

Department of
Earth and Environmental Sciences

PhD program in Chemical, Geological and Environmental Sciences
Cycle XXXVII

Curriculum: Chemical

Computational study on the bioluminescence of the NanoLuc-Furimazine system

TESI IN COTUTELA



Surname: Bonardi Name: Alessandro

Registration number: 812370

Tutor: Claudio Greco

Supervisors: Ugo Renato Cosentino (UniMiB), Carlo Adamo (Chimie ParisTech - ENSCP)

Coordinator: Marco Giovanni Malusà

ACADEMIC YEAR 2023-2024

*“Una tesi c’ha bisogno di un contrario per capirla
una lucciola alla luce è un insettaccio, al buio brilla.”*

Nemico - Kid Yugi feat. Ernia

Contents

Abstract	8
Abbreviations and acronyms	9
1. Introduction	12
1.1 Bioluminescence, an intriguing natural phenomenon	12
1.1.1 Bioluminescence biological functions	
1.1.2 The evolutionary history of bioluminescence	
1.1.3 Examples of bioluminescent organisms	
1.2 Luciferin-luciferase systems	21
1.2.1 General mechanism of the luciferin-luciferase reaction	
1.2.2 Examples of notable natural luciferin-luciferase pairs	
1.2.3 Applications of the two most popular luciferins and derivatives	
1.3 Computational chemistry and bioluminescence	30
1.3.1 Modeling of reaction mechanisms	
1.3.2 Luciferin and luciferase rational design	
1.4 NanoLuc-furimazine pair	32
1.4.1 The enzyme - from OLuc to NanoLuc	
1.4.2 The substrate - furimazine	
1.4.3 Bioluminescent properties	
1.4.4 Applications	
1.4.4 Open challenges and unanswered questions	
1.5 Aim of the project	37
2. Computational methods	38
2.1 Molecular mechanics (MM)	38
2.1.1 Stretching and bending potential functions	
2.1.2 Torsional potential function	
2.1.3 Non-bonding potential	
2.2 The phase space	45
2.3 Molecular simulation methods	46
2.3.1 Molecular Dynamics (MD)	
2.3.2 Analysis of simulations	
2.4 Docking calculations	53
2.4.1 Searching	

2.4.2 Scoring	
2.4.3 Induced Fit Docking (IFD)	
2.5 Ab initio methods	60
2.5.1 Hamiltonian operator approximations	
2.5.2 Wavefunction approximations	
2.6 Density Functional Theory (DFT)	64
2.6.1 Time-Dependent Density Functional Theory (TD-DFT)	
2.7 Hybrid approaches - Quantum Mechanics/Molecular Mechanics (QM/MM)	71
2.7.1 ONIOM scheme	
2.8 Applied protocols	74
2.8.1 Modeling of NanoLuc-furimazine and NanoLuc-furimamide complexes	
2.8.1 Prediction of UV/Vis and NMR spectra of furimazine in DMSO	
2.8.2 Modeling of furimazine oxidation mechanism in DMSO	
3. Results	76
3.1 Modeling of NanoLuc-furimazine and NanoLuc-furimamide complexes	76
3.1.1 Identifying the substrate: furimazine enol and keto forms	
3.1.2 Building the NanoLuc-furimazine complexes	
3.1.3 Furimamide possible protonation states	
3.1.4 Building the NanoLuc-furimamide complexes	
3.1.5 Prediction of the emission spectra	
3.1.6 Conclusions	
3.2 Furimazine and furimamide characterization	103
3.2.1 Furimazine	
3.2.1 Furimamide	
3.2.3 Conclusions	
3.3 Modeling of furimazine oxidation mechanism in DMSO	113
3.3.1 Imidazopyrazinone core oxygenation	
3.3.2 Decomposition of dioxetanone intermediate	
3.3.3 Conclusions	
4. Conclusions and perspectives	134
References	136
A1. Appendix 1 - Basics of FRET photophysics	152
A2. Appendix 2 - Computational details	154
A2.1 Modeling of NanoLuc-furimazine and NanoLuc-furimamide complexes	154

A2.2 Furimazine and furimamide characterization_____	156
A2.3 Modeling of furimazine oxidation mechanism in DMSO _____	156
Acknowledgements _____	158

Abstract

The study of bioluminescence, a fascinating natural phenomenon, has generated a fervent scientific interest because of its various biological functions and potential applications in many different fields like chemical biology, imaging, and therapeutic procedures. This thesis work focuses on unravelling the bioluminescent mechanism of the NanoLuc-furimazine system, a luciferase-luciferin pair that recently became popular and widespread in bioanalytical applications due to its outstanding brightness and stability, but whose molecular processes are still unclear.

Section 1 revolves around bioluminescence diffusion, evolution and role in the natural reigns, including a detailed description of the most popular luciferin-luciferase systems and their applications. The role of computational chemistry in the struggle to get deeper insights of the mechanisms underlying their bioluminescence is then introduced, as well as NanoLuc-furimazine pair main features and characteristics, with a particular emphasis on the obscure aspects of its light emitting process.

Section 2 gives a small-scale cross-section of the state of the art of different computational methods, including molecular mechanics, molecular dynamics, docking calculations, and quantum mechanics/molecular mechanics (QM/MM) approaches, which were all applied to model the system of interest.

In the subsequent third section, then, the results of the research project are reported, with a three-parted structure. Section 3.1 main focus is modeling of the NanoLuc-furimazine and NanoLuc-furimamide complexes to identify the most likely protonation states of furimazine, the reactant of the catalytic process, and furimamide, the final emissive product. The results, which were published on *Physical Chemistry Chemical Physics*¹, revealed that the relative stability of the enol and keto forms of furimazine in aqueous solutions are compatible with the involvement of the former in the bioluminescent reaction. Basing on this hypothesis, the zwitterionic form of furimamide (FurC) emerged as the most probable final product, with its emission spectrum being the best fitting one with respect to experimental data. Even if the anionic form (FurD) was shown to have a quite fitting emission spectrum, it is less likely to be the culprit of the bright light deriving from NanoLuc activity due to its high protonation tendency.

Section 3.2 reports the spectroscopic characterization of furimazine and furimamide in polar solvents, combining UV/Vis and NMR spectroscopy, which was carried out to further support the involvement of the enol form as substrate of the NanoLuc catalytic cycle and check the exclusion of FurD from the pool of possible emitters. Experimental findings point at the predominance of the enol tautomer of furimazine in polar solvents such as dimethyl sulfoxide, which is consistent with the results obtained from quantum mechanical calculations. The observed extreme resistance of neutral furimamide (FurA) to deprotonation, then makes it possible to reasonably state that FurC, rather than FurD, is the primary emissive species.

Section 3.3 investigates the oxidation mechanism of furimazine, focusing on the formation and decomposition of the dioxetanone intermediate, a crucial species in the reaction, whose chemiluminescent decomposition accounts for many bioluminescent systems emission. Two possible pathways for furimazine oxidation, at the C2 and C3 positions, were explored, both of which are feasible, with energy barriers suggesting that C2 oxygenation is the more favourable one. The subsequent decomposition of dioxetanone was studied for three protonation states (being the precursors of FurA, FurC, and FurD), revealing insights into the mechanism by which these products may form.

All the results obtained are then summarized in the final section, to help delineating a comprehensive picture of the NanoLuc-furimazine system. This section also discusses some implications of these findings giving suggestions for further developments and a direction for future design and optimization of the NanoLuc-furimazine pair towards improved bioluminescence applications.

Abbreviations and acronyms

1A-MMPBSA - One average MM/PBSA

3A-MMPBSA - Three average MM/PBSA

ACN - Acetonitrile

ADP - Adenosine diphosphate

AMP - Adenosine monophosphate

AP - Alchemical perturbation

APN - Aminopeptidase N

ATP - Adenosine triphosphate

BLI - Bioluminescence imaging

BRET - Bioluminescence-resonance-energy transfer

BS - Broken symmetry

CBR - Click beetle red luciferase

CPCM - Conductor-like polarizable continuum model

CT - Charge transfer

DCM - Dichloromethane

DE - Dioxetanone

DFT - Density functional theory

DMSO - Dimethyl sulfoxide

DSSP - Dictionary of the secondary structure of proteins

DZ - Double zeta

ECP - Effective core potential

ES - Excited state

ESPT - Excited state proton transfer

FEP - Free energy perturbation

ff - Force field

FFT - Fast Fourier transformed

FLuc - Firefly luciferase

FMN - Flavin mononucleotide

FRET - Förster-resonance-energy transfer

FurA - Furimamide neutral form

FurC - Furimamide zwitterionic form
FurD - Furimamide anionic form
GA - Genetic algorithms
GAFF - General Amber Force Field
GFP - Green fluorescent protein
GGA - Generalized gradient approximation
GLuc - Gaussia luciferase
GS - Ground state
GTO - Gaussian type orbital
HOMO - Highest occupied molecular orbital
IFD - Induced fit docking
iLBP - Intracellular lipid binding protein
IMOMM - Integrated Molecular Orbital + Molecular Mechanics scheme
IMOMO - Integrated Molecular Orbital + Molecular Orbital scheme
ImPy - Imidazopyrazinone
IRC - Intrinsic reaction coordinate
LA - Link atom
LCAO - Linear combination of atomic orbitals
LDA - Local density approximation
LOD - Limit of detection
LUMO - Lowest unoccupied molecular orbital
MC - Monte Carlo
MD - Molecular dynamics
MLuc - Metridia luciferase
MM - Molecular mechanics
MMC - Metropolis Monte Carlo
MMGBSA - Molecular mechanics with Generalized Born and Surface Area solvation
MMPBSA - Molecular mechanics with Poisson-Boltzmann and Surface Area solvation
MO - Molecular orbital
NanoBiT - NanoLuc Binary Technology
NIR - Near infrared
Nluc - NanoLuc luciferase

NTR - Nitroreductase

ONIOM - Our own N-layered Integrated molecular Orbital and Molecular mechanics

PBC - Periodic boundary conditions

PDB - Protein data bank

PDB - Protein data bank

PDT - Photodynamic therapy

PES - Potential energy surface

PME - Particle Mesh Ewald

QD - Quantum dots

QM - Quantum mechanics

QM/MM - Quantum Mechanics/Molecular Mechanics

QST3 - Quadratic Synchronous Transit 3

RLuc - Renilla luciferase

RMSD - Root mean square deviation

ROS - Reactive oxygen species

RSH - Range separated hybrids

SCF - Self consistent field

SET - Single electron transfer

SLC - Split luciferase complementation

ssDNA - Single stranded DNA

STO - Slater type orbital

TD-DFT - Time-dependent density functional theory

TS - Transition state

TZ - Triple zeta

vdW - van der Waals

ZVE - Zero-point vibrational energy

1. Introduction

1.1 Bioluminescence, an intriguing natural phenomenon

Bioluminescence is a captivating natural phenomenon which consists of light production as a consequence of biochemical reactions within living organisms, thus meaning that it can be classified as a specific form of chemiluminescence^{2,3}. This phenomenon has intrigued human beings for centuries, even since Aristotle.

“Not everything that is visible depends upon light for its visibility. This is only true of the 'proper' colour of things. Some objects of sight which in light are invisible, in darkness stimulate the sense; that is, things that appear fiery or shining. This class of objects has no simple common name, but instances of it are fungi, flesh, heads, scales, and eyes of fish. In none of these is what is seen their own proper' colour.”

Aristotle - *De Anima*, II, 7

It is the result of a highly specialized reaction between a substrate, known as luciferin, and an enzyme called luciferase. Since these names (luciferin and luciferase) are not specific, substrates and enzymes are often identified with the species name or group denomination⁴. The interaction between the two, generally involving oxygen, generates light with minimal heat production. This cold light, as it is sometimes referred to, has been fine-tuned by evolution to serve a multitude of purposes across various species, that is why this phenomenon has been identified in at least 700 genera, not necessarily close in the phylogenetic tree⁵. Indeed, bioluminescence is found in a broad range of habitats, either terrestrial or marine, but it is most prominently associated with marine ecosystems, especially the sea-floor ones. It is estimated that over 75% of sea organisms have developed the ability to bioluminesce, making bioluminescence a key factor in the dynamics of this ecosystem and, given that the deep ocean is the largest habitat on earth by volume, a biological trait of major relevance on the planet⁶. The size of the aquatic environment provides for the ideal conditions to the variety of bioluminescent organisms. Oceans' pelagic zone can be divided into five vertical sections, shown in Figure 1.1, each presenting unique characteristics, leading to different bioluminescent abundance and function: the epipelagic zone, characterized by sunlight, goes up to 200 m of depth, followed by the mesopelagic one, where light dims quickly until almost disappearing in the bathypelagic zone, from 1000 m of depth to 4000 m, and abyssopelagic one (4000 m of depth and below). In these aphotic zones, where sunlight does not penetrate, bioluminescence plays a vital role in communication, predation, and survival and that is the reason why some species maintained functional eyes, even if small or colour-blind⁷.

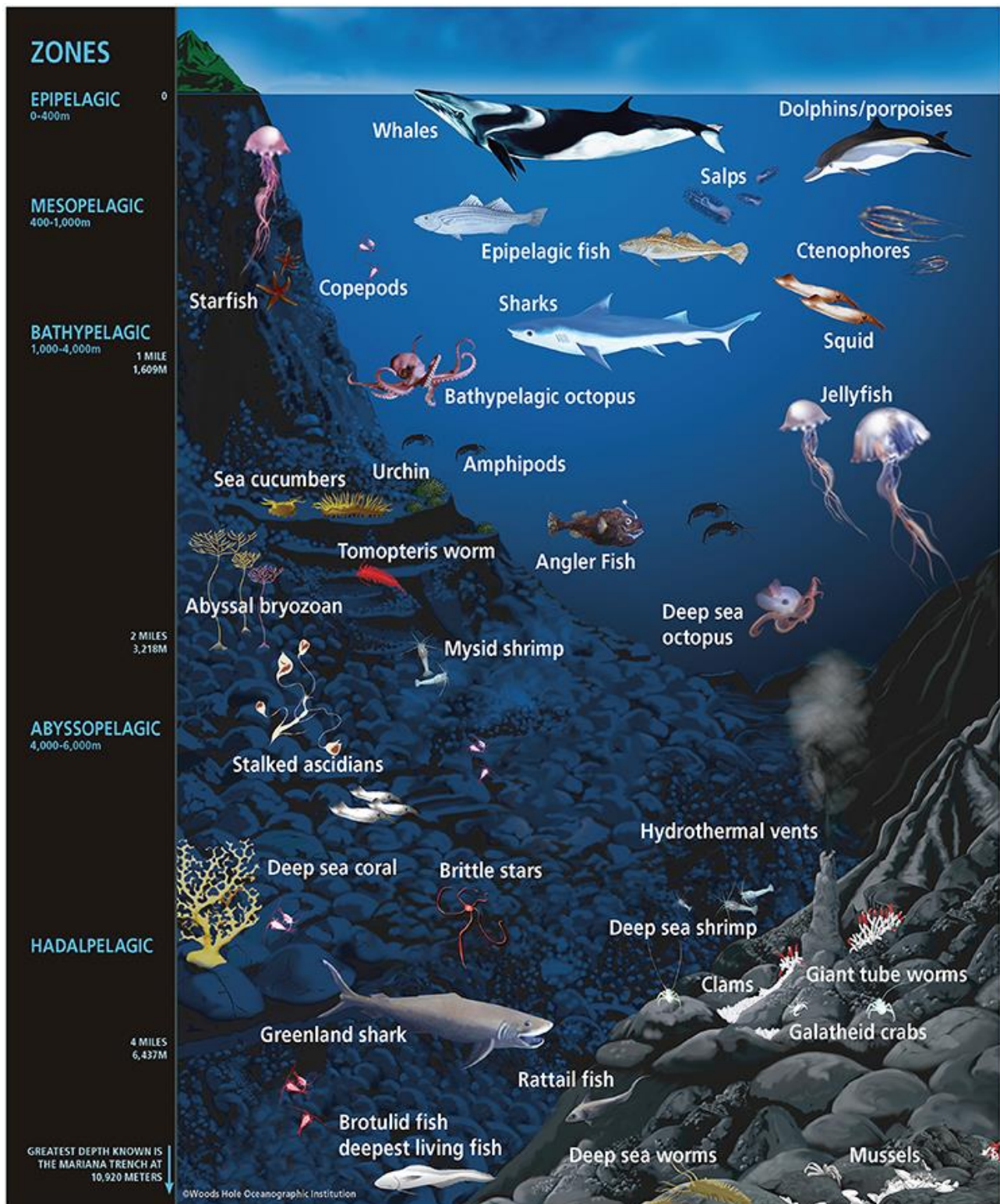


Figure 1. 1 - Pelagic zones and some of their inhabitants, image from reference⁸

1.1.1 Bioluminescence biological functions

Analysing marine bioluminescence provides for the exploration of all the numerous biological functions of this phenomenon, including:

- Camouflage
- Prey luring
- Defence
- Communication
- Mating
- Search for food

Camouflage:

One of the most popular purposes for natural bioluminescence is camouflage. In the oceans' depth, light may be scarce and not perceptible to human eyes, but a residual brightness is still perceptible to some organisms. That is the reason why many deep-sea fish use bioluminescence as a method of counter-illumination, a camouflage strategy characterizing some organisms, like lanternfish⁹, consisting in matching the light from above to blend with the background and become invisible to predators or prey beneath them¹⁰ (Figure 1.2). Preys can either try to avoid encountering predators or to scare them and bioluminescence can be helpful in both cases, since predators avoid preys with a "warning" colour pattern and this is the case of brittle stars and jellyfish like the *Aequorea victoria*¹¹.

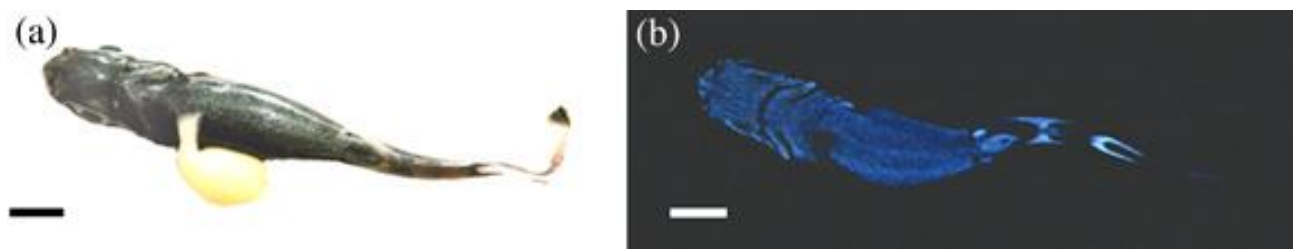


Figure 1. 2 - Spontaneous luminescence of a 115 mm total length *Etmopterus spinax* embryo. (a) Digital picture of the embryo when the light was turned on. (b) Digital picture of the same embryo when the light was turned off. Scale bars = 1 cm.

Figure taken from reference¹⁰

Prey luring:

While bioluminescent camouflage can be exploited by both predators and preys, deep-water anglerfish displays a predator-only light emission application. This lurking hunter uses a moving extension of its body, which resembles a fishing rod, to attract preys (Figure 1.3). The dangling lure is bright, but not emitting light itself: this organ contains bioluminescent bacteria and represents a perfect example of symbiosis, since the anglerfish exploits light to hunt, while bacteria get a safe environment to grow in return¹².

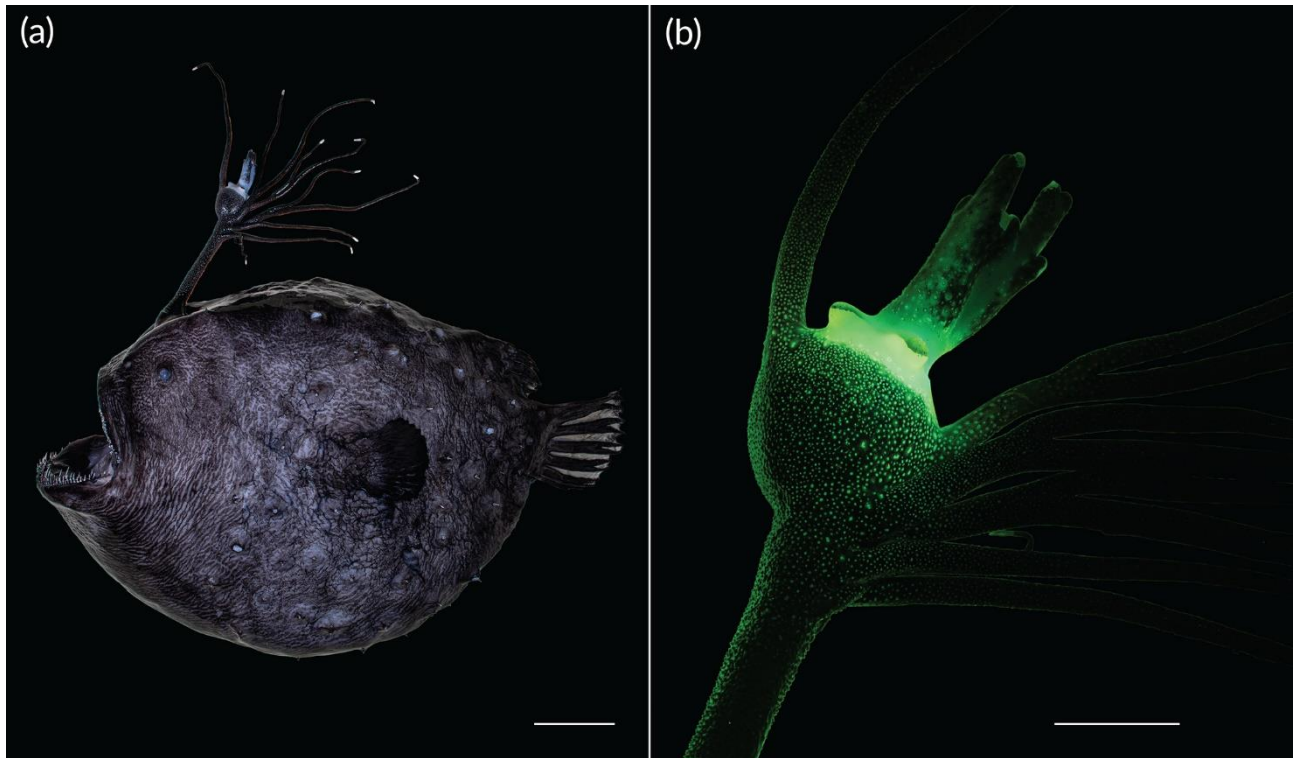


Figure 1. 3 - LACM 60082-1, a female *Himantolophus sagamius* prior to preservation (a). The scale bar represents 50 mm. The esca under royal blue excitation light (b). The scale bar represents 15 mm. Image taken from reference¹³

Defence:

Preys can exploit bioluminescence not only to avoid being attacked by predators either hiding or by aposematism, but also to defend themselves from the assault, usually expelling a bright ink cloud which usually sticks to the predator. This behaviour is found in some deep-sea squids¹⁴ or small ostracods like *Vargula hilgendorfii*¹⁵. Scare the predator is not the prey's only way to defend from an already started attack and *Noctiluca scintillans*¹⁶ (Figure 1.4) is a perfect example of the “burglar alarm” defensive strategy: after being attacked, or even when undergoing a shear stress, this dinoflagellate expels a shiny cloud not aiming at blinding the predator, but at attracting a larger one, which may put to flight the aggressor¹⁷.



Figure 1. 4 - Sea sparkle caused by *Noctiluca Scintillans* at the yacht port of Zeebrugge, Belgium. © Hans Hillewaert, CC BY-SA 4.0, <https://commons.wikimedia.org/w/index.php?curid=10711494>

Communication:

Bioluminescence can be useful not only to interact with members of other species, but also to communicate amongst organisms of the same species. A typical example is the quorum sensing communication in Pyrosomes, holoplanktonic colonial tunicates able to emit and detect light (Figure 1.5). The mechanism of luminescence propagation within the colony is not neural, but only light-dependent¹⁸. Each individual is able to detect light emitted by the others, even nearby colonies, with a receptor and emits itself as a response. This allows *Pyrosoma* zooids to coordinate swimming, for example, and may cause a marvellous spectacle like the one reported by T.H. Huxley in his diary¹⁸.

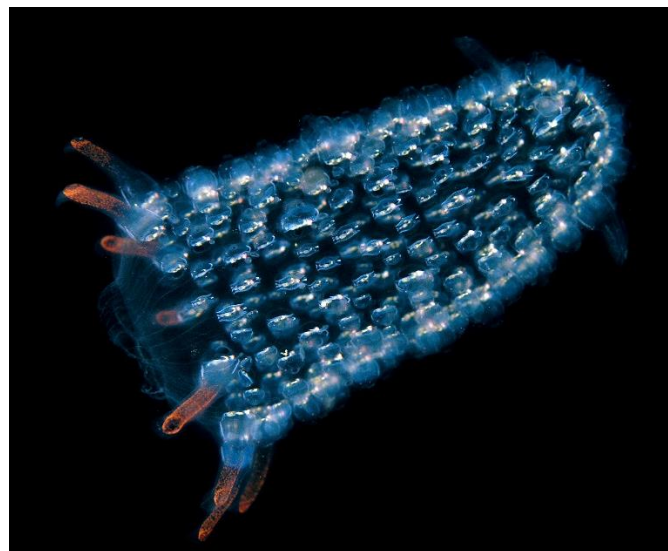


Figure 1. 5 - Pyrosome colonial tunicate encountered at Atauro island, East Timor. Photographer: Nick Hobgood - Opera propria, CC BY-SA 3.0, <https://commons.wikimedia.org/w/index.php?curid=4576657>

Mating:

Finding a partner is another crucial function of natural bioluminescence, in both types of habitats, terrestrial and marine ones. On land, in fact, bioluminescence is less common and mainly occurring in insects, but still remarkable, especially as part of the mating process. Fireflies, like Lampyridae firefly, the most well-known example of terrestrial bioluminescence, use their light-producing ability primarily for mating purposes, with different species emitting specific light patterns to attract mates¹⁹. The intricate flashing patterns are a result of complex neurobiological control, enabling an efficient communication with potential mates.

Search for food:

Since the depth of the ocean are almost totally dark, no light is available for hunting, but light emission may attract bigger predators. That is the reason why some organisms display red bioluminescence, difficult to be detected in water, to illuminate their way while searching for food without attracting potential predators¹¹.

These strategies exemplify the fine-tuned adaptation of bioluminescence, especially in deep ocean waters, where it flourishes. Despite its popularity, anyway, the question of bioluminescence origin and evolution remains relatively uncertain, and the evolutionary causes of the phenomenon are still under investigation.

1.1.2 The evolutionary history of bioluminescence

A credited theory is the one proposed by Seliger and McElroy²⁰, which presents light emission as just a byproduct of the oxygen removal reactions enacted by primitive anaerobes for survival purposes. The main evidence supporting this theory is the capability of many luciferins to quickly react with reactive oxygen species (ROS) and the universal requirement for oxygen in all known bioluminescent reactions, accompanied by the widespread character of bioluminescence, which is displayed by organisms in unrelated taxa, probably reflecting this shared survival strategy. This is further supported by the presence of coelenterazine, a typical luciferin, in non-bioluminescent organisms²¹, highlighting the fact that it is not only employed for light emission purposes. The evolution of bioluminescence from a byproduct of the detoxification process to the key element of the aforementioned survival, mating and communication strategies is a billion-year affair.

One hypothesis is that the previously antioxidant purposes have progressively become less relevant due to the migration of the involved organisms to less challenging conditions for them, such as the oceans depth, where the concentration of ROS is less than 2% if compared to the epipelagic zone, thus allowing the bioluminescent processes to be harnessed for different reasons, becoming selective traits²².

Another possible explanation is the enzyme-centric explanation, which suggests that luciferases derived from oxygenases employed by past organisms to get free of pigments or toxic substances linked to oxidative stress. A mutation in such an enzyme, catalysing the breakdown of a pigment, could result in an immediate selective pressure for the organism it belonged to, if it took place in a spot displayed to attract a mate or repel a predator⁵. The adaptive advantages conferred by bioluminescence, especially in environments where light is scarce, such as the deep ocean, would have driven its widespread evolution across multiple lineages. Today, bioluminescence is found in bacteria, fungi, invertebrates, and vertebrates, showcasing its versatility as an evolutionary trait.

Another critical aspect of bioluminescence is the range of colours produced by different organisms. The colour of bioluminescent light is determined by the specific luciferin and luciferase involved in the reaction, as well as the environment in which the reaction takes place, with emission wavelengths mainly ranging from 430 nm to 620 nm as shown in Figure 1.6¹¹.

Marine organisms, for instance, typically emit blue or green light, which travels the farthest in water. In contrast, some terrestrial organisms, like fireflies, emit yellow or orange light, which is more suitable for communication in air. The spectral tuning of bioluminescence to specific ecological niches highlights the intricate interplay between biochemistry and evolution.

