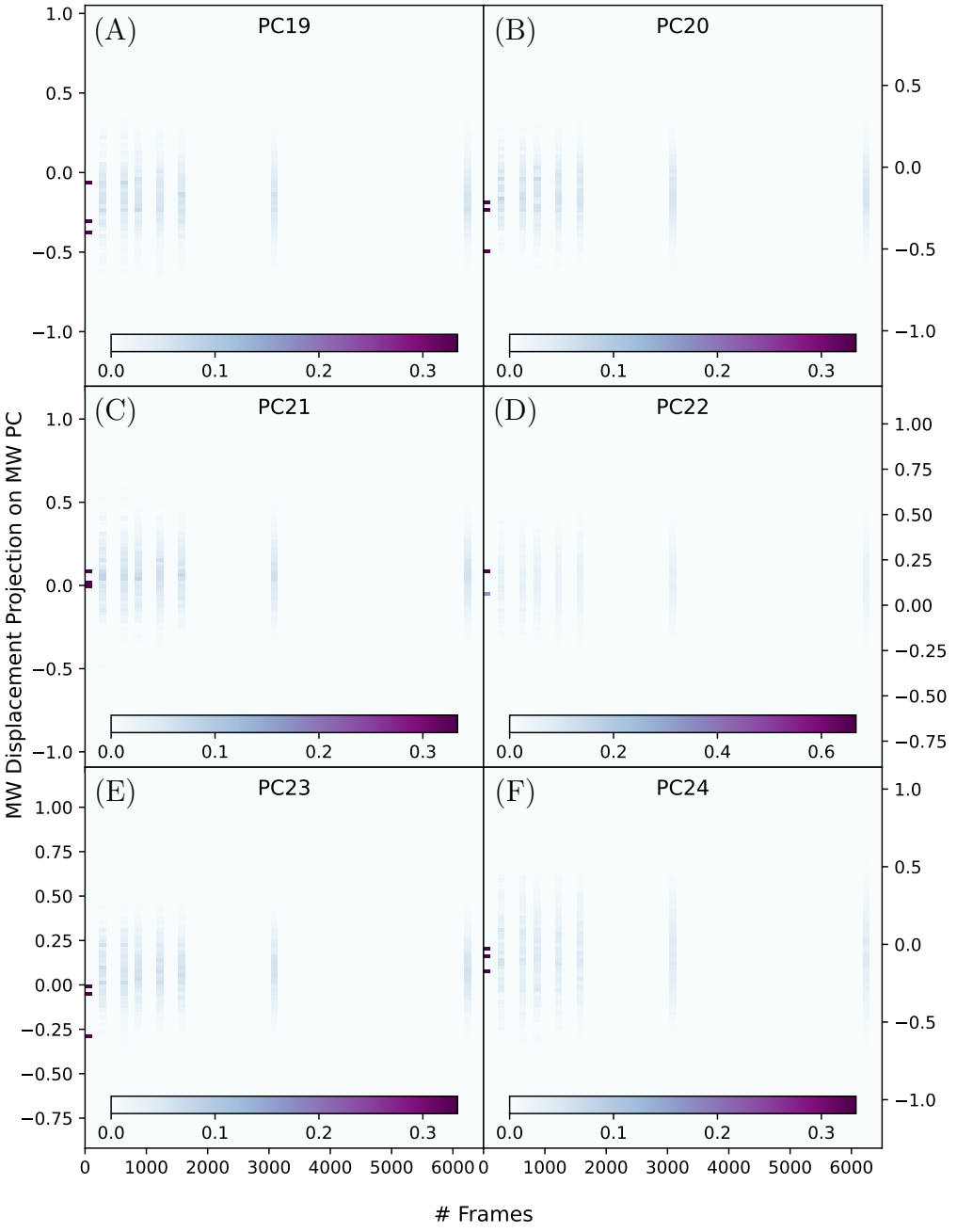


Figure E.11: Convergence of the sampling of PCs in the s-trans conformation. The colour bar represents the probability distribution of the projection. Part 4 of 7.

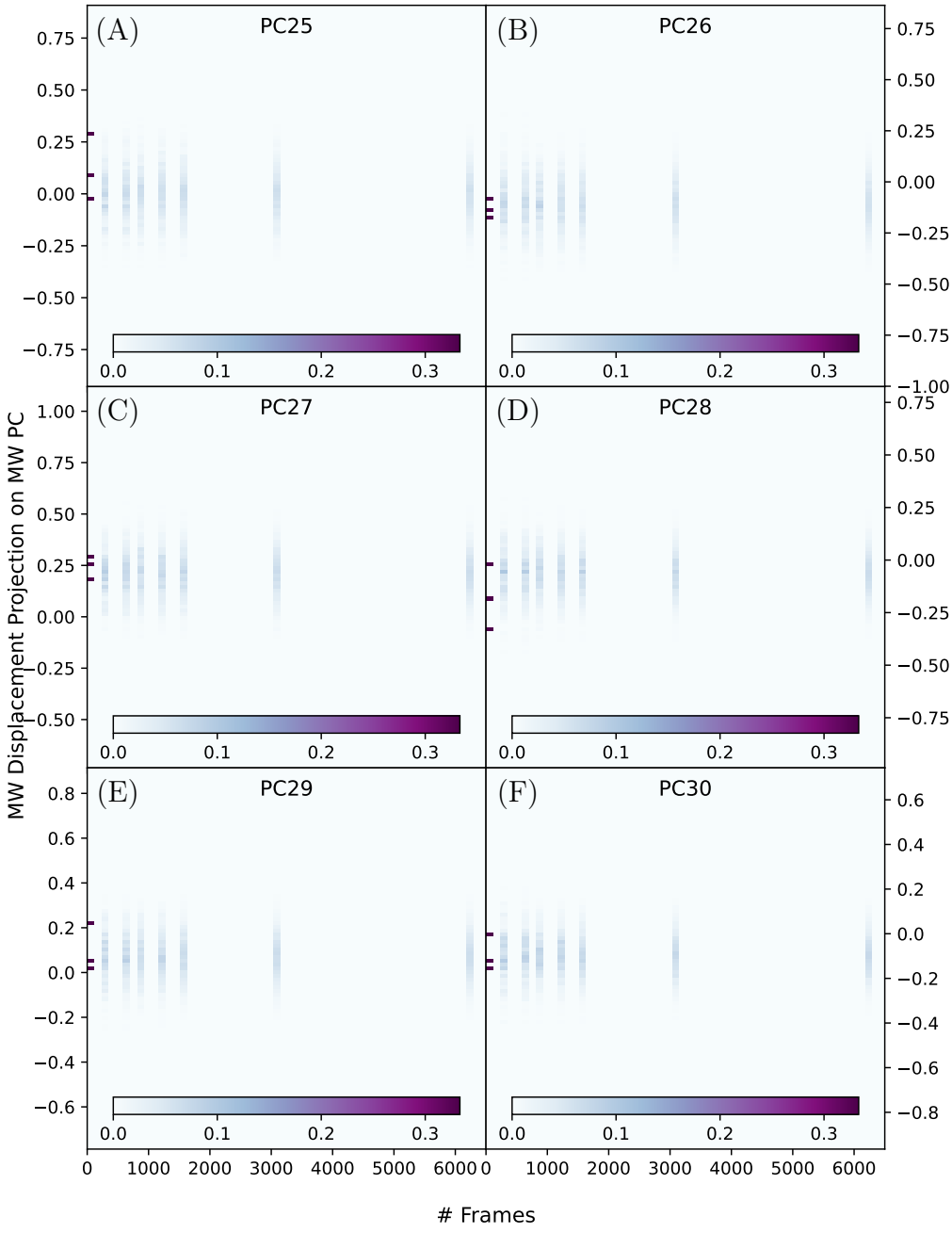


Figure E.12: Convergence of the sampling of PCs in the s-trans conformation. The colour bar represents the probability distribution of the projection. Part 5 of 7.

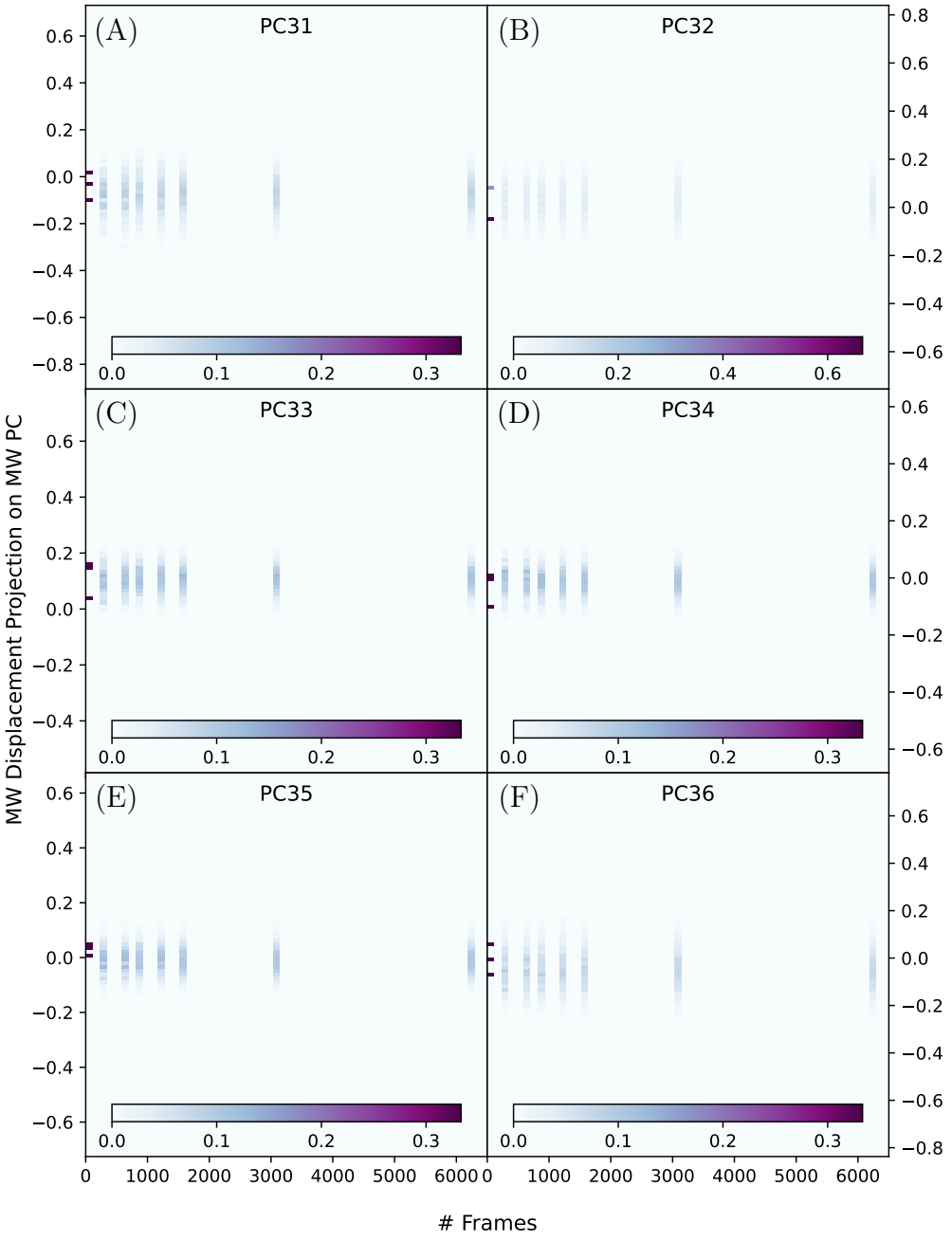


Figure E.13: Convergence of the sampling of PCs in the s-trans conformation. The colour bar represents the probability distribution of the projection. Part 6 of 7.

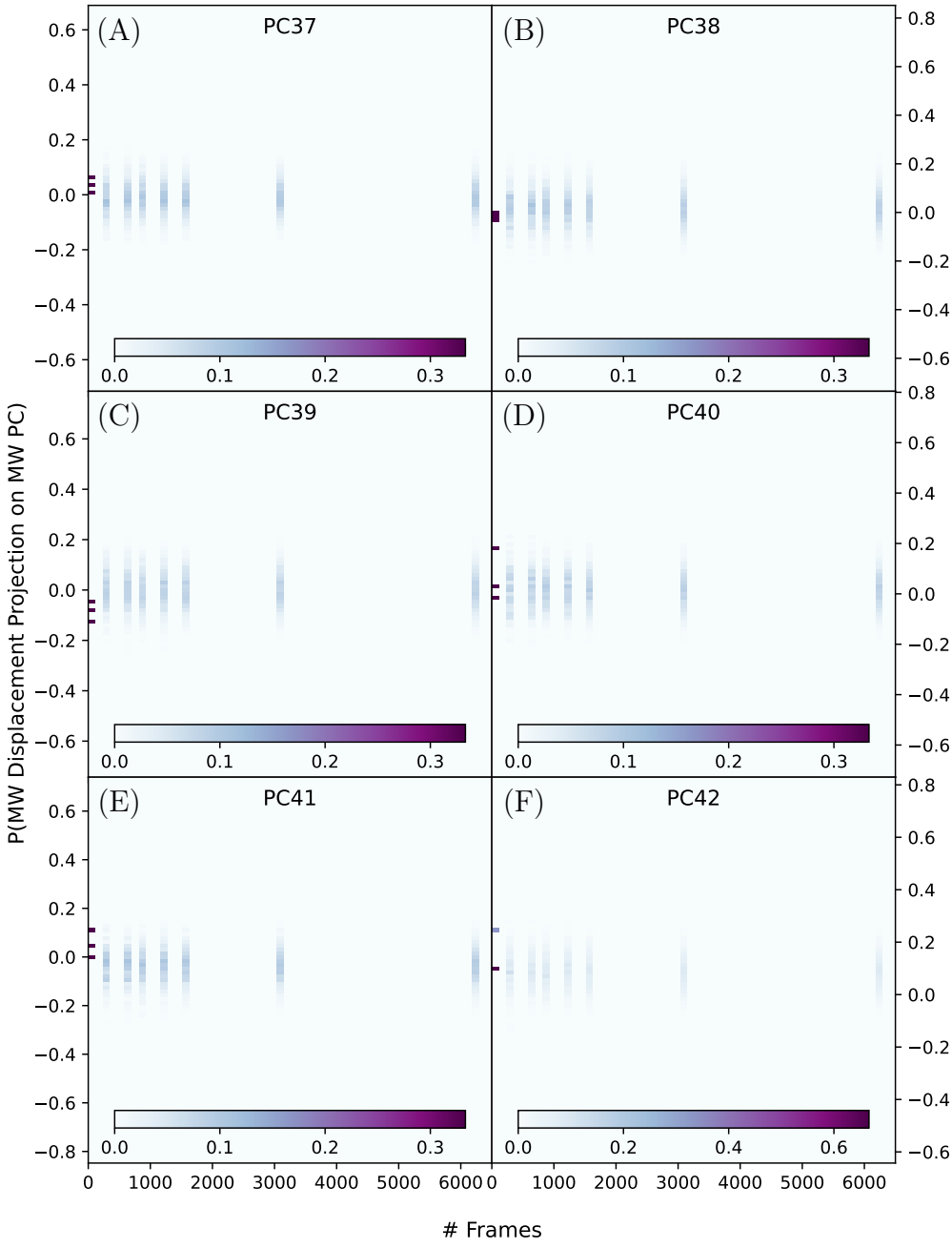


Figure E.14: Convergence of the sampling of PCs in the s-trans conformation. The colour bar represents the probability distribution of the projection. Part 7 of 7.

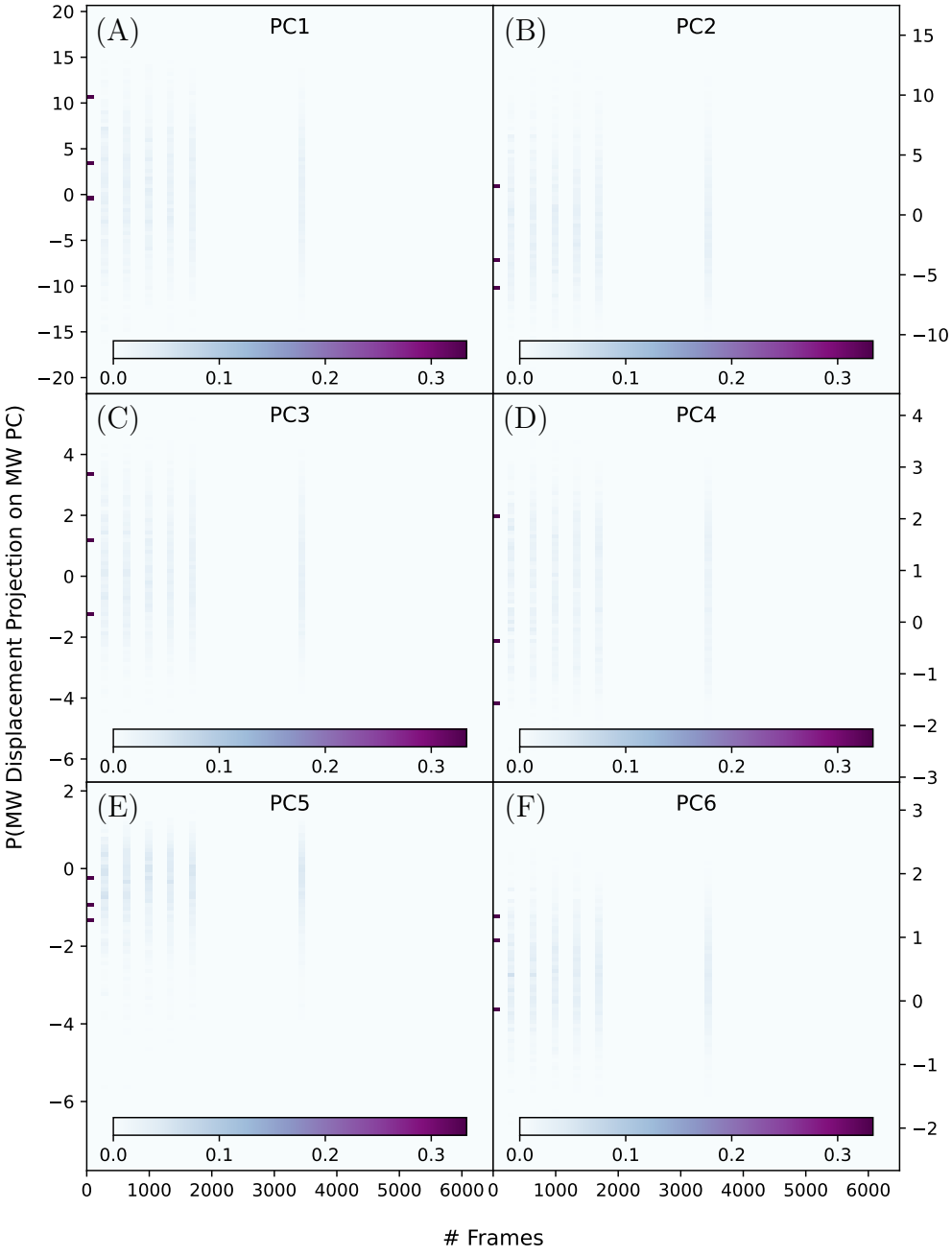


Figure E.15: Convergence of the sampling of PCs in the 270º conformation. The colour bar represents the probability distribution of the projection. Part 1 of 7.