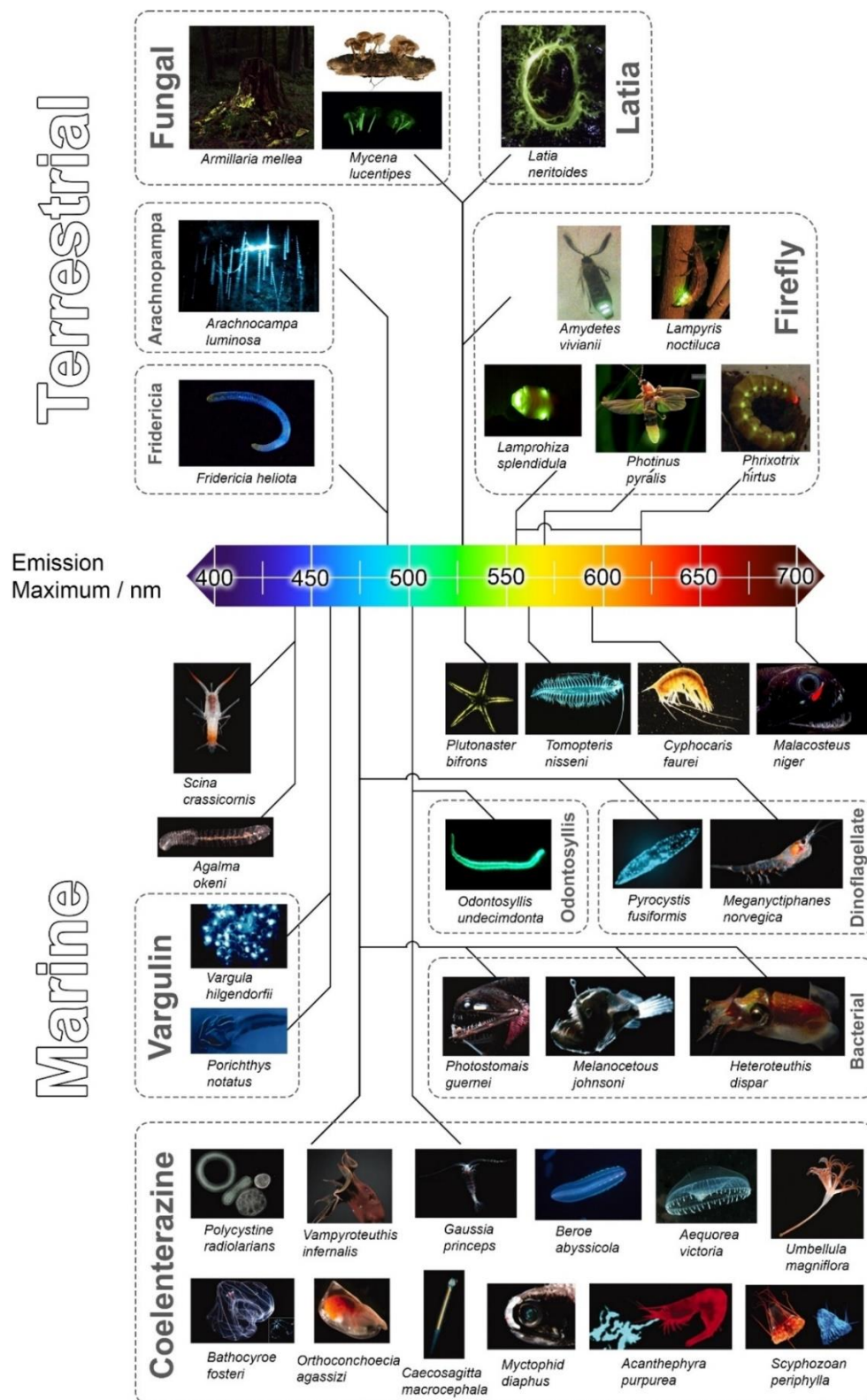


Figure 1. 6 - Spectra distribution of some bioluminescent organisms in the terrestrial (top) and marine (bottom) habitat at their approximate emission maximum. Systems with knowledge on the underlying structural chemistry are grouped with dotted frames. Image from reference¹¹

1.1.3 Examples of bioluminescent organisms

Bioluminescence is found across a wide range of taxonomic groups, a great indicator of its evolutionary effectiveness and versatility. One of the most famous bioluminescent organisms is the jellyfish *Aequorea victoria*, whose bioluminescence is due to the photoprotein aequorin²³. Indeed, aequorin is usually included in scientific articles and reviews dealing with luciferin-luciferases pairs, but the proper denomination should be “photoprotein”. In 1961 it was discovered and characterized, and it showed a unique behaviour, with light emission as a consequence of the addition of calcium ions to the environment, whether in presence of oxygen or not. This is due to the fact that aequorin is a photoprotein, an enzyme-substrate complex that is more stable than the corresponding dissociated components²⁴. Essentially, this protein, in absence of calcium ions, tends to bind coelenterazine and oxygen, being highly rigid and stable^{23,25}. Once the calcium ions are added to the environment, the coelenterazine oxidation reaction is triggered and results in the decomposition of aequorin in coelenteramide, i.e. the oxyluciferin, CO₂ and apoaequorin, i.e. the unbound luciferase²⁴. Apoaequorin is then able to reform the original photoprotein once it is put in presence of coelenterazine and oxygen²⁵. The bioluminescence of *Aequorea victoria* is not only important for understanding the organism's ecology but has also paved the way for other crucial findings in biotechnology and medicine, most notably the discovery of green fluorescent protein (GFP, Figure 1.7), which has become a fundamental tool in cellular imaging and gene expression studies^{26,27}.



Figure 1. 7 - "Steel Jellyfish" by Julian Voss-Andreae, 2006. Stainless steel, height: 4'7" (1.40 m). Location: Friday Harbor Labs, San Juan Island, WA. The sculpture is inspired by the structure of the GFP.

The marine environment, whom *Aequorea victoria* is a notable inhabitant, offers a great variety of other popular bioluminescent organisms, like the aforementioned anglerfish, a deep-sea predator well-known for the use of a bioluminescent lure, actually produced by symbiotic bacteria housed in a modified dorsal spine, to attract prey in the dark ocean depths¹². The anglerfish's bioluminescent lure is an excellent example of how this fascinating phenomenon can evolve to serve a specific predatory function. The light produced by the lure mimics the bioluminescence of small prey species, drawing potential prey close enough for the anglerfish to capture them.

The anglerfish is not a unique case of bioluminescence and symbiosis. Bioluminescent bacteria, such as *Vibrio fischeri*²⁸, often form symbiotic relationships with marine animals, such as the Hawaiian bobtail squid (*Euprymna scolopes*)^{29,30}. In particular, this squid harbours these bacteria in specialized light organs and uses their light for counter-illumination³¹. This symbiotic relationship, like the one between the anglerfish and the bacteria hosted in the lure, is a classic example of mutualism, where both the bacteria and the host organism benefit. The bacteria gain a protected environment and a source of nutrients, while the squid gains a powerful tool for avoiding predators. The study of bioluminescent bacteria has also provided insights into microbial ecology, symbiosis, and the regulation of gene expression in this taxon.

Another striking example of marine bioluminescence is the lanternfish, which uses bioluminescence for communication and mating^{9,10}. Lanternfish have photophores, specialized light-producing organs, distributed across their bodies. These organs produce blue-green light, which is ideal for communication in the ocean mesopelagic zone, where blue and green wavelengths penetrate most effectively. The patterns of light produced by these photophores are species-specific, allowing individuals to identify each other and facilitating reproductive behaviours in the vast darkness of the ocean. The ecological importance of lanternfish is profound, as they are a major component of the marine food web, and their bioluminescence plays a key role in their survival and reproduction³².

Survival strategies based on bioluminescence may vary and one of the most notable group of organisms exploiting them is dinoflagellates, single-celled marine plankton that produce light in response to mechanical stimulation. Dinoflagellates are responsible for the mesmerizing glow seen in ocean waves at night, particularly in regions where they are abundant¹⁶. The bioluminescence of these organisms is thought to serve as a defence mechanism, deterring predators by either startling them or attracting larger predators to the scene, with a typical “burglar alarm” strategy. In certain conditions, dinoflagellate blooms can result in phenomena known as “red tides,” where large concentrations of these organisms produce bioluminescent displays visible over wide areas, visible even from space (Figure 1.8)³³. The environmental conditions that trigger these blooms, and the ecological impacts of bioluminescent dinoflagellates, are the subject of ongoing research, especially as climate change alters marine ecosystems^{34,35}.

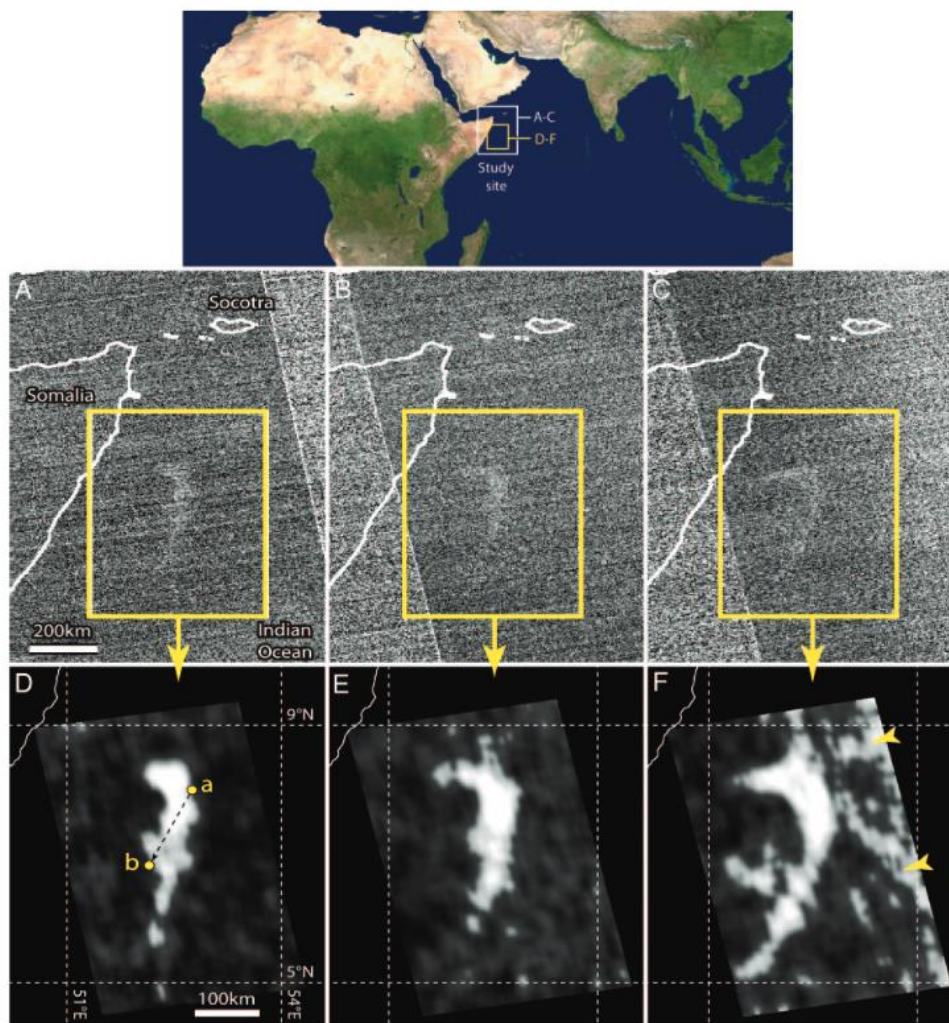


Figure 1. 8 - Satellite images of a bioluminescent tide caused by dinoflagellates. Image taken from reference³³

In terrestrial ecosystems, bioluminescence is most commonly associated with fireflies, but other organisms, such as certain fungi, also exhibit the trait. This is already known since Pliny the elder, who reported in his *Naturalis Historia*, about the characteristic glow of damp wood³⁶. This bioluminescence, often seen in decaying wood and creating an eerie glow known as "foxfire", is typically linked to *Armillaria mellea*, commonly known as honey fungus. It produces light from their mycelium, the root-like structures of the fungus. The ecological function of fungal bioluminescence is still not fully understood, but it is hypothesized to play a role in spore dispersal or in attracting insects that aid in reproduction^{37,38}.

Beyond these well-known examples, bioluminescence is present in a wide variety of other organisms, including shrimp¹⁴, worms³⁹, and even some species of sharks^{10,40}. The deep-sea environment, in particular, is a hotbed of bioluminescent activity, with species evolving novel ways to utilize light in the extreme darkness of the ocean floor. Some species use bioluminescence to lure prey, while others take advantage of it for mating purposes or to disorient predators. The diversity of bioluminescent organisms and their uses of light make bioluminescence one of the most intriguing subjects in evolutionary biology.

1.2 Luciferin-luciferase systems

1.2.1 General mechanism of the luciferin-luciferase reaction

Despite being a multifaceted process, which permeates the entire phylogenetical tree, the beating heart of bioluminescence, i.e. the fundamental biochemical process behind it, is the oxidation of a luciferin catalysed by a luciferase. The interaction between the enzyme (luciferase) and its substrate (luciferin), in presence of molecular oxygen, results in the production of light as a consequence of the formation of an oxidized product (oxyluciferin) in its excited state (ES), subsequently involved in radiative decay to the ground state (GS)²⁴. Sometimes this reaction requires additional cofactors, like adenosine triphosphate (ATP) in the case of all beetles luciferases⁴¹ or flavin mononucleotide (FMN), typical of bacteria⁴². The luciferin-luciferase system has evolved multiple times in nature, leading to a variety of bioluminescent mechanisms differing in their specific chemical pathways but sharing the same basic process of light generation.

At its core, the luciferin-luciferase reaction is a chemiluminescent process, i.e. luminescence arising from chemiexcitation². In most luciferin-luciferase systems, the reaction proceeds through a series of well-defined steps:

1. *Activation of Luciferin:* In ATP-dependent systems, like the well-known firefly luciferase, luciferin is first activated by the enzyme luciferase, which catalyses the formation of a luciferyl-adenylate intermediate⁴³. This step involves the attachment of an AMP (adenosine monophosphate) group to the luciferin molecule, creating a highly reactive intermediate. In some cases not only ATP, but also other cofactors, like FMN or Mg^{2+} are required, in others the activation is simpler, consisting of a deprotonation, and does not involve other chemical species^{43,44}.
2. *Oxidation:* Once activated, the luciferin undergoes oxidation in the presence of molecular oxygen. This step is critical, as the molecular oxygen reacts with the activated luciferin to form an excited-state oxyluciferin⁴³⁻⁴⁵. The energy released during this reaction is temporarily stored in the excited-state molecule.
3. *Light Emission:* As the oxyluciferin returns to its ground state, the stored energy is released in the form of visible light. The specific wavelength of the emitted light depends on the chemical structure of the luciferin and the environmental conditions, such as pH and the presence of cofactors²⁴.

The reaction typically produces light in the blue to green spectrum, although some species can emit yellow or red light¹¹. One of the key aspects of the luciferin-luciferase reaction is its requirement for molecular oxygen, O₂. In all known systems, it plays a crucial role in the oxidation of luciferin, leading to the generation of light. This dependency on molecular oxygen suggests that bioluminescence evolved in aerobic environments, where it was readily available, and supports the aforementioned hypothesis of the emergence of bioluminescence as a way for anaerobes to get rid of the toxic element (see chapter 1.1.2). In addition to O₂, some luciferin-luciferase systems require cofactors or other molecules to start or modulate the reaction. For example, calcium ions (Ca²⁺) are necessary for the bioluminescence of certain marine organisms, such as the aforementioned *Aequorea victoria*^{24,27}. In this system, Ca²⁺ ions act as a trigger, activating the luciferase enzyme and initiating the light-producing reaction. In absence of calcium ions, the enzyme-substrate complex is reprimed²⁵.

The luciferin-luciferase reaction is highly regulated in living organisms, ensuring that light is produced only when needed. This regulation can occur at multiple levels, from the synthesis of luciferin and luciferase to the activation of the enzyme and the control of oxygen availability. In some species, light production is under circadian control, with bioluminescence occurring only during specific times of day, such as at night⁴⁶. In other species, bioluminescence is triggered by external stimuli, such as mechanical disturbance or changes in environmental conditions^{14,16,34,47}.

The molecular structure of luciferin varies between different species, but it typically consists of a small, highly conjugated organic molecule⁴⁸. Part of the conjugated structure is conserved even after the oxidation process, which is usually associated with the elimination of a CO₂ molecule, as shown in Figure 1.9. Conjugation reduces the energy gap between the highest occupied molecular orbital (HOMO) and the lowest unoccupied molecular orbital (LUMO), making it possible for excited oxyluciferins to emit visible light, spacing in the electromagnetic spectrum depending on the molecular structure⁴⁹.

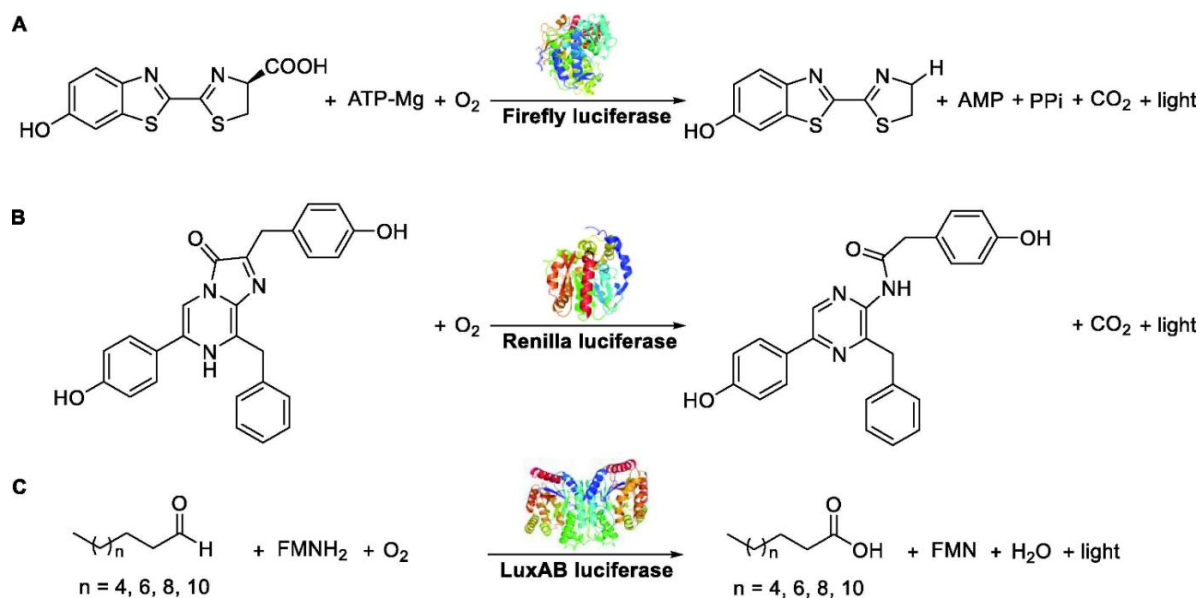


Figure 1. 9 - Chemical equation of three different luciferin oxidation mechanisms. (A) Pathway of firefly luciferase (PDB: 1LCI). (B) Pathway of Renilla luciferase (PDB: 2PSD). (C) Pathway of LuxAB luciferase (PDB: 1XKJ). Image taken from reference⁵⁰

Luciferase, the enzyme responsible for catalysing the reaction, is equally diverse. While its basic function is conserved across species, the structure of the enzyme can vary significantly. This variation in enzyme structure is due to the selectivity of the enzyme with respect to the corresponding substrate and is responsible for the differences in reaction kinetics, and overall efficiency observed in different bioluminescent systems. Even when working with identical substrates, luciferases may present different structural features, thus strengthening the hypothesis that bioluminescence has multiple parallel origins⁵¹.

1.2.2 Examples of notable natural luciferin-luciferase pairs

Bioluminescence has evolved a minimum of 40 times throughout the history of life, leading to the development of a variety of luciferin-luciferase pairs⁵². Even if some bioluminescent organisms share the same luciferin, each luciferin-luciferase pair has its own unique chemical characteristics, tailored to the ecological needs of the organism. Amongst the great variety of natural bioluminescent systems, the luciferin and luciferase structures have been identified for only 11 of them⁵³. The most popular ones are the systems from the following organisms:

Photinus pyralis: The firefly luciferin-luciferase system is perhaps the most famous and widely studied example of bioluminescence. In this system, firefly luciferase, also known as FLuc catalyses the oxidation of D-luciferin in presence of ATP, Mg^{2+} and O_2 , producing a yellow-green light. Light emitted by fireflies is used primarily for mating and communication, with different species displaying distinct flashing patterns to attract mates¹⁹. Firefly bioluminescence is a remarkably efficient one, with a quantum yield of $41 \pm 7.4\%$, higher than those of other bioluminescent organisms⁵⁴. D-luciferin is one of the most popular luciferins, since more than 40 different species of fireflies, click-beetles and railroad worms rely on it for their bioluminescence⁵³. *Photinus pyralis*, commonly known as North American firefly, is characterized by a 62 kDa luciferase, which catalyses the oxidation of the luciferin in two different steps, reported in Figure 1.10. First, D-luciferin reacts with adenosine-5'-triphosphate- Mg^{2+} , forming the adenylyl intermediate, then this intermediate is oxidized to give the corresponding oxyluciferin, AMP and CO_2 ⁵⁵. The reaction of firefly luciferin-luciferase system has been extensively studied and it was demonstrated that it starts with the adenylation of luciferin, followed by a single electron transfer (SET) from the negatively charged adenylyl intermediate to an oxygen molecule, leading to the oxygen attack to form a peroxide anion⁵⁶. The intramolecular cyclization of the peroxide leads to the elimination of AMP and gives a dioxetanone anion. Its decomposition is still matter of debate, since it was believed to happen from the anionic form of it, but, according to results by Pinto da Silva's group⁵⁵, it exists in the negatively charged form only in polar environments and at a basic pH, while the reaction occurs in the protein environment and in a wide range of pH (5 - 8). Thus, the dioxetanone anion is protonated, then a CO_2 molecule is eliminated to form oxyluciferin in its excited state. At this point, an excited state proton transfer (ESPT) takes place, since this type of molecules display a strong photoacidity⁵⁵, leading to the formation of the anionic emitter which is protonated again after bioluminescence. The entire process has been object of many studies due to its potential applications in biotechnology, especially to measure ATP concentration in cellular environment, monitor ATP-forming and ATP-degrading reactions exploiting the ATP-dependence of the oxidative process⁵⁷⁻⁵⁹.

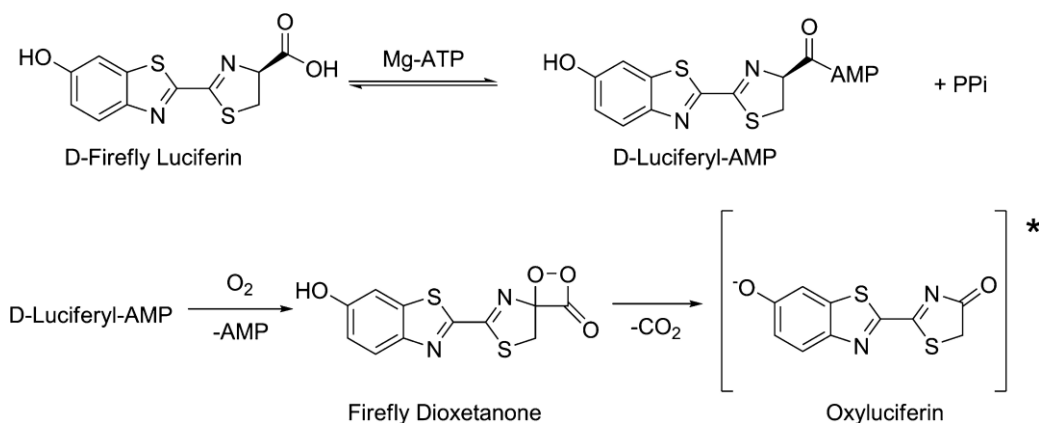


Figure 1. 10 - Mechanism of firefly luciferin excitation. Image taken from reference⁶⁰

- *Renilla reniformis*: Coelenterazine represents one of the most widely distributed luciferins in the marine environment, being found in a variety of organisms, around 15, including jellyfish, shrimp, and fish⁵³. It is a larger molecule, if compared to D-luciferin, is scarcely soluble in water and it is also more prone to auto-oxidation in solution due to the fact that no cofactors are required for the prosecution of the reaction⁶¹. Coelenterazine has also been a critical molecule in scientific research, particularly in the study of calcium signaling, as it is the substrate of the aequorin, a calcium-sensitive photoprotein which has been used to monitor intracellular calcium levels²⁷. Oppositely to D-luciferin, this molecule undergoes direct oxidation in the presence of oxygen, catalysed by luciferase, to produce blue light, without any cofactor being required. The first coelenterazine-based luciferase was discovered by Shimomura from the deep-sea marine shrimp *Oplophorus gracilirostris*¹⁴, while the *Renilla reniformis* luciferase, also known as RLuc, was one of the earliest to be cloned⁶² and a possible mechanism for its catalytic activity, reported in Figure 1.11, has been recently proposed⁴⁴. The reaction is partially analogous to the one observed in *Photinus pyralis*, as coelenterazine is converted into the corresponding excited state oxidized luciferin, coelenteramide, through a dioxetanone intermediate. In this case, the reaction occurs without adenylation though. Coelenterazine is a larger molecule, if compared to D-luciferin, is scarcely soluble in water and it is also more prone to auto-oxidation in solution due to the fact that no cofactors are required for the prosecution of the reaction⁶¹.

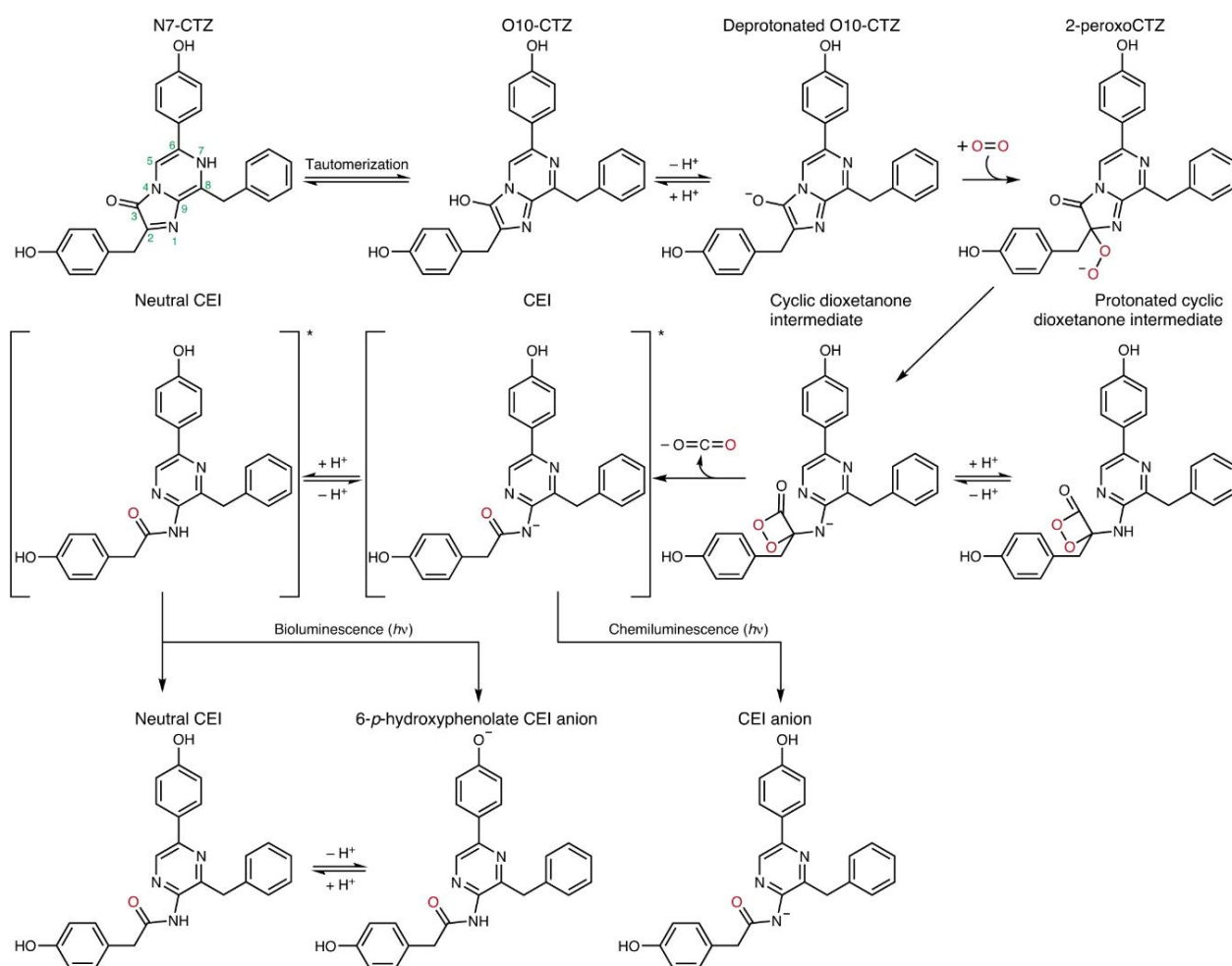


Figure 1. 11 - Coelenterazine oxidation mechanism, both chemiluminescent and bioluminescent path represented. Image taken from reference⁴⁴

- *Vibrio fischeri*: Bacterial bioluminescence is another well-studied example, particularly in species such as *Vibrio fischeri*. This bacterium uses a luciferin-luciferase system to produce blue-green light and is a perfect example of the symbiotic relationships involving bacteria and marine animals, such as squid and fish, where the bacteria provide light in exchange for nutrients and a protected environment³⁰. One of the most well-known examples of this symbiosis is the relationship between *Vibrio fischeri* and the Hawaiian bobtail squid (*Euprymna scolopes*)²⁹. The squid houses the bacteria in specialized light organs, and the bacteria produce light that the squid uses for counter-illumination, a form of camouflage. Bacterial bioluminescence is characterized by a unique mechanism, depicted in Figure 1.12, based on FMN. FMN is firstly reduced to form FMNH, which is then deprotonated and involved in an oxygenation to give the FMN-4a-hydroperoxide. This represents the precursor of the emitting species. After the reaction between this intermediate and a long-chain aldehyde, FMN-4a-peroxyhemiacetal is formed. This leads to the monooxygenation of the aldehyde to give a long-chain fatty acid, subsequently eliminated to give the FMN-4a-hydroxide in an excited state, which is the emitting compound. After the emission, a water molecule is eliminated and FMN is reformed⁴².