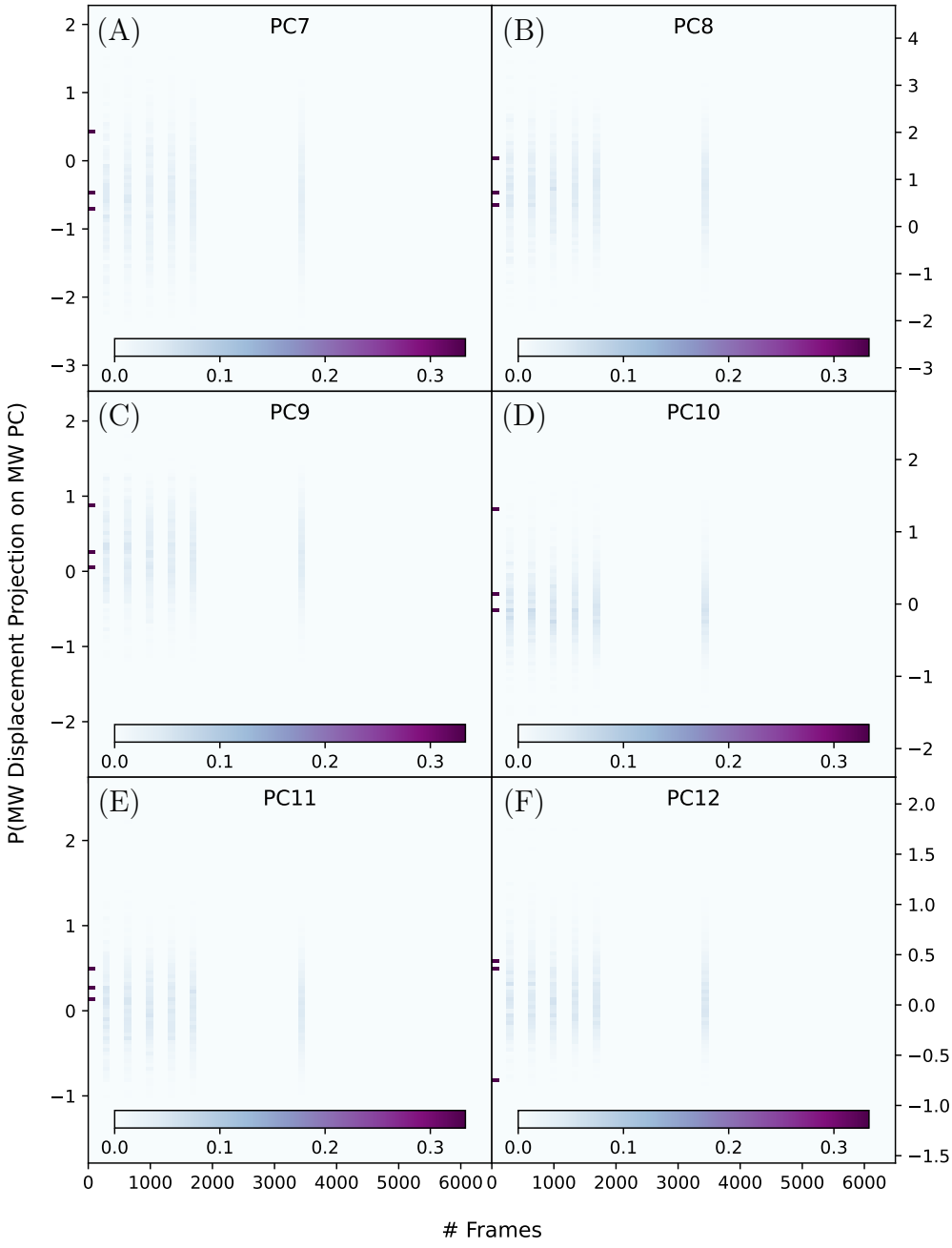


Figure E.16: Convergence of the sampling of PCs in the 270º conformation. The colour bar represents the probability distribution of the projection. Part 2 of 7.

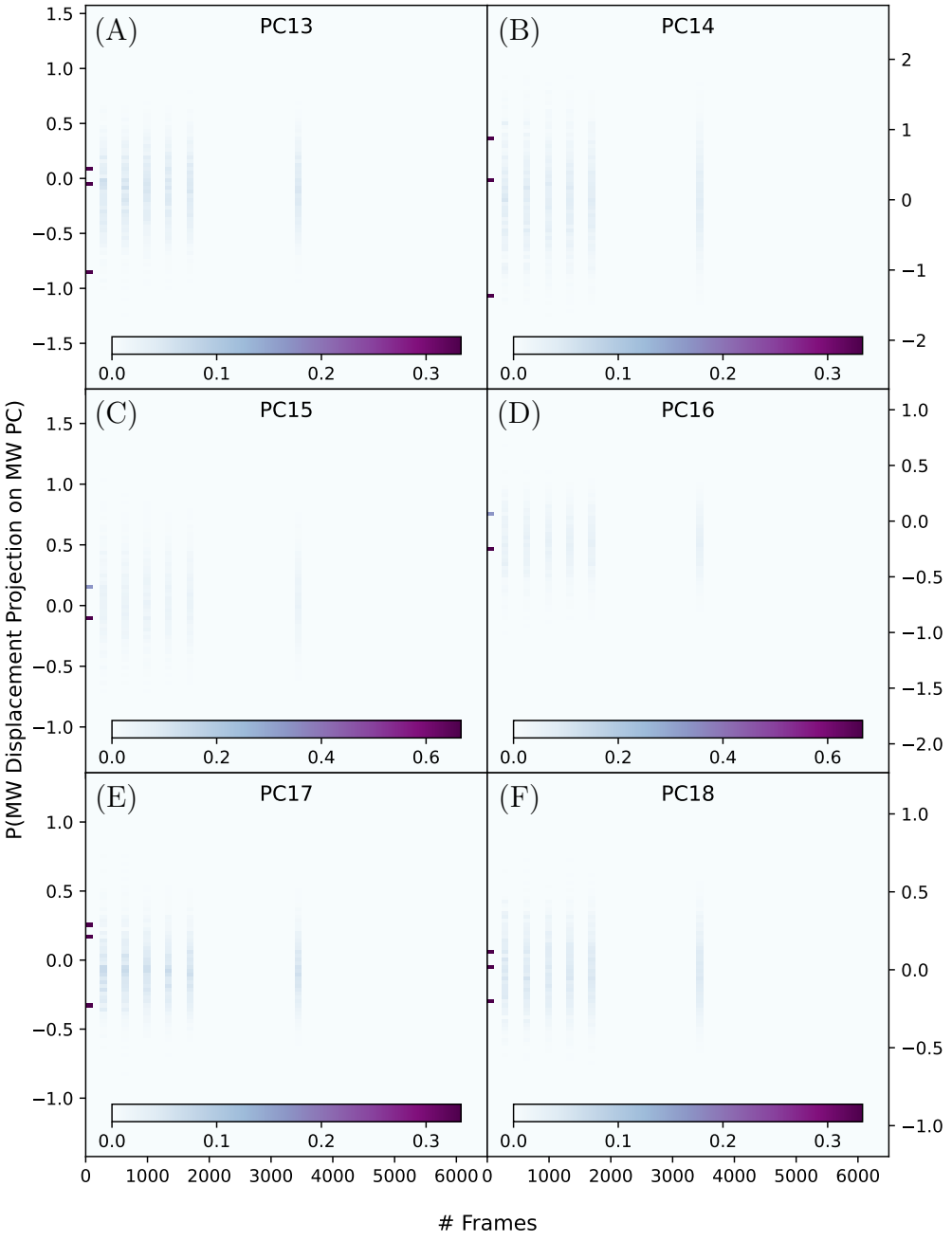


Figure E.17: Convergence of the sampling of PCs in the 270º conformation. The colour bar represents the probability distribution of the projection. Part 3 of 7.