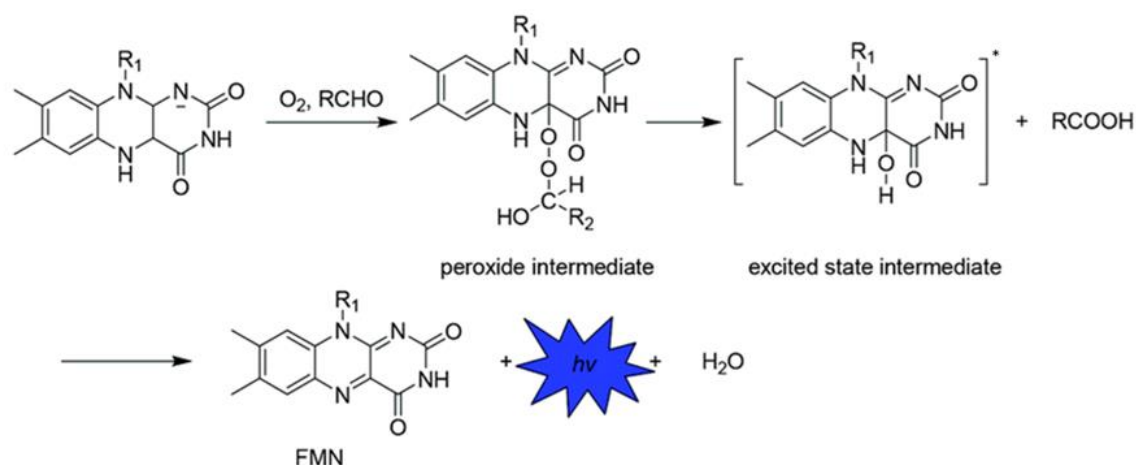


Figure 1. 12 - Mechanism of the bacterial bioluminescent system. Image taken from reference⁶³ and modified

1.2.3 Applications of the two most popular luciferins and derivatives

The luciferin-luciferase reactions have become an invaluable tool in bioimaging and biosensor development due to the relatively high sensitivity, specificity, lack of background noise and relatively low cost⁶⁴. Unlike fluorescence, bioluminescence does not require external excitation since it occurs naturally through the luciferin-luciferase reaction, raising the signal to noise ratio (S/N) and making it an ideal phenomenon to exploit for low-abundance targets detection in complex biological systems and whenever avoiding false-positive signals is desirable^{58,65,66}. Nowadays, bioluminescence is the underlying phenomenon to a wide range of applications which require high sensitivity, varying depending on luciferin-luciferase properties. The two most popularly employed luciferins are D-luciferin and coelenterazine, with respective derivatives, and both of them have multiple usages.

1.2.3.1 D-luciferin & derivatives

The ATP-dependence of D-luciferin oxidation process is a unique feature of this luciferin, but it is far to be the only one, since the *Photinus pyralis* - D-luciferin pair displays an impressive quantum yield ($41 \pm 7.4\%$ ⁵⁴) and a notable maximum emission wavelength ($\lambda_{\max} = 558 \text{ nm}$ ⁵⁴), respectively the highest and the longest amongst natural luciferases. Those properties (Table 1.1) place D-luciferin based luciferin-luciferase pairs amongst the pool of the best options of bioluminescent systems for bioluminescent imaging (BLI), especially when blood is involved. This is due to the fact that haemoglobin and mammalian tissues absorb mainly light at a shorter wavelength ($<500 \text{ nm}$)⁶⁷. The more red-shifted emission compared to the one from other pairs does not solve completely the light absorption and scattering in tissues though, since the transparent window for mammalian tissues is located at longer wavelengths, in the near infrared (NIR) region ($650\text{-}900 \text{ nm}$)⁶⁷⁻⁶⁹, and, despite the great variety of natural D-luciferin dependent luciferases, none of them is suitable for *in vivo* BLI.

	FLuc
$\lambda_{\max}$	558 nm
Optimal pH	7.8
Optimal T	28°C
Quantum yield	$41 \pm 7\%$ at pH 8.5

Table 1. 1 - Properties of firefly luciferase (Fluc) from *Photinus pyralis*. Table taken from reference⁵³

In addition, there are other relevant drawbacks to the utilisation of D-luciferin for BLI, in particular the low cell permeability⁷⁰, which forces large amounts dosing⁷¹ and the inhomogeneous bio-distribution displayed in rodent models⁷², topped by a scarce uptake in some organs of interest⁷³. The main issue, the overlap between luciferin emission and tissues absorption, partially attenuated by the notable shift observed *in vivo* ($\lambda_{\max} = 610 \text{ nm}$ ⁷⁴), has been tackled with the design of new substrates, some of them reported in Figure 1.13, leading to red-shifted emission, with best results obtained by extending the π system by adding double bonds between the thiazole moieties or adding condensed rings to one of the two^{70,75}. This has been possible since Fluc catalytic activity was not too strongly affected by substrate modifications⁷⁶. Focus has then been moved from the luciferin engineering to the luciferase one, not only to obtain more red-shifted variants, but also to make the emission signal stable at a wider range of conditions, including variation of pH range, temperature or concentration of different chemical species⁵³. These efforts were usually paired with the design of new substrates (Figure 1.13), to get better performances, but in no case it was possible to get a red-shifted emission without loss of light intensity^{53,77}.

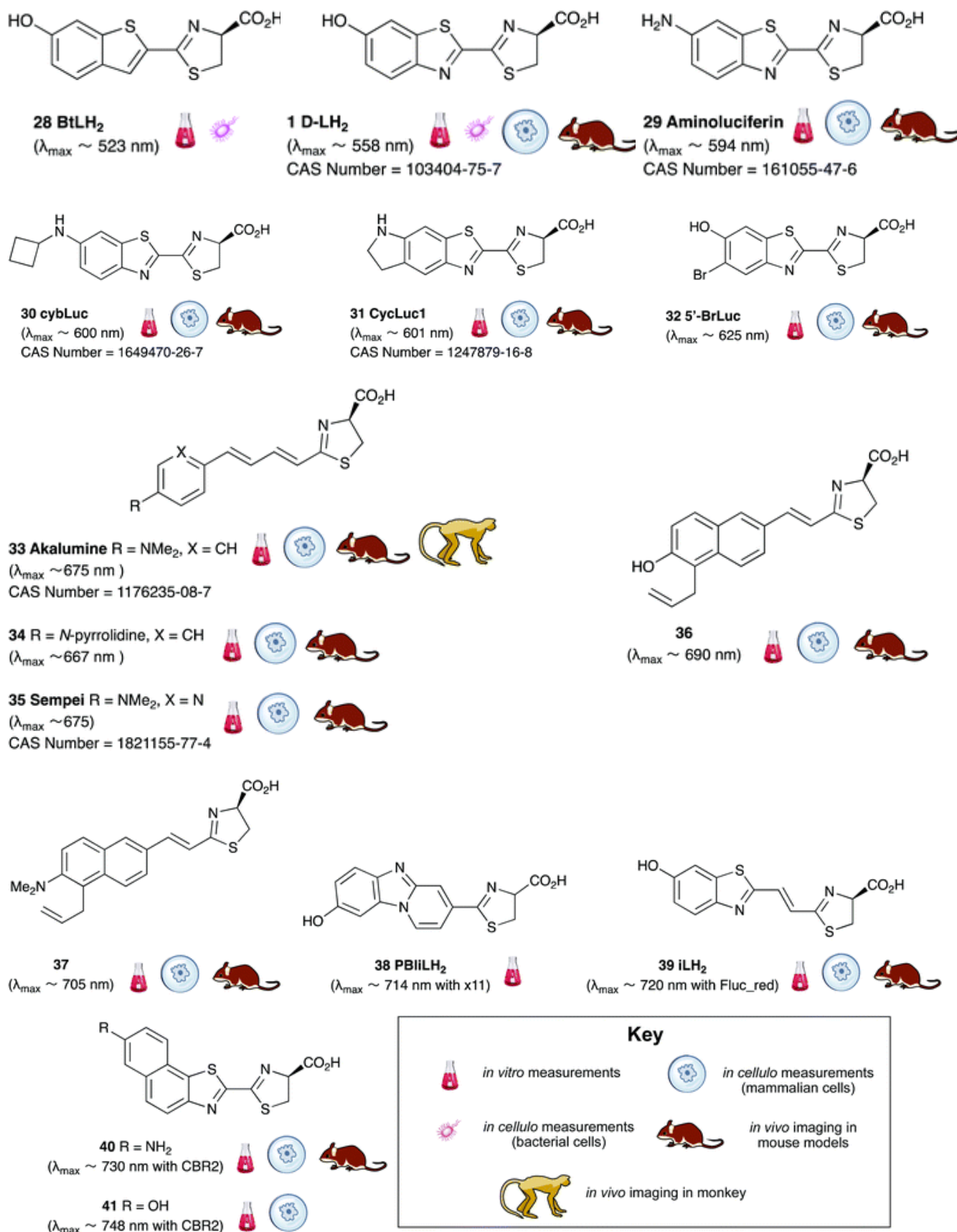


Figure 1. 13 - A selection of synthetic D-luciferin analogues and their λ_{max} . Image taken from reference⁵³

Nonetheless, the FLuc-D-luciferin derived pairs are still widely used, with a vast variety of usages:

- *Enzyme activity evaluation*: luciferin structure can be modified not only to generate a different emission after luciferase action, but also to get a prosubstrate of the reaction, which can be transformed into the active form only after metabolism by another enzyme. This is the rationale behind the design of caged luciferins, i.e. inactive luciferin derivatives activated by the action of an enzyme, like galactosidase⁷⁸ and α -chymotrypsin⁷⁹, whose activity can then be measured basing on the bioluminescence started by the uncaging of the substrate. This has a myriad of declinations, like the detection of cancer basing on the tumour-specific enzyme activity, like aminopeptidase N (APN)⁸⁰, nitroreductase (NTR)⁸¹ and CYP450 reductase⁸², which is not a typical tumour-specific enzyme, but its activity is strongly influenced by hypoxia conditions, a characteristic of the less vascularized tumour micro-environment⁸³.
- *Small molecules and ions detection*: caged luciferins can also be activated by the presence of specific small molecules or ions, directly or indirectly. This is the reason why specific modified D-luciferins, with a boronic acid (H_2O_2 sensitive) at the 6' position, were designed⁸⁴ to monitor the concentration of hydrogen peroxide molecules, typically higher in tumour tissues and in case of a neurodegenerative disease⁸⁵. In presence of hydrogen peroxide, D-luciferin is reformed, and bioluminescence is detected. A similar idea is behind the design of a Cu^{2+} detection method⁶⁵. In fact, some metal ions are involved in an altered homeostasis as a consequence of metabolic disorders, neurodegeneration and infection, and copper is among them⁸⁶.
- *Cell-cell interaction visualization*: specific tissues can be detected also exploiting the interaction between them and other cells. If a cell population is designed to express specific enzymes like galactosidase and another one is engineered to express luciferase, it is possible to monitor and visualize their interaction. It is enough to modify D-luciferin to get a prosubstrate activatable via enzyme specific activity and signal is then generated only when the two populations interact with each other, since luciferin is inactive until the specific enzyme action and its migration to luciferase-expressing cells is more probable the closer the two populations are⁸⁷.
- *ATP concentration measurement*: as previously mentioned, since the D-luciferase oxidation is ATP-dependent, it can be exploited to monitor ATP formation and depletion reactions, as the ATP availability is directly linked to the strength of the bioluminescent emission^{41,57,58}.

1.2.3.2 Coelenterazine & derivatives

If D-luciferin can be considered as the most representative luciferin for terrestrial habitats, coelenterazine plays the same role for marine ones²⁴. It is a versatile luciferin used by a variety of organisms, more than 15, including *Renilla* (RLuc), *Metridia longa* (MLuc) and *Gaussia* (GLuc) species^{23,62,88}. The coelenterazine-based bioluminescence systems do not need ATP to emit, as the oxidation requires only oxygen and activation of the substrate by luciferase enzyme to occur. Coelenterazine-based systems' drawbacks are several: the emission quantum yield is generally lower than the one of FLuc and mutants^{89,90} and associated emission is blue-shifted ($\lambda_{max} = 480-495$ nm), making them generally unsuitable for BLI (properties shown in Table 1.2).

	RLuc	Gluc	Mluc2
λ_{max}	480 nm	485 nm	488 nm
Optimal pH	7.4	7.5–8.4	7.5–8.4
Optimal T	32°C	15–20°C	5°C
Optimal [NaCl]	0.5 M	1.6 M	1.0–1.5 M
Quantum yield	5.3 $\pm$ 0.1%	n.a.	n.a.

Table 1. 2 - Properties of common marine luciferases. Table taken from reference⁵³

On top of that, coelenterazine is bigger, less soluble and more toxic than D-luciferin and also less stable, due to the auto-oxidation process that can be initiated from environmental conditions, since it does not require adenylation to be activated⁶¹. Despite this, the broad range of organisms using coelenterazine makes its luminescence adaptable for various contexts, especially in situations where energy conservation is a concern. One of the main advantages of coelenterazine-employing luciferases is their size, a desirable property for applications involving small vectors, as both RLuc, Gluc and MLuc are smaller than FLuc⁵³. As for D-luciferin, in an effort to mitigate the main issues of coelenterazine-based bioluminescent pairs, new substrate derivatives have been synthesized, aiming to enhance stability and red-shift emission, yet often at the cost of lower light intensity⁸⁹. Nevertheless, coelenterazine-luciferase pairs have been widely applied in diverse fields, including:

- *Gene reporter assays*: *Gaussia princeps* luciferase was cloned in 2002 and quickly became popular for its small size (19.9 kDa) and its secretory character. This permits non-invasive and long-term monitoring of biological phenomena, since neither a tissue lysis step nor a frequent substrate injection is necessary. This is why this coelenterazine-based luciferase, like *Metridia longa* luciferase, another luciferase which presents the same behaviour is commonly used as a reporter protein⁹¹.
- *Protein-protein interaction monitoring*: many luciferases operating with coelenterazine are exploited to monitor protein-protein interaction since it is possible to realise the so called split luciferase complementation (SLC) assays⁹². This method consists of splitting a luciferase into two different subunits (C-terminal and N-terminal), both unable to single-handily catalyse the oxidation process, which are bound to the two proteins whose interaction is the object of the assay. Once the two proteins interact with each other, the two luciferase subunits are brought close enough to enable the restoration of the luciferase, thus resulting in bioluminescent activity, if the substrate is available⁹³. The same process can be exploited to study cell-cell interaction, with a population of cells containing the C-terminal luciferase and another expressing the N-terminal one⁹⁴.
- *Calcium ions sensing*: as previously discussed, aequorin is not a luciferase, but, more properly, a photoprotein operating with coelenterazine as a substrate. It is popularly used to evaluate the calcium concentration in cells since its bioluminescence is triggered by a higher value of it. The Ca^{2+} ions start the coelenterazine oxidation by changing the conformation of the protein, probably changing the hydrogen bonding network in the pocket⁹⁵, leading to the formation of apoaequorin, coelenteramide and CO_2 .
- *BRET based biosensors realization*: The sea pansy, *Renilla reniformis*, has played a crucial role in redefining the boundaries of bioluminescence imaging and processes monitoring due to the bioluminescence resonance energy transfer (BRET) process which involves its luciferase RLuc and the green fluorescent protein (GFP). BRET is the non-radiative excitation transfer between two molecular entities, the bioluminescent emitter, called donor, and another species, the acceptor. It essentially analogous to a Förster-resonance-energy transfer (FRET), the only difference being the origin of the transferred energy, not deriving from previous light absorption, but being a product of a biochemical process. As in a FRET process², there is no light emission and subsequent reabsorption, but a direct energy transfer resulting from the long-range coupling of the donor and acceptor dipoles (photophysics better described in Appendix 1)⁹⁶. In *Renilla* luciferase the BRET happens to involve the luciferin-luciferase pair as the donor and the GFP as the acceptor. This is the reason behind the detected green luminescence ($\lambda_{\text{max}} = 510 \text{ nm}$) which is red-shifted if compared to the blue light emitted by the luciferin-luciferase pair alone. The BRET logic was applied for the design of several biosensors and to improve bioluminescence imaging, allowing to partially circumvent the tissue absorption issue^{97–101}.