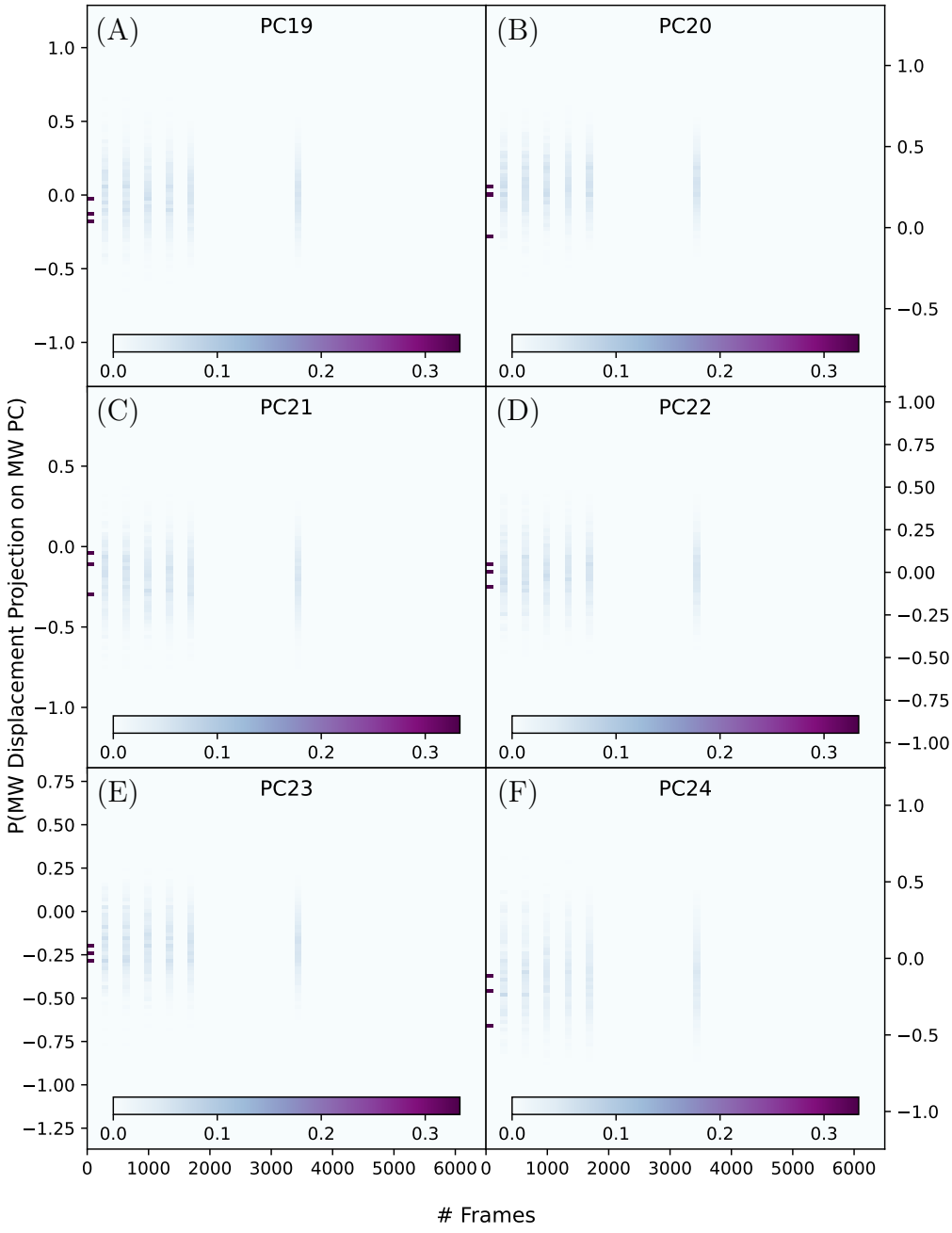


Figure E.18: Convergence of the sampling of PCs in the 270º conformation. The colour bar represents the probability distribution of the projection. Part 4 of 7.

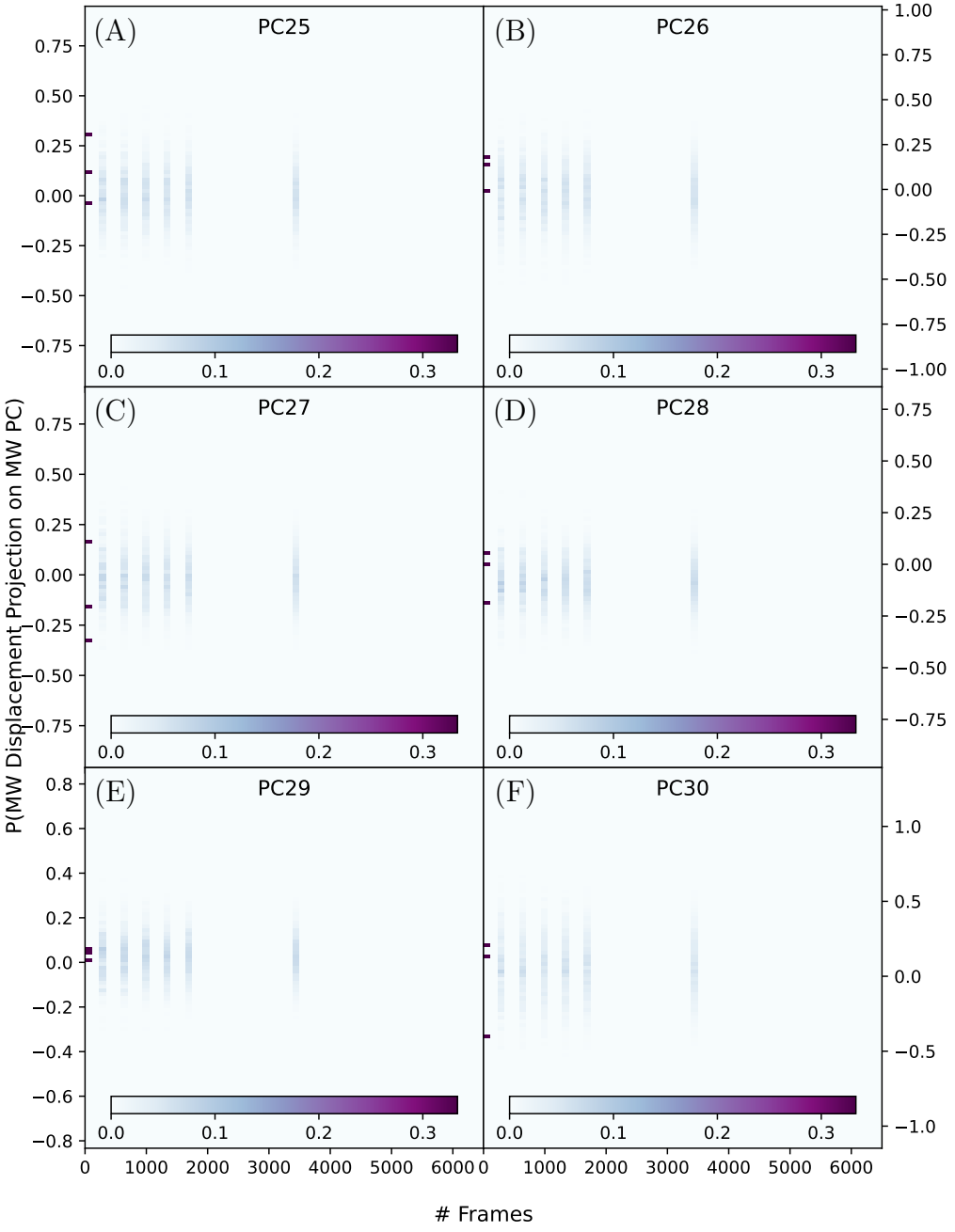


Figure E.19: Convergence of the sampling of PCs in the 270º conformation. The colour bar represents the probability distribution of the projection. Part 5 of 7.

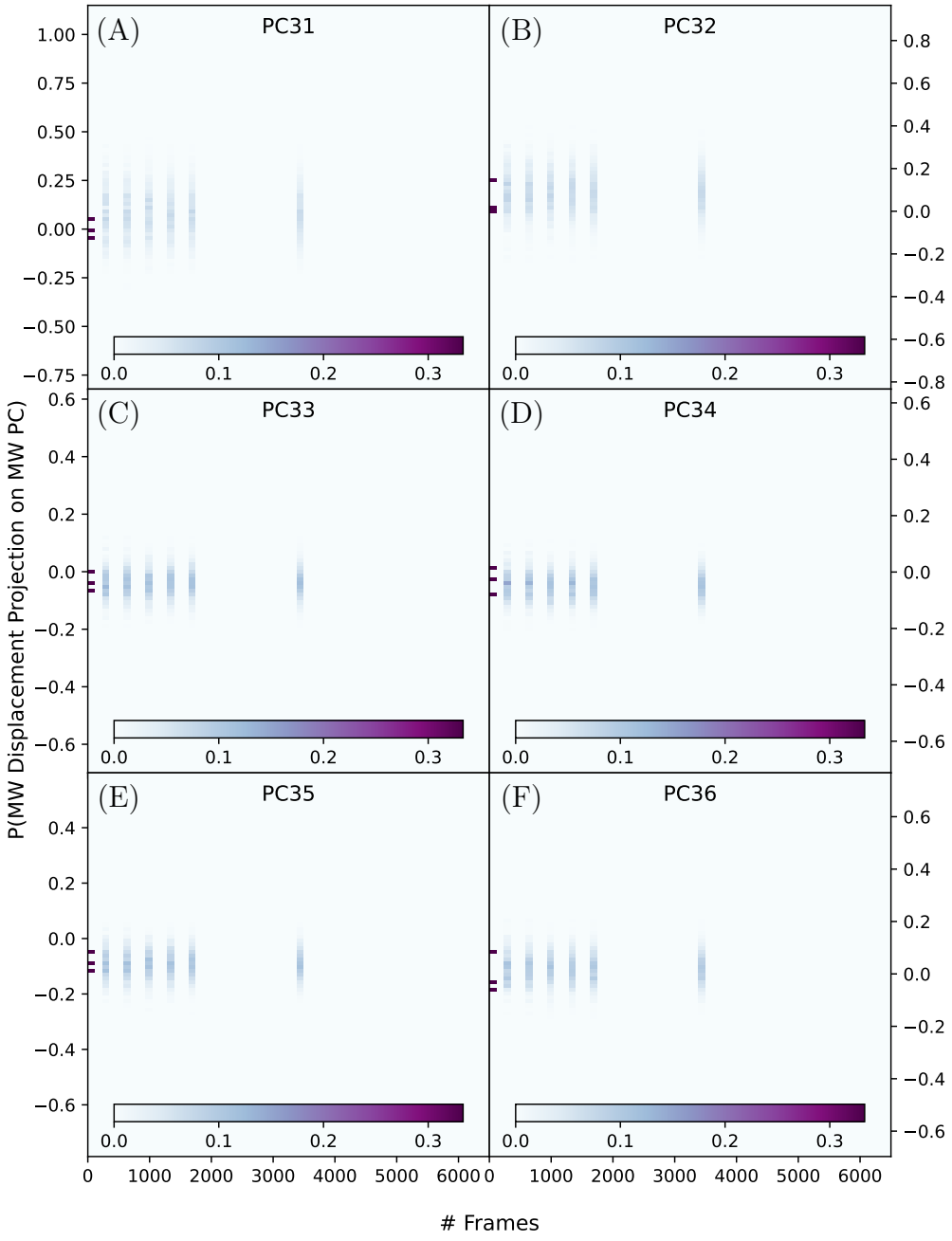


Figure E.20: Convergence of the sampling of PCs in the 270º conformation. The colour bar represents the probability distribution of the projection. Part 6 of 7.

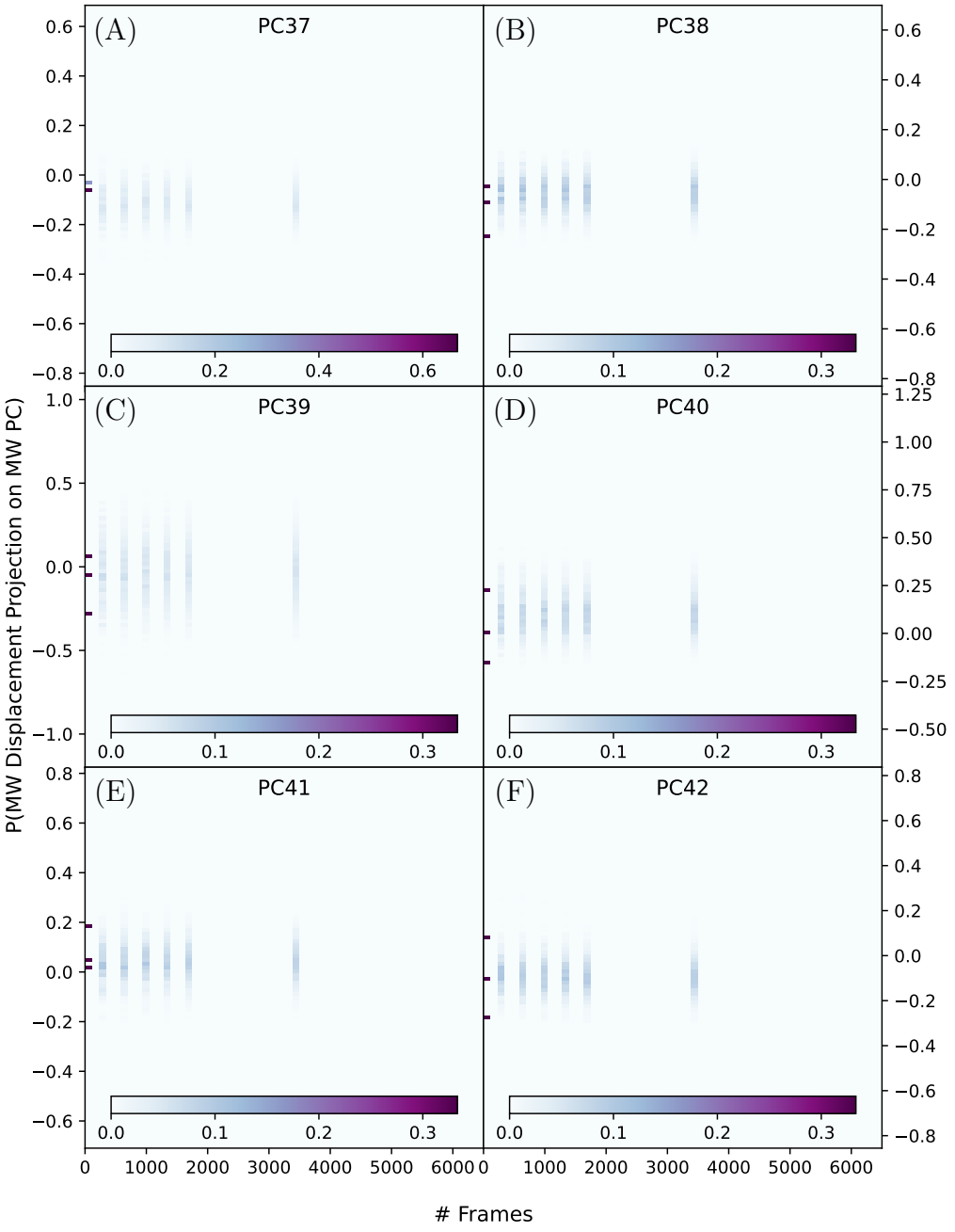


Figure E.21: Convergence of the sampling of PCs in the 270º conformation. The colour bar represents the probability distribution of the projection. Part 7 of 7.

E.2 Convergence of the Electronic Properties

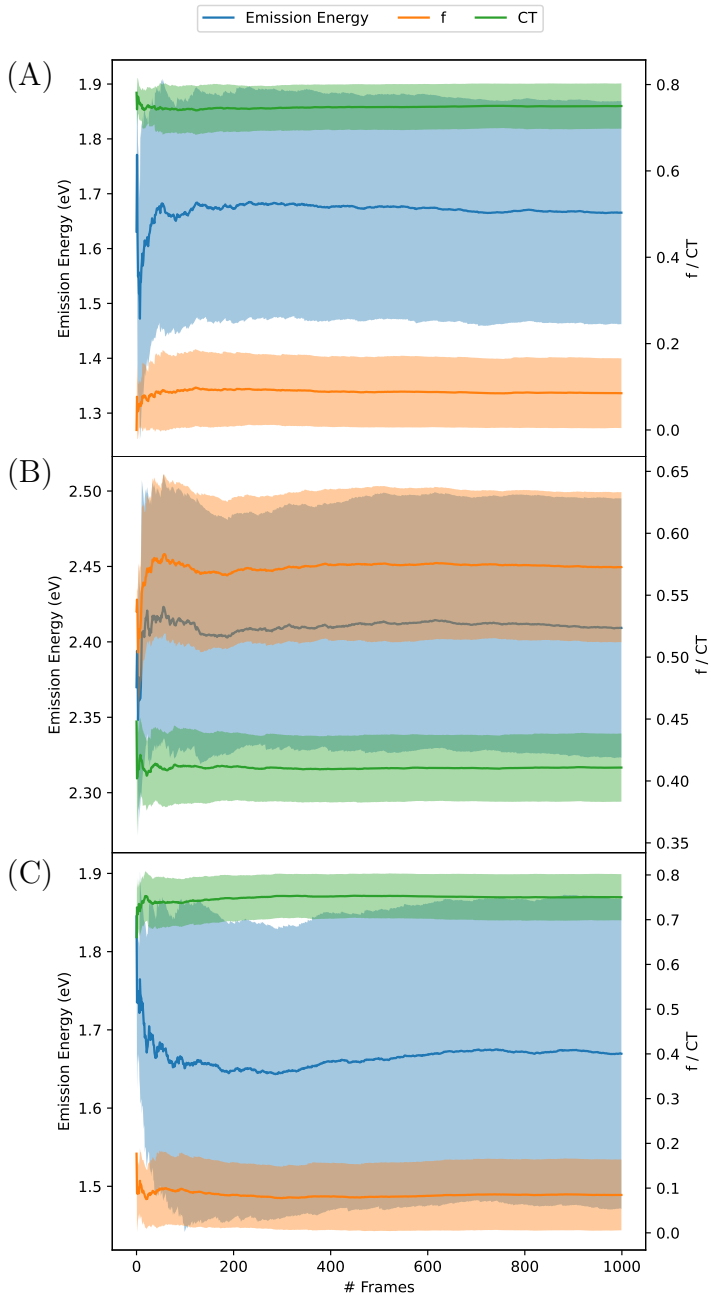


Figure E.22: Convergence of the emission energy, the oscillator strength f and the CT number of (A) the 90º, (B) s-trans and (C) 270º conformations.

E.3 Variances of the Principal Components

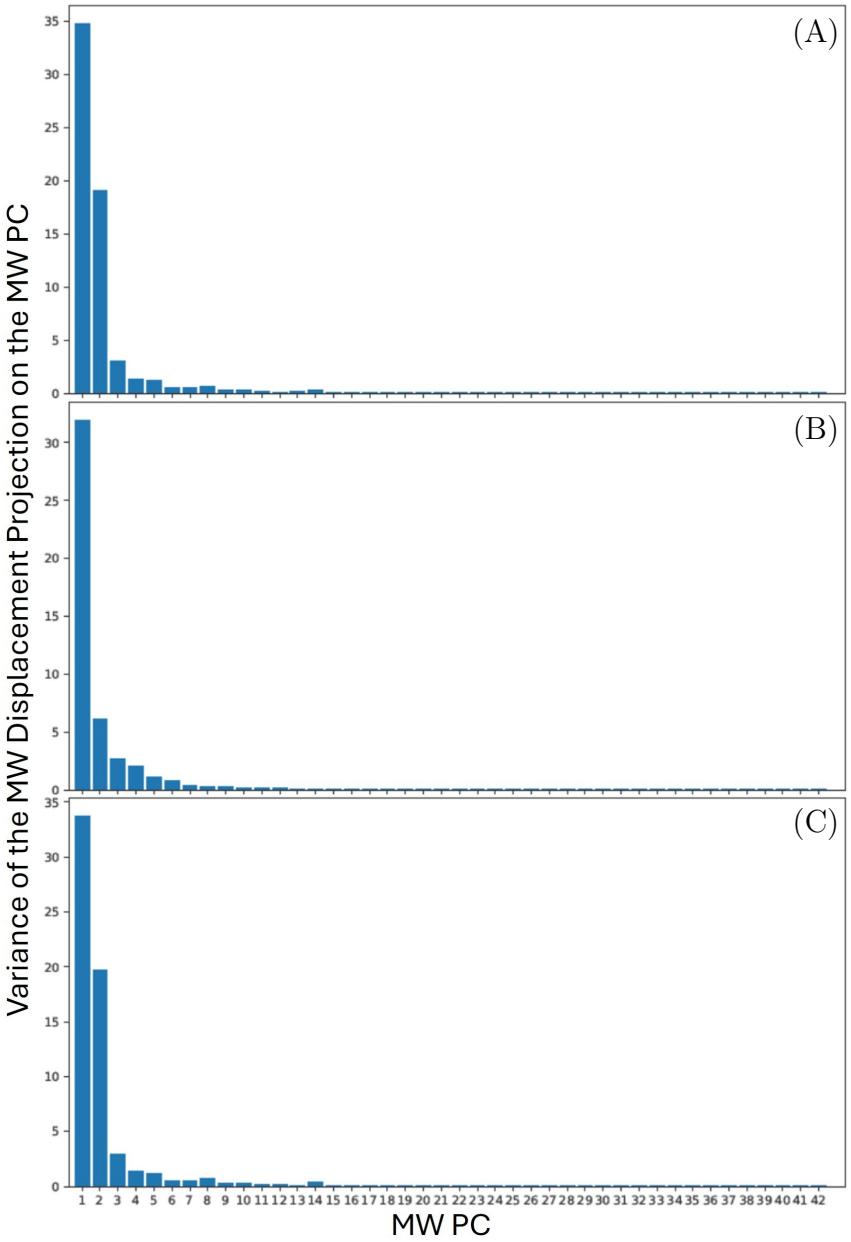


Figure E.23: Variances of the PC of (A) the 90º, (B) s-trans and (C) 270º conformations.