1.3 Computational chemistry and bioluminescence

The wide range of applications of bioluminescence, spanning from medical imaging and diagnostics to biological processes' monitoring, is the main reason for the deep scientific interest towards this phenomenon. This interest has translated into many experimental works over the past years, with almost no theoretical contribution up to 21st century, as reported by Shimomura²⁴. Experimental chemistry is obviously crucial to explore bioluminescent systems, but it cannot give the answers to critical questions related to the processes. No experiment can explain how the spin-forbidden reaction between triplet O₂ and luciferase is the first step of one of the most efficient processes in nature or how does the reaction lead to the formation of an excited state product or even which is the final product. To find the answers to these (and other) questions and harness the potential of these processes to widen the already vast range of applications, researchers increasingly turn to computational chemistry. Computational chemistry brings into play the exact nature of atomic and molecular dynamics associated with bioluminescent reactions, thus helping to bridge the gap between theoretical description and experimental observation. By doing this, it offers a unique opportunity to study the details of the underlying reaction mechanisms and, consequently, represents a valuable tool for the systematic design of both luciferins and luciferases.

1.3.1 Modeling of reaction mechanisms

The core of bioluminescence is the luciferin oxidation process, catalysed by luciferase, eventually in presence of cofactors. Although the overall reaction may appear straightforward, complex intermediates and excited-state dynamics are involved that are not always completely accessible using experimental techniques. First of all, the key stage of bioluminescence is the emission of light, caused by the radiative decay of an excited state molecule caged in a protein environment. Neither excited states, nor enzymatic environments are easy-to-do studies and sometimes there are even additional hurdles. For example, cases where the crystal structure of an intermediate or of the luciferin-luciferase pair are rare and, in most cases, neither the enzymatic pocket has been clearly identified. In these situations, the only solution is to appeal to docking techniques, i.e. computational methods which allow to get hypothetical structures for a ligand and a receptor combined (more details in Chapter 2.1)¹⁰². Getting the luciferin-luciferase complex structure is not the only valuable contribution from computational chemistry: quantum mechanical methods (see Chapter 2.5) allow the identification of key transition states and intermediates involved in the oxidation process, making it easier to rationalise.

For instance, in the well-studied firefly luciferin-luciferase system, computational studies have been carried out to detail the steps of the oxidation process^{55,103–105}. These works have been particularly useful in mapping out the energy profile of the reaction, unveiling the relative stability of intermediates and the energy barriers between them. These models have shown that the formation of the luciferyl-adenylate intermediate is one of the key energy-consuming steps^{102,106}, but once this complex is formed, the reaction proceeds efficiently toward the final product, excited state oxyluciferin^{55,56,103–105}. In addition, density functional theory (DFT) calculations (see Chapter 2.5) brought to the identification of the electronic transitions responsible for the colour variation of the emitted light^{54,107,108}.

In coelenterazine-based systems, luciferin is directly oxidized by oxygen, in presence of luciferase, to produce blue light. Computational models have been instrumental in making inferences about still unknown reaction mechanisms⁴⁴ and in understanding the role of calcium ions in modulating reaction kinetics and light-emission efficiency of aequorin. Indeed, it was suggested that calcium binding may induce conformational changes in the luciferase enzyme, leading to variations in the H-bonds pattern and being the trigger for emission¹⁰⁹.

Another important aspect of using computational chemistry in modeling bioluminescent reaction mechanisms is related to the possibility to probe different reaction pathways and explore different protonation states options for the product. For example, firefly luciferase has been shown to emit light in both the yellow-green and red spectrum, depending on the exact conditions of the reaction, in terms of temperature and pH. Computational models have helped to explain the molecular basis for this dual emission, showing that changes in the protonation state of oxyluciferin can lead to different electronic transitions and different maximum emission wavelengths^{54,108}.

1.3.2 Luciferin and luciferase rational design

A deeper understanding of the reaction mechanism underlying light emission, in terms of both the oxidation process and the electronic transitions involved in the excited state product decay, is invaluable, since it opens the door for the rational design of luciferin and luciferases. Rational design refers to systematically changing a molecule structure to achieve specific properties, such as greater stability, altered substrate specificity, or improved catalytic efficiency. This approach is popular not only in the bioluminescence field, but whenever an enzyme is involved¹¹⁰. Few enzymes' reaction are already optimized for industrially convenient conditions or for specific purposes and enzyme optimization is a field bustling with new studies and developments^{111–113}. In bioluminescence, rational design is applied to both, luciferin and luciferase.

The former is the light-emitting molecule in the bioluminescent reaction, and its structure plays a critical role in determining the wavelength and intensity of the emitted light, especially in coelenterazine-based systems, where the protein environment seems to play only a marginal role in shifting the emission⁸⁹. For example, in BLI, it is quite desirable to work with luciferins that emit in the red or near-infrared range, since these radiations are associated with higher penetration properties in biological tissues than higher energy photons, like the ones of blue or green light^{69,74}. Computational chemistry provides the tools needed to predict how luciferin structure modifications will affect its electronic properties and, consequently, the colour of the corresponding emission¹⁰². One of the main challenges in designing new luciferins is ensuring that the modified molecule remains compatible with the luciferase enzyme and computational docking studies are useful tools to predict whether the modified substrate will fit within the enzyme active site. Docking calculations are often followed by molecular dynamics (MD) simulations to assess the luciferin-luciferase complex behaviour over time, providing insights into the stability of the system¹¹⁴.

Besides designing new luciferins, computational chemistry plays a crucial role in the rational design of luciferase enzymes too, as luciferase structure determines the efficiency, speed, and specificity of the reaction. Altering the amino acid sequence of luciferase can enhance its catalytic activity or change its substrate specificity¹¹⁵. For example, mutations within the enzyme firefly luciferase have extended its thermal stability, rendering the enzyme more suitable for industrial applications that require increased temperatures¹¹⁶.

Before experimental verification of changes in property of interest, it is possible, with protein design techniques, such as homology modeling and MD simulations, to predict how specific mutations will affect the structure and function of luciferase. Homology modeling allows building a three-dimensional model of a protein basing on a related and already known one, which means that mutants of the luciferase to be redesigned can be easily obtained, as the 3D structure can be reproduced with ease¹¹⁷. As it is done with docking results, also in this case, MD simulations can be useful to provide insights into how the enzyme structure changes over time, i.e. verifying its stability or structural modifications. The so-obtained models can then be used to identify key amino acids that are involved in substrate binding or catalysis, making it possible to target them for mutagenesis⁹⁰.

1.4 NanoLuc-furimazine pair

Rational design is a powerful tool to obtain better fitting bioluminescent systems for a specific application, but it requires a detailed and accurate knowledge of the target luciferin-luciferase pair. This kind of deep insight can only be the result of years of research, employing both experimental and computational methods, an effort which is typically oriented towards potentially ground-breaking innovations, like the ones involving the NanoLuc-furimazine pair, which is the object of this section and has been the system of interest of the entire project.

1.4.1 The enzyme - from OLuc to NanoLuc

In 1978, a coelenterazine-oxidizing luciferase was isolated for the first time, from the deep-sea shrimp *Oplophorus gracilirostris*, by Osamu Shimomura¹⁴. This luciferase, also known as OLuc from the name of the organism itself, is part of the *Oplophorus* defensive strategy, which exploits it as the key component of the glowing cloud spit in front of eventual predators to disorient and confuse them. OLuc is an ATP-independent luciferase which catalyses the oxidation of coelenterazine to coelenteramide leading to the emission of blue light ($\lambda_{\text{max}} = 454 \text{ nm}$) like many marine luciferases^{24,89}. The quaternary structure of OLuc was found to be a complex of 35 kDa and 19 kDa proteins, bound in a non-covalent manner⁴⁷. In particular, the luciferase is composed of two subunits each of the two proteins. After cloning each of the fragments singularly, the oxidising activity was found to be an exclusive characteristic of the 19 kDa subunit (known as KAZ⁴⁷ or Oluc-19)¹¹⁸. This, together with the OLuc improved luminescence intensity and quantum yield if compared to *Renilla* luciferase and aequorin, was the spark which ignited further studies, oriented at the optimization of this subunit to recover the stability of the original complex towards changing in temperature and pH of the environment. Despite having no crystal structure for Oluc-19, Hall et al. were able to find similarities between it and the well-characterized intracellular lipid binding proteins (iLBPs) with fold-recognition¹¹⁹. This was a turning point, since those proteins are known to be stable and show a strongly conserved structural motif, with an Arg or Lys residue forming an hydrogen bond and packing against a Trp¹²⁰. This peculiar motif is not perfectly reproduced in Oluc-19, since the Arg/Lys residue is substituted by an Asn, and the punctual mutation N166R, to reintroduce the desired pattern, gave the expected results: the mutant was 0.5 times more stable and 3 times brighter¹²¹. After this step was achieved, since neither the detailed characteristics of the enzyme structure nor the reaction mechanism were known, it was not possible to adopt rational design to improve the performance of the newly synthesized luciferase, so directed evolution was the choice. Directed evolution is a powerful tool for protein engineering which can be described as a fast-forward version of natural evolution. It emerged in 1967 as a possible way to optimize enzymes without having to know the sequence-structure-function of the protein itself, by mimicking natural evolution¹²². Directed evolution consists of the generation of a library of variants of the protein of interest and then a selection of the best performing one with respect to a specific desired characteristic¹²³.

Oluc-N166R, the best performing variant of Oluc-19 was taken as the starting point for a directed evolution approach which first led to the variant C1A4E, already 29000 times brighter than Oluc-N166R in HEK293 cell lysates due to the increased production paired with a better thermal stability at 37°C¹²¹. C1A4E did not satisfy all the requirements for an ideal version of the luciferase itself, since it was strongly prone to aggregation, as was demonstrated with gel permeation chromatography. While proceeding with another round of directed evolution aimed at obtaining a monomeric luciferase without tendency to aggregation, Hall et al. decided to consider also alternative substrates to coelenterazine and opted to screen each of the best performing C1A4E mutants with them, for brighter luminescence. The best performing pair got to be a mutated version of C1A4E working with a coelenterazine analogue: 2-furanylmethyl-deoxy-coelenterazine, named furimazine. Once it was demonstrated that it was more stable in the cell culture media and displaying lower autoluminescence than coelenterazine, a final round of directed evolution was performed using it as a substrate.

The final result was a Oluc-19 mutant displaying 16-amino acid mutations if compared to the natural protein and consisting of a 10 stranded β -barrel capped by several α -helices which was called NanoLuc (also known as Nluc, Figure 1.14)¹²¹.

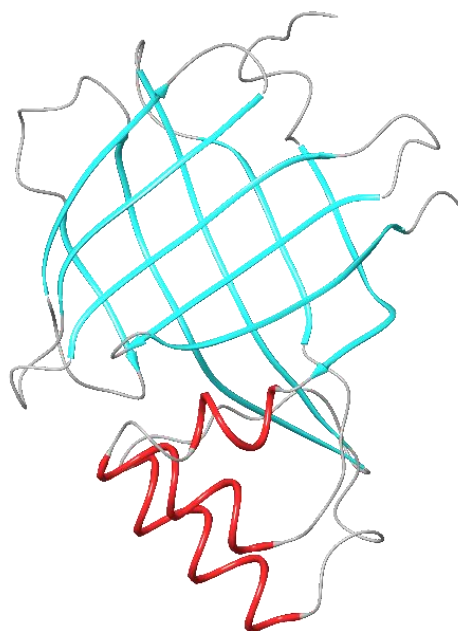


Figure 1. 14 - NanoLuc crystal structure (PDB: 5IBO)

1.4.2 The substrate - furimazine

The newly synthesized luciferase was able to produce light with the naturally occurring substrate coelenterazine, but emitted light was 25 times dimmer than the signal detected when using furimazine, a small molecule which is very similar to coelenterazine, since they both present an imidazopyrazinone core with aryl substituents in positions 2, 6 and 8 (Figure 1.15). The main difference is the modification of two of the peripheral groups, the two p-hydroxyphenilic moieties of coelenterazine in position 2 and 6, which have been modified into a furan ring and a simpler phenyl one, respectively. Those modifications led to a change in the molecular properties, in particular, furimazine was found to be three times less prone to autoluminescence and eight times more stable in cell culture media²⁴. These desirable properties, together with the improved bioluminescence that NanoLuc displayed when working on this new luciferin made furimazine the new standard substrate, despite being a purely synthetic molecule with a higher production cost if compared to coelenterazine¹²⁴.

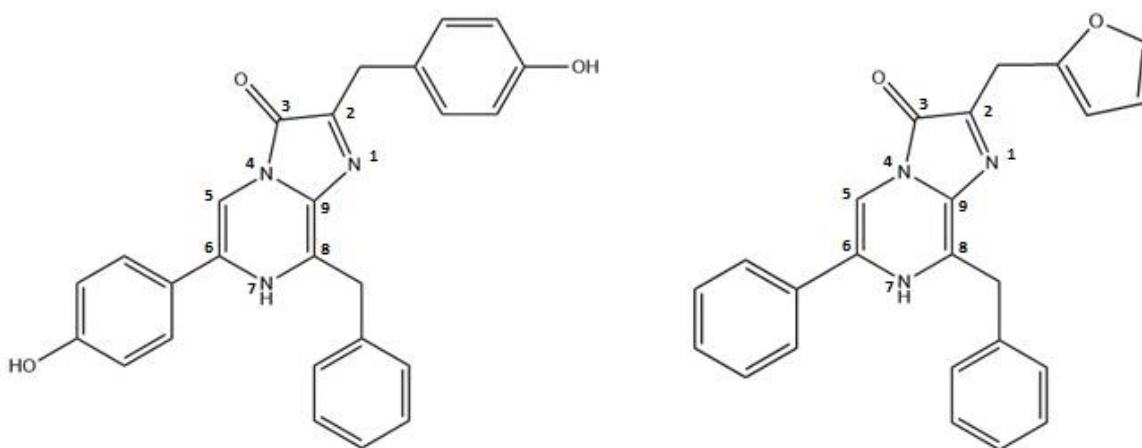


Figure 1. 15 - Sketches of coelenterazine (left) and furimazine (right)

1.4.3 Bioluminescent properties

NanoLuc-furimazine pair has several unique bioluminescent properties that set it apart from the other luciferin-luciferase systems and the main one is brightness. The intensity of the light output recorded from cell lysates is 150 times higher than the corresponding value for either FLuc or RLuc on a per mole basis. Two orders of magnitude of difference is too much to be explained only with an higher quantum yield, since RLuc^{90,125} and FLuc⁵⁴, especially, have both quite good values of it. As a consequence, improved brightness is probably related to an increase in catalytic turnover¹²¹.

As for almost all the coelenterazine-based luciferases, both natural and engineered, even NanoLuc-furimazine light emission consists of blue photons, with a maximum emission wavelength of approximately 460 nm, thus it is not directly suitable for BLI *in vivo*, but the emission spectrum is 20% narrower than the one of either RLuc, FLuc or click beetle red luciferase (CBR). The width of the peak is a crucial feature in the design of multiplexed luciferase assays, since it is often difficult to find two emitting systems with non-overlapping emission spectra and the narrowness of the NanoLuc-furimazine one makes it suitable for applications involving multiplexing, for example if involving CBR or similar luciferases, with a $\lambda_{\text{max}} > 600$ nm.