E.4 Frequencies of the Principal Components

	Conformer				Conformer		
PC	90º	s-trans	270º	PC	90º	s-trans	270º
1	14.06	14.72	14.29	22	559.75	544.92	557.44
2	19.11	33.79	18.79	23	590.23	571.93	597.62
3	48	50.89	48.7	24	634.64	630.62	627.66
4	74.78	62.6	74.1	25	684.28	681.43	675.46
5	77.49	85.56	78.8	26	695.51	686.03	697.42
6	116.62	94.19	116.75	27	769.45	750.63	781.82
7	122.74	141.34	122.3	28	812.59	804.92	797.9
8	155.02	159.3	153.92	29	894.32	848.93	895.11
9	171	174.41	175.25	30	930.52	928.81	913.53
10	206.32	193.16	206.41	31	967.03	1008.4	967.8
11	216.14	225.24	217.1	32	1033.58	1070.31	1057.53
12	279.05	280.65	278.44	33	1777.72	1776.6	1772.84
13	303.61	299.64	302.36	34	1714.7	1710.05	1716.29
14	309.44	302.45	313.53	35	1651.46	1661.22	1640.83
15	332.53	342.69	334.85	36	1444.14	1129.8	1454.21
16	393.94	367.64	393.97	37	1148.61	1462.38	1335.61
17	424.65	407.75	428.11	38	1321.32	1198.73	1293.46
18	438.3	402.7	446.43	39	1293.73	1220.89	1140.61
19	472.05	462.69	471.34	40	1172.95	1341.66	1239.28
20	516.33	487.03	518.09	41	1232.12	1278.69	1175.29
21	533.59	520.86	537.55	42	1208.32	1299.96	1194.04

Table E.1: Frequencies, in cm^{-1}, calculated from the eigenvalues of the PC analysis.

Appendix F

Classical and QM/MM Energy Decomposition Analysis of Oxyluciferin/Luciferase in the Electronic Excited State

F.1 Benchmark of Functionals and Basis Sets

Employing the representative structure of the trajectory, not only the DFT functional and the basis set were selected but also the convergence with the number of amino acids in the QM region of the protein was studied, as explained in Section 7.2.1. To determine the most suitable DFT functional/basis set pair for the QM/MM-EDA calculation, different combinations of the CAM-B3LYP and the M062X DFT functionals and 6-31G(2d,p), 6-31+G(2d,p), 6-311+G(2d,p) and 6-31G(2df,p) basis sets with and without empirical dispersion described by the D3BJ version of Grimme's dispersion with Becke-Johnson damping (CAM-B3LYP) or by the D3 version of Grimme's dispersion with the original D3 damping function (M062X) due to implementation constrains, have been studied. As shown in Figure F.1(A-D), CAM-B3LYP, which is not parametrized to properly describe dispersion, presents very consistent results. However, when empirical dispersion is included (Figure F.1(D)), the polarization term gains participation to the total interaction energy due to a better description of dispersion (Figure F.2(D)). In the case of the M062X functional (Figure F.1(E-G)), all the studied basis set present very similar contributions to the total interaction energy of the EDA components. However, the inclusion of diffuse functions results in an underestimation of the dispersion component,

as shown in Figure F.2(E-F).

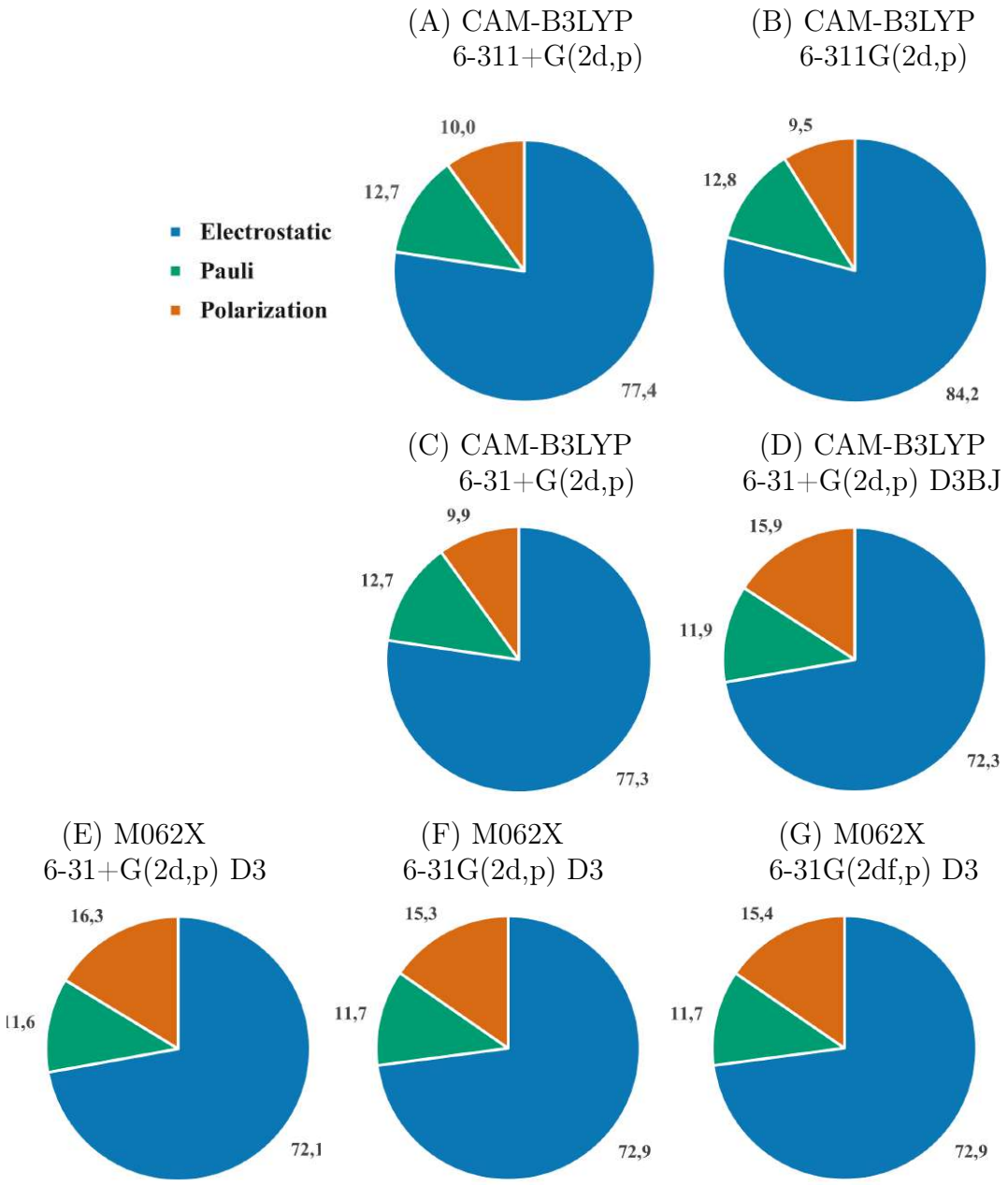


Figure F.1: Percentages of contribution to the total interaction energy of the EDA components relative to the sum of the absolute values of the three represented components calculated on the representative geometry.

F.2 Number of Amino Acids in the QM Region

For the convergence of the EDA with the number of amino acids, the twenty amino acids with the largest binding Gibbs free energy have been included in the QM region of the TD-DFT calculation one by one in decreasing order of contribution according to the 1A-MM-GBSA analysis. These calculations were performed using the M062X functional and the 6-31+G(2d,p), 6-31G(2d,p) and 6-31G(2df,p) basis sets, together with the D3 version of Grimme's dispersion with the original D3 damping function. Upon the inclusion of diffuse functions in the basis set, any of the EDA components seem to be near convergence when the first fourteen amino acids are included in the QM region, as shown in Figure F.3(A). Moreover, the dispersion term becomes divergent reaching to an unrealistic positive value even though it is an attractive term. Thus, including diffuse functions in the basis set leads to a completely wrong description of the interaction energy. When considering a different amount of polarization functions on the heavy atoms (see Figure F.3(B-F)), any difference but the 2.5 times larger computational time can be observed, and thus it is more suitable to use the 6-31G(2d,p) basis set in order to perform more efficient calculations.

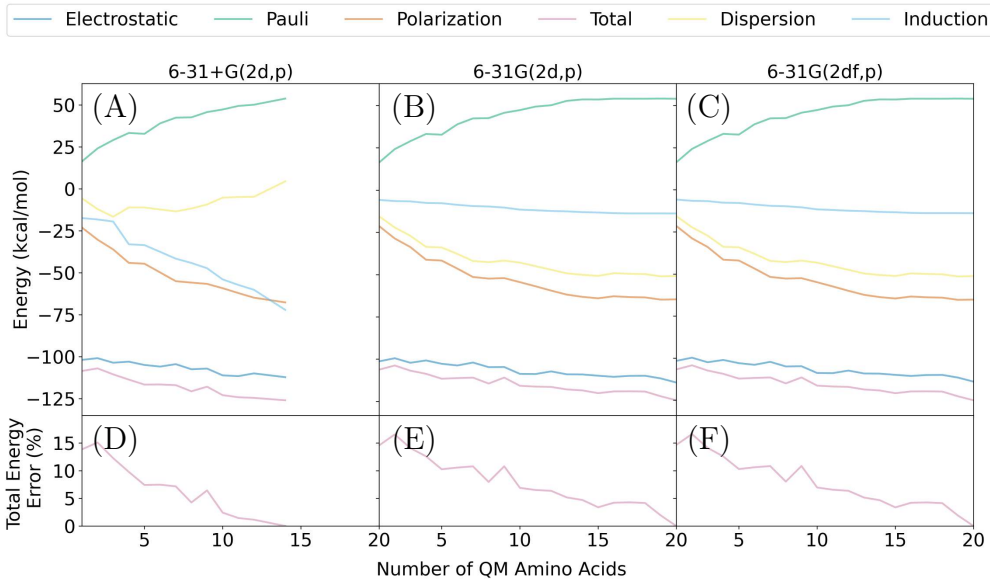


Figure F.3: (A-C) Convergence of the interaction energy and the EDA components and (D-F) error of the total interaction energy with respect to the calculation with twenty amino acids in the QM region as the number of amino acids in the QM region increases with the M062X functional and (A,D) 6-31+G(2d,p), (B,E) 6-31G(2d,p) and (C,F) 6-31G(2df,p) basis sets.

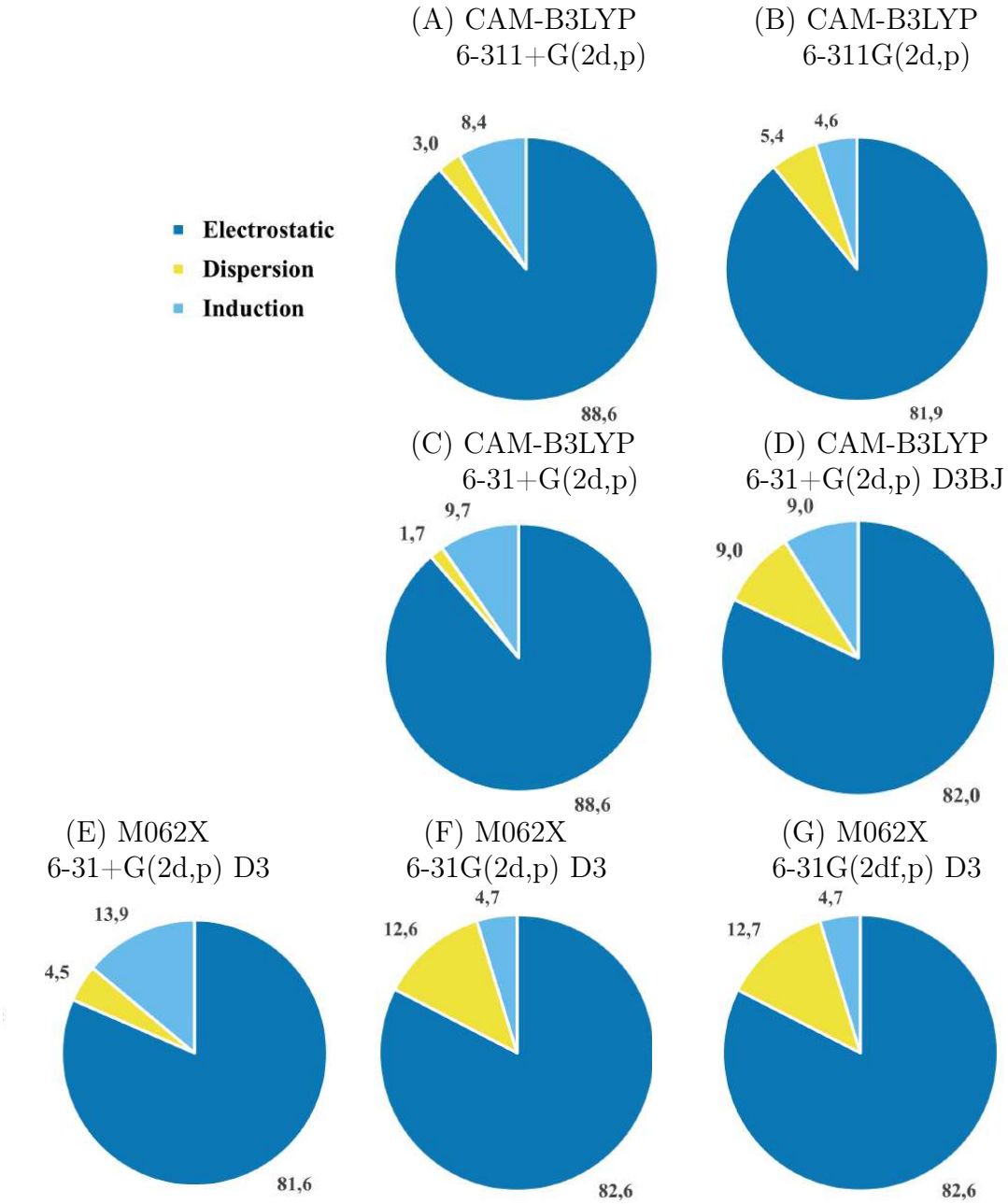


Figure F.2: Percentages of contribution to the attractive terms of the interaction energy of the EDA components relative to the three represented attractive components calculated on the representative geometry.

F.3 Charge Transfer along the MD trajectories

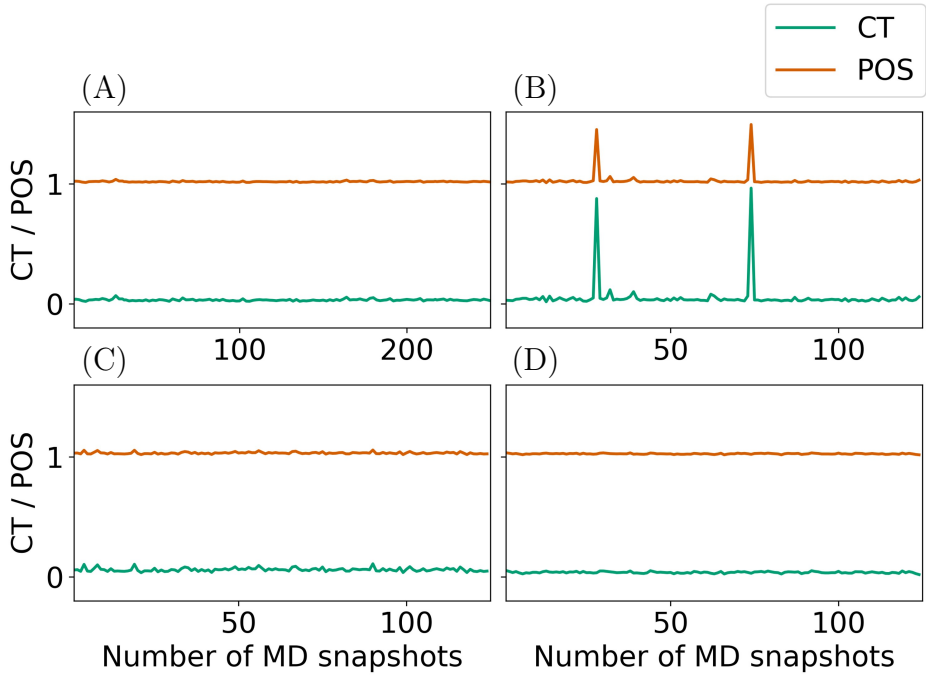


Figure F.4: Number of transferred electrons between the two monomers (green) and averaged position of the orbital before and after the electronic transition (orange) along the MD trajectory of the (A) phenolate-keto, (B) phenol-keto, (C) phenol-enolate and (D) phenolate-enolate OLU.

F.4 Convergence with the Number of Frames of Phenol-keto, Phenol-enolate and Phenolate-enolate Oxyluciferins

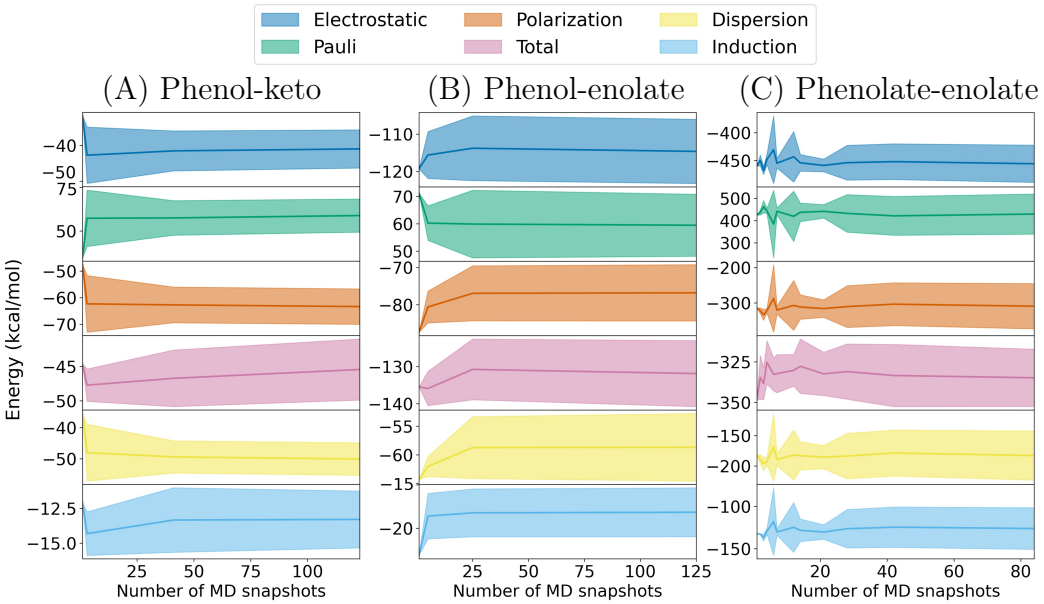


Figure F.5: Convergence of the interaction energy and each of the EDA components with respect to the number of sampled geometries along the MD trajectory of the (A) phenol-keto, (B) phenol-enolate and (C) phenolate-enolate OLU.