NanoLuc-furimazine performances have been tested under different conditions to furtherly check the improved stability of the enzyme and where does it place itself with respect to the popularly used FLuc and results were stunning. As reported in Figure 1.16 NanoLuc showed a better thermal resistance, since its activity is maintained even after 30 minutes of heating at 50°C, while FLuc started to lose it already at 30°C. Dependency on pH of the buffer was investigated too and while FLuc showed to be efficient only at pH higher than 8, while NanoLuc retained its optimal performance up to pH = 6.5 and displayed a non-negligible residual activity even at lower values of pH. Similarly, urea sensitivity was measured and NanoLuc came out again as the best performer, its bioluminescent activity being unaltered even at high urea concentrations (8 M), while FLuc started losing it at 2 M. The effects of an increasing concentration of NaCl instead, gave similar results for the two luciferin-luciferases pairs, with a better performance of NanoLuc only at extreme conditions ([NaCl] = 10 M)¹²¹.

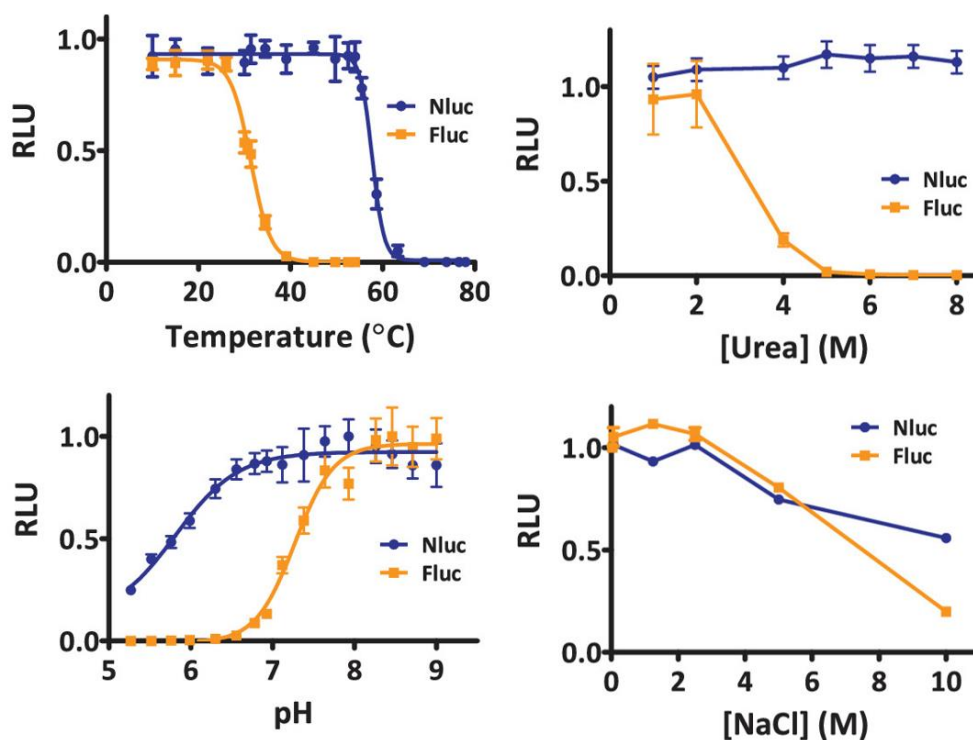


Figure 1. 16 - Comparison between NLuc and FLuc sensitivity to various conditions. Image taken from reference¹²¹

1.4.4 Applications

Looking at the astonishing bioluminescent properties of NanoLuc it is easy and safe to say that it is one of the most interesting bioluminescent systems on the market and this is confirmed by the vast number of possible applications of this technology or derived ones, which space from blue-sky pure research to applied biotechnologies.

1.4.4.1 Bioluminescence Imaging (BLI)

Probably the most important field of NanoLuc application is in bioimaging, although the λ_{max} represents the greatest limiting factor for the incredible potential of this system. Traditionally, FLuc is the reference luciferase for BLI, due to its emission spectrum having its peak at 560 nm, allowing for centimetres deep imaging in tissues. Despite that, Stacer et al. demonstrated that NanoLuc could be directly used for BLI *in vivo* in superficial tissues by testing it in mice¹²⁶. The advantages of employing a small luciferase like NanoLuc for BLI are mainly related to the behaviour of some viral agents, which are able to detect relevant modifications to their genome, thus making it impossible for traditional reporters' sequences to be inserted. The superiority of smaller reporters, like this 19 kDa luciferase, was demonstrated with the first *in vivo* tracking of influenza infection dynamics in a murine model¹²⁷.

1.4.4.2 Split complementation sensors

Ligand-receptor interactions are the core of molecular biology and assays to assess it play a major role in delving into it. Traditional assays are effective but are carried out with the use of radioactive compounds, thus being dangerous and sometimes limiting for the short half-life of many radionuclides. Significant efforts have been made in order to design nonradioactive-based assays and bioluminescent tracers emerged as great potential alternatives, especially in terms of split luciferase complementation assays. These are designed to allow the study of protein-protein interactions without hindering them, which is not an easy task since they can be influenced by microenvironmental surroundings, other small molecules or structural modifications. A SLC assay is based on two complementary luciferase fragments, which are unable to catalyse the reaction when separated, but recover their activity when brought close to each other. These fragments are thus fused to the two molecular structures whose interaction has to be monitored and, after luciferin addition, the luminescence provides instantaneous quantitative information about the process of interest¹²⁸. NanoLuc-based NanoBiT (NanoLuc Binary Technology) is a luminous example of this technology (Figure 1.17). A 13 amino acids fragment, corresponding essentially to the tenth strand of the secondary structure, was cut from the luciferase structure and both the fragment and the remaining part of the luciferase were singularly optimized in order to maximize the preservation of the tertiary structure for the bigger block and minimize, but not remove, the affinity to it for the smaller part¹²⁸. The shorter polypeptide was furtherly tested to verify whether it was possible to reduce its length without altering the performances and, in the end, a 11 amino acid length fragment was found to be optimal. NanoBiT has shown so many favourable properties that it has rapidly become one of the most popular SLC assays on the market^{129–132}.

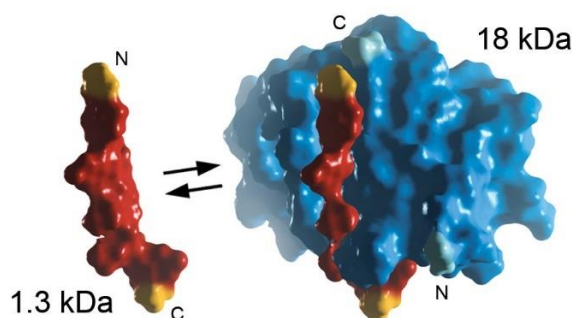


Figure 1. 17 - NanoBiT structure, smaller unit represented in red. Image taken from reference¹²⁸

1.4.4.3 Gene regulation sensor

NanoLuc has become widely employed also for confirming gene delivery. Reporter gene approaches consist in adding a sequence encoding for a protein conferring, on organisms expressing it, an easily detectable property to the gene sequence that is wanted to be delivered. If gene delivery is successful, not only the desired gene is expressed, but also the additional one, making it possible to monitor it by monitoring the detectable property. Luminescence is one of the most popular properties conferred by reporter genes and NanoLuc was shown to be an excellent gene regulation sensor *in vitro* as it allowed for validation of gene delivery even when other approaches lacked the necessary sensitivity¹³³. Monitoring of gene expression is pivotal in getting insights into viral infection too as it allows to follow the infection spread¹³⁴.

1.4.4.4 BRET-based systems

Similar to the SLC assays, ratiometric sensors allow the evaluation of protein-protein interactions and ligand-protein interactions. These sensors are almost always relying on BRET (process schematized in Figure 1.18), either bimolecular or intramolecular.

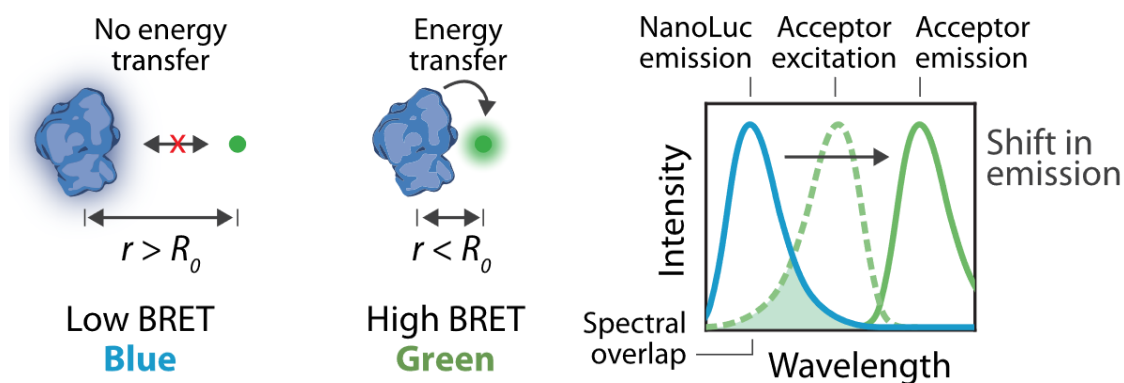


Figure 1. 18 - Schematic representation of BRET-based systems. R is the average distance between donor and acceptor and R_0 is the Förster distance. Image taken from reference¹³⁵

The former are realized by tagging a protein with NanoLuc and a ligand, which is supposed to be interacting with the protein, with a fluorescent dye with an absorption spectrum overlapping with the NanoLuc-furimazine emission one. Measuring the ratio between NanoLuc emission and the one associated with the fluorescent emission of the dye it is possible to evaluate the amount of interacting protein-ligand pairs. On the other hand, intramolecular BRET sensors are built up differently, with the acceptor directly linked to NanoLuc structure. A similar result can be achieved, for example, by genetically fusing NanoLuc and a fluorescent protein. As a consequence, the ratio between NanoLuc and BRET acceptor emissions is fixed and changes only upon structural modifications such as the cleavage of the linker. Such a behaviour can be exploited to monitor the concentration of a specific enzyme, as it was demonstrated by den Hamer et al. with a NanoLuc-mNeonGreen BRET pair¹³⁶. The two subunits have been connected with a linker containing caspase recognition sequences and caspase concentration could be measured basing on the detected decrease in BRET signal intensity, which was due to the cleavage of the linker by the enzyme action. A similar scenario is present in BRET-beacons employed for the detection of small concentrations of nucleic acids. In this case, NanoLuc is fused to a single stranded DNA (ssDNA) and a complementary ssDNA, displaying an internal stem-loop was fused to a BRET acceptor unit. Once the interaction is established, the BRET signal can be detected. In presence of a nucleic acid strand called input, the DNA structure is altered, with the loss of the internal stem-loop motif, resulting in a decrease in BRET output. These BRET-based applications of NanoLuc are all inspired by previously designed analogous FRET-based technologies. Each of them is suitable for different scenarios, with FRET preferred for higher absolute brightness whenever excitation is not too difficult to achieve and BRET being appreciated for the lower Limit of Detection (LOD)¹³⁵.

1.4.4.5 Photodynamic therapy (PDT)

As photodynamic therapy has become more and more popular in biomedicine, NanoLuc has increasingly been used as a trigger for photosensitizers. PDT consists in employing a photosensitizer to generate ROS in target tissues, resulting in cell death. It is an approach typically oriented towards cancer treatment, especially for superficial tissues, since the photosensitizer must be excited by external radiation¹³⁷. Delivering the excitation to target tissues is not an easy task and a possible solution to effectively excite the photosensitizers is the application of quantum dots (QD) BRET-based systems¹³⁸, but neither the delivery of these conjugates is simple and the QD safety is still debatable¹³⁹. Precisely for this reason, using only luciferin-luciferase pairs could be a gamechanger: with bioluminescence as a trigger, external light supply will no more be necessary, making it possible to target also deeper tissues and avoiding QD the safety issues can be bypassed. The main limitations of luciferin-luciferase pairs as light source to excite the photosensitizers are the same encountered in designing BRET-based pairs: spectral overlap between donor emission and acceptor excitation and brightness of the donor are required. NanoLuc happens to be not only characterized by a really strong emission, but its emission spectra is almost perfectly matching the absorption wavelength of a popular photosensitizer, miniSOG¹⁴⁰, making it a perfect candidate for PDT when paired with it¹⁴¹. This is confirmed by the fact that the phototoxicity of the miniSOG sensitizer excited by NanoLuc-Furimazine emission was shown to be comparable with the one observed for the LED-induced process¹⁴². This resulted in obtaining a PDT agent which is completely genetically encodable, a non-trivial aspect, since it allows for its delivery to any target tissue.

1.4.4 Open challenges and unanswered questions

Although being a possible solution for in vivo imaging in some specific situations, NanoLuc-furimazine pair emission is still too blue-shifted to be employed in mammalian tissues, especially if not superficial. Tuning the emission spectrum is a non-trivial challenge which has been tackled mainly by either redesigning the substrate with a trial and error approach^{143,144} or building BRET-based pairs^{98,136,145–147}. No rational design for substrate or luciferase has been possible up to now due to the fact that the reaction mechanism is still unclear, and the protein residues involved in it are still to be identified. The current knowledge is limited to the general concept that molecular oxygen reacts with furimazine, forming an excited state product that decays to emit light. This mechanism has yet to be fully elucidated for furimazine, either in the isolated enzyme or within cellular environments, but has already been tackled for other luciferases with the same imidazopyrazinone core⁴⁵. The oxidation of imidazopyrazinone substrates involves a complex multi-step reaction mechanism which can be broken down into several key steps. First, the molecule is involved in a deprotonation, necessary to activate the substrate itself, and then in the oxygenation, leading to the formation of a peroxy intermediate. Previous studies suggest this step requires a single electron transfer from the anionic luciferin to oxygen, forming a superoxide anion which then reacts with the luciferin itself to form the peroxide^{45,56}. The prosecution of the reaction after the formation of the peroxide is still unknown, especially in the protein environment, thus meaning that the emitting species have not been identified and the activation of the substrate is triggered by a similarly unknown base.

1.5 Aim of the project

This thesis' work is an attempt to fill in this knowledge gap with the help of computational methods, aiming at helping to get a deeper understanding of how does this unique luciferase work, in particular identifying the chemical species involved in the bioluminescent process and the underlying reaction mechanisms key steps, making it possible to furtherly explore and exploit its enormous potential.

2. Computational methods

In this chapter, the employed computational methods are introduced. In particular, in sections 2.1 to 2.3 the Molecular Mechanics and Molecular Dynamics approaches, and crucial concept of Phase Space, are described basing on the contents of Jensen's book¹⁴⁸ and other quoted articles^{149–157} and monographies¹⁵⁸. Section 2.4 is then dedicated to docking methods, with a focus on the induced fit docking protocol. In the following sections, i.e. 2.5 and 2.6, ab initio methods are introduced, with the main reference for the mathematical equations being Cramer's book¹⁵⁸, Jensen's one¹⁴⁸ and Engel and Dreizler's one¹⁵⁹, and density functional theory cardinal principles enounced. Section 2.7 revolves around the concept of hybrid approaches, based on both quantum mechanical and molecular mechanics calculations, with a particular attention on the ONIOM subtractive scheme. Finally, section 2.8 introduces the protocols used for the research activity, with the corresponding computational details reported in Appendix 2.

2.1 Molecular mechanics (MM)

The molecular mechanics method can be considered as a direct product of classical chemistry and the simple atom and bond representation of molecules. In the fifties, plastic models of complex molecular systems started to be built, and their computer representations are still used nowadays. The MM approach, which is commonly applied for the evaluation of complex molecular systems, is based on the laws of classical mechanics and can be seen as the mathematical translation (and refinement) of the ball-and-stick visualization of molecules. In this methodology, in contrast to what happens in quantum mechanical (QM) methods, electrons are not explicitly considered¹⁴⁸. They are responsible for the shape and properties of the potential energy surface (PES), where the nuclei move, but, in MM, the PES is not defined on the basis of the electronic energy, as it is in ab initio methods where it is calculated by solving the Schrödinger equation. Instead, it is described by means of a set of functions, derived empirically, which have the mathematical form typical of classical mechanics laws¹⁴⁹. Since it is a classical-based approach, the MM method is particularly advantageous in the study of complex molecular systems which are closer to the macroscopic world than small molecules. A molecule is described as a set of bond lengths, bond angles and torsional angles and its energy depends on the geometric deformations described in terms of these coordinates with the addition of a non-bonding (van der Waals and electrostatic) contribution representing the forces acting between unbound atoms. The fundamental idea behind molecular mechanics is that there are *natural* values of the geometric parameters, which would characterize the molecule in absence of non-bonding interactions between the various atom groups, but these are not displayed in real molecular systems since each atom interacts with all the others leading to the molecule *strain*, i.e. values of the geometric parameters are deviating from the natural ones. The MM method assumes that the energy associated with these deformations can be calculated. The set of functions used to describe the PES and the empirical parameters appearing in the functions is called force field (ff). The parameters used in the functions to calculate the potential energy are modifiable ones; they are optimized to reproduce a range of experimental or calculated molecular properties, such as geometries, conformational energies, heats of formation and vibrational frequencies. The fundamental assumption of MM is that the parameters optimized for a certain molecular fragment, e.g. the value of the C-N single bond distance, are transferable, for the same fragment, from one molecule to another; in other words, the parameters are obtained from a reduced set of simple molecules and then used in calculations involving larger, more complex molecular systems¹⁴⁸. The hypothesis of parameter transferability cannot be proven *a priori*, but is confirmed by the results obtained, which are generally in good agreement with the experimental evidence, justifying the hypothesis that individual molecular fragments have similar properties in both simple and complex systems. Moreover, the transferability of parameters allows the extension of ffs, starting from an already optimised parameter set, to include new molecular fragments or new classes of molecules. In MM, a single standard transferable value, the natural value, corresponds to a certain type of bond, bond angle or dihedral

and the geometry of equilibrium that is found by relaxing the molecule to its energy minimum. Each coordinate (bond distances, bond angles, torsional angles and non-bond distances) tries to assume its natural value: the equilibrium geometry is derived from the balance of forces associated with each coordinate. In the MM method, the molecule is thus seen as a mechanical model in which the atoms are represented by Newtonian point particles bound together by springs (bonds) whose behaviour is described by Hooke's generalised law.

When a molecule consisting of n atoms and defined in terms of $3n$ coordinates x_i undergoes a deformation with respect to its reference geometry, $\{x_i^0\}$, corresponding to a potential energy minimum, V_0 , its potential energy can be written as a Taylor expansion:

$$V_{\text{pot}} = V_0 + \sum_i \left(\frac{\partial V}{\partial x_i} \right)_0 \Delta x_i + \frac{1}{2} \sum_{i,j} \left(\frac{\partial^2 V}{\partial x_i \partial x_j} \right)_0 \Delta x_i \Delta x_j + \dots \dots \dots \quad (2.1.1)$$

The term V_0 is a constant for a particular molecule and can be regarded to as a reference value, which is equivalent to setting it to zero. The first derivative of V , calculated for the equilibrium geometry, is null by definition, since the equilibrium geometry corresponds to a minimum in the PES. Furthermore, by considering small rearrangements, terms of order higher than the second can also be neglected (harmonic approximation). In the first approximation, therefore, the potential energy will only depend on the third term of the expansion, which is the second derivative of V with respect to the coordinates $\{x_i\}$. By replacing the expression of the second derivatives, which are the force constants, with the symbol f_{ij} , we obtain the relationship corresponding to a simple harmonic force field:

$$V_{\text{pot}} = \frac{1}{2} \sum_{i,j}^{3n} f_{ij} \Delta x_i \Delta x_j \quad (2.1.2)$$

This equation exactly defines, within the harmonic approximation, a system of coupled harmonic oscillators. The force constants are typically represented as a matrix in which the diagonal terms correspond to $i = j$. If all non-diagonal terms are zero, then the oscillator system is completely decoupled, which simplifies Hooke's law:

$$V_{\text{pot}} = \frac{1}{2} \sum_i^{3n} f_{ii} \Delta x_i^2 \quad (2.1.3)$$

The harmonic approximation and the assumption of decoupled oscillators are at the basis of the development of the MM method. It must be kept in mind that, in order to have a better description of the nuclear motions, and thus obtain high-quality equilibrium geometries, it is necessary to add mixed terms to the harmonic equations, i.e. those outside the diagonal of the matrix. Furthermore, it has been verified that the inclusion of mixed terms is a necessary condition for the development of transferable force fields. The harmonic approximation is strictly valid only for small displacements relative to the reference position, otherwise terms of higher than second order must also be included in the potential energy expression in order to more satisfactorily take anharmonic effects into account. In MM, the energy of the molecule is defined as steric energy, E , given by the sum of M different potential energy functions, V , each dependent on the value of N geometric coordinates of i -th type (bond lengths, bond angles, torsions...), q_{ik} :

$$\begin{aligned} E_s &= \sum_i^M \sum_k^{N_i} V_i(q_{ik}) = \\ &= \sum_{\text{bond}} V_{\text{stretching}}(q_i) + \sum_{\text{angle}} V_{\text{bending}}(q_j) + \sum_{\text{torsion}} V_{\text{torsional}}(q_k) + \\ &\quad + V_{\text{Van der Waals}} + V_{\text{el}} + V_{\text{cross}} \end{aligned} \quad (2.1.4)$$

Within the harmonic approximation, a generic potential function V is expressed by the generalized Hooke's law:

$$V_i(q_{ik}) = \sum_k^N \frac{1}{2} K_{ik} (q_{ik} - q_{ik}^0)^2 \quad (2.1.5)$$

Where K_{ik} are the force constants, q_{ik} is the value of the k -th geometric coordinate and q_{ik}^0 is the natural value the coordinate would take if there were no deformation.

The interactions which are generally considered in the potential energy functions are:

- 1-2 interactions (bond lengths, stretching)
- Interactions 1-3 (bond angles, bending)
- Interactions 1-4 (dihedral angles between bond pairs, twists)
- Interactions between non-bonded atoms, or between atoms separated by more than two bonds. They generally include an electrostatic contribution and a Van der Waals' one.

Steric energy is therefore given by:

$$E_s = V_{\text{stretching}} + V_{\text{bending}} + V_{\text{torsional}} + V_{\text{non-bonded}} + \text{other terms} \quad (2.1.6)$$

Among the “other terms”, the polar term V_{pol} and cross terms, such as the “stretch-bending” term $V_{\text{s-b}}$, may be particularly relevant.

2.1.1 Stretching and bending potential functions

For these two potential functions the oscillators are considered independent since the harmonic approximation is generally assumed as valid, so:

$$V_{\text{str}}(r) = \frac{1}{2} K_s (r - r_0)^2 \quad (2.1.7)$$

$$V_{\text{ben}}(\theta) = \frac{1}{2} K_b (\theta - \theta_0)^2 \quad (2.1.8)$$

In these two expressions, r and θ are the values assumed by a particular bonding distance or angle of bonding and r^0 and θ^0 are the corresponding natural values in the absence of deformation. As seen above, the harmonic approximation may not be sufficient for values of r and θ far from natural values, as can be clearly seen in Figure 2.1.

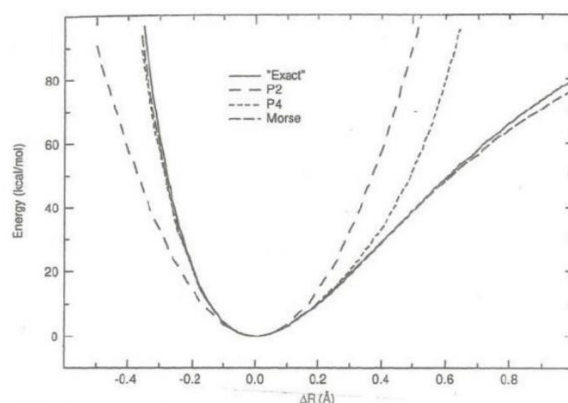


Figure 2. 1 - Comparison between exact energy curve and the profile of three different potential expressions to calculate it. Image taken from reference¹⁴⁸

In equations 2.1.7 and 2.1.8, r and θ are the values assumed by a particular bonding distance or bond angle and r_0 and θ_0 are the corresponding natural values in the absence of strains.

Thus, higher-order corrective terms were introduced with third- and sixth-degree functions for stretching and bending, respectively.

$$V_{\text{str}}(r) = \frac{1}{2}K_s(r - r_0)^2 + K'_s(r - r_0)^3 \quad (2.1.9)$$

$$V_{\text{ben}}(\theta) = \frac{1}{2}K_b(\theta - \theta_0)^2 + K'_b(\theta - \theta_0)^6 \quad (2.1.10)$$

In the case of stretching, the inclusion of a third-degree term leads to a better description of the potential energy curve for distances close to or less than r^0 . However, for a certain value of $r > r^0$, the function tends to negative values as r increases. The use of a function with a cubic term beyond a certain value of r would therefore lead to an unrealistic situation in which the atoms move away from each other. Consequently, the cubic correction is only used in the final stages of geometry optimisations, when values are closer to equilibrium. The use of functions of even degree, as in the case of bending, does not present any serious problems as it reflects the shape of the potential well. Since the deformation of a bond angle leads to a stretching in response to the strain, the introduction of a mixed term (the stretch-bending potential) is required:

$$V_{s-b} = \sum_k^N \frac{1}{2}K_{sb}(r - r_0)(\theta - \theta_0) \quad (2.1.11)$$

2.1.2 Torsional potential function

The second crucial contribution is given by the torsional potential function, which must be periodic: after a 360° rotation of the torsional angle, the potential must return to its initial value. Furthermore, torsional motion requires rather small energies compared to stretching and bending. Thus, the molecules may exhibit significant torsional distortions compared to the minimum. For this reason, it is not good practice to use a Taylor expansion for the torsional potential, but it is preferable to use a Fourier series expansion. For the torsional angle ABCD, the torsional potential is given by:

$$V_{\text{tors}}^{\text{ABCD}} = \sum_{n=1} k_n^{\text{ABCD}} \cos(n\omega) \quad (2.1.12)$$

In equation 2.1.12 $n=1$ describes a periodic rotation with period 360° ; $n=2$ is periodic with period 180° ; $n=3$ is periodic with period 120° and so on. The value of the constant k_n determines the height of the barrier: depending on the case, some constants may be 0.

A similar expression to the previous one is:

$$V_{\text{tors}}^{\text{ABCD}} = \frac{1}{2}V_1^{\text{ABCD}}[1 + \cos(\omega^{\text{ABCD}})] + \frac{1}{2}V_2^{\text{ABCD}}[1 - \cos(2\omega^{\text{ABCD}})] + \frac{1}{2}V_3^{\text{ABCD}}[1 + \cos(3\omega^{\text{ABCD}})] \quad (2.1.13)$$

The signs are chosen so as to have a minimum of V_1 at 180° , of V_2 at 0° and 180° , of V_3 at 60° , 180° and 300° , as shown in Figure 2.2.

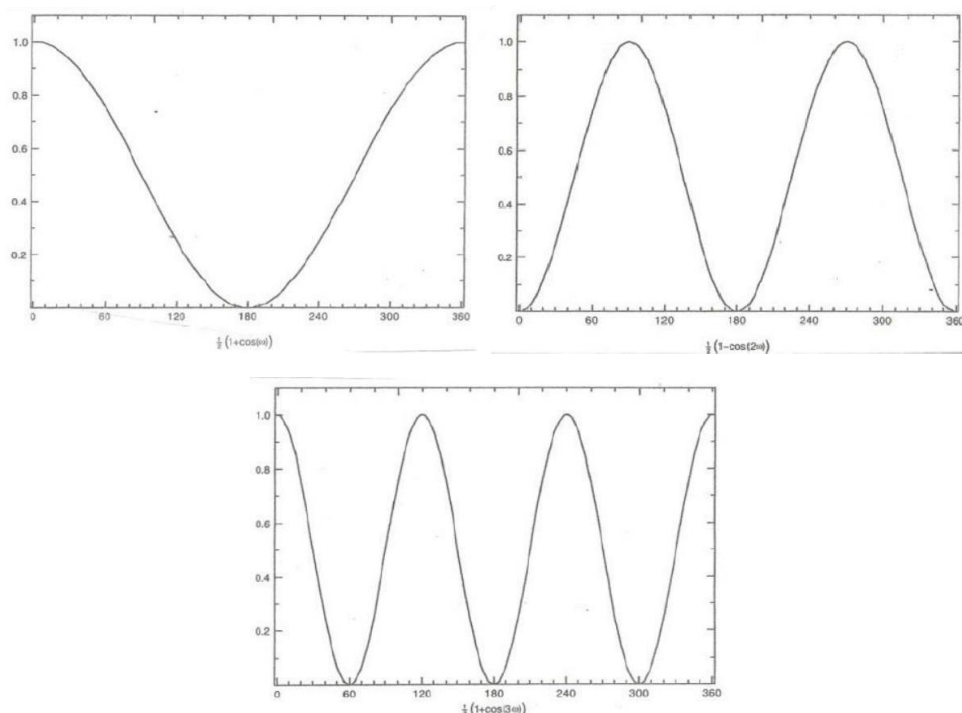


Figure 2. 2 - Different contributions of torsional potential functions. Image taken from reference¹⁴⁸

2.1.3 Non-bonding potential

The interaction energy between the unbound atoms is calculated as the sum of two contributions: the Van der Waals term, V_{vdW} , and the electrostatic term V_{el} .

2.1.3.1 Van der Waals potential

Non-bonding interactions greater than 1-3 are explicitly treated in the MM method. In general, it is assumed that the same laws used to describe intermolecular interactions are valid for intramolecular interactions too. The general form of any non-bonding potential function is given by the sum of two contributions: one repulsive, short-range active; one attractive, long-range active, tending to zero asymptotically as r increases. The former contribution, the repulsive one, derives from the repulsive force that is established between the electronic distributions of two sufficiently close atoms. This force is known as the “exchange force” or “overlap force” and is established between electrons with the same spin. The latter, the attractive contribution, results from molecular interactions between instantaneous dipoles, even if the interacting fragments do not have a permanent dipole moment. These