Appendix G

Structure of the Phenolate-Keto Oxyluciferin Emitting Species inside the SER286ASN Mutated Luciferase

G.1 RMSD of the Protein

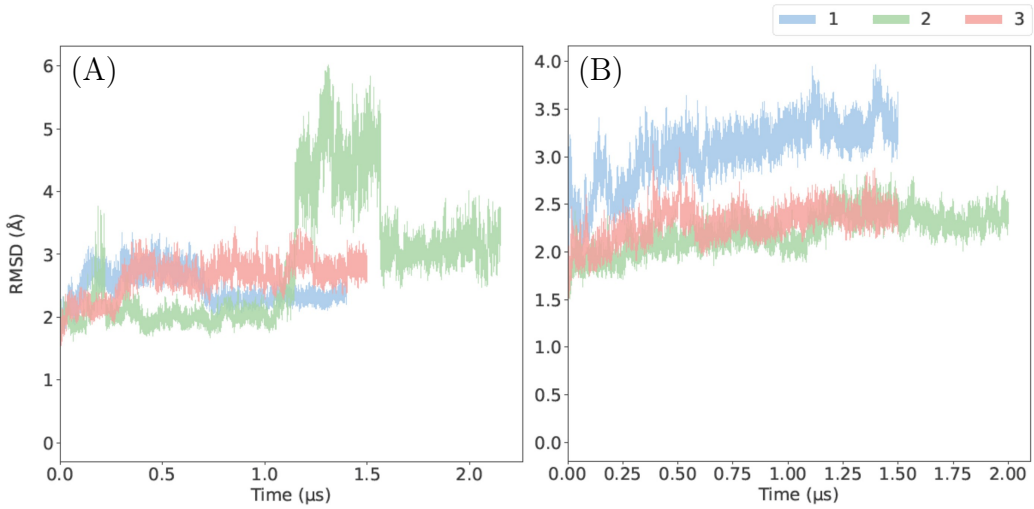


Figure G.1: RMSD of the (A) WT and (B) SER286ASN luciferases for each of the replicas of the MD trajectories with the phenolate-keto OLU in the S_1 state.

G.2 Increased Temperature Simulation

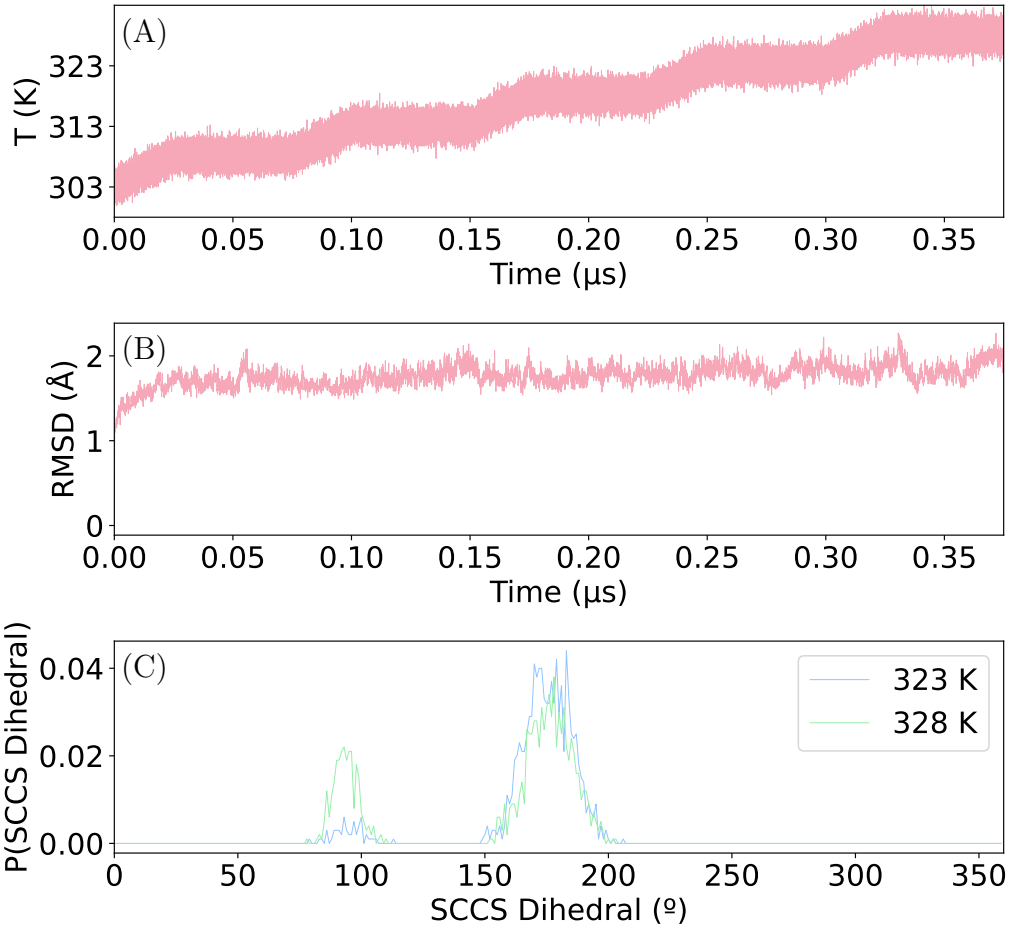


Figure G.2: (A) Temperature ramp and (B) RMSD of the SER286ASN luciferases during the simulation. (C) Probability distributions of the SCCS dihedral angles of the phenolate-keto OLU at two different temperatures.

G.3 SCCS Angle of the Oxyluciferin

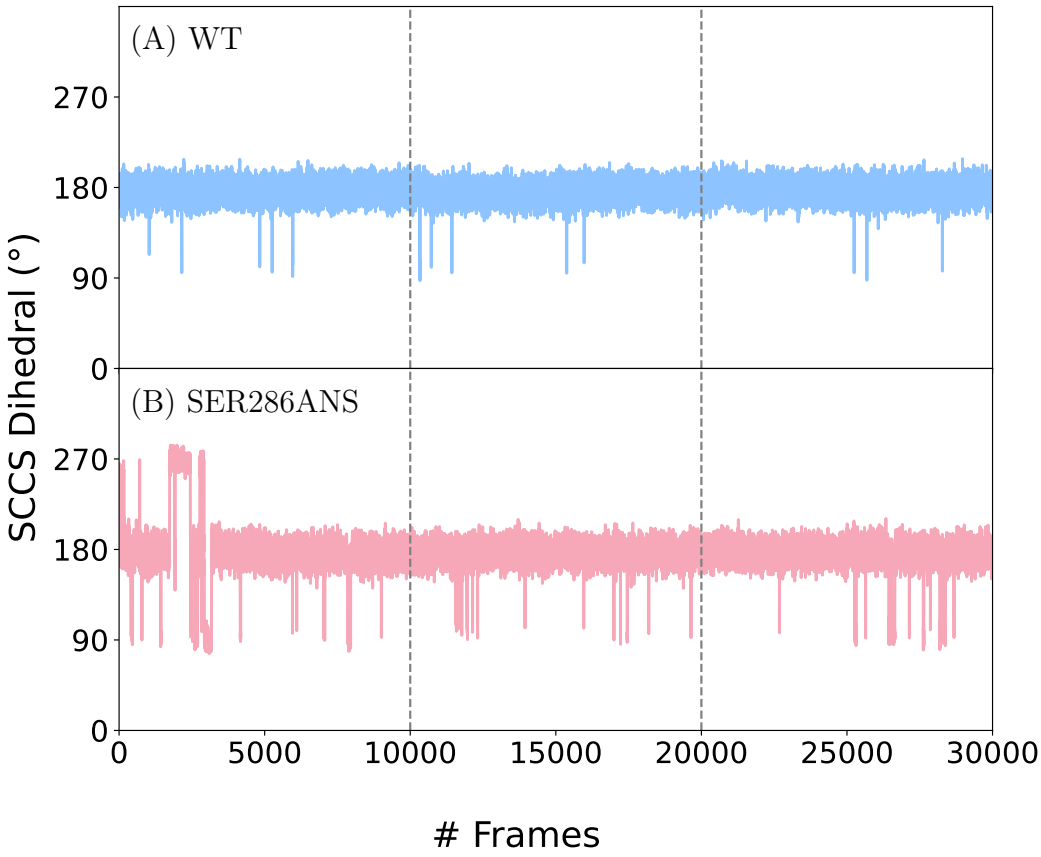


Figure G.3: SCCS angle of the phenolate-keto OLU in the S_1 state along time inside (A) the WT and (B) the SER286ASN luciferases. Vertical dashed lines represent the separation into the three different replicas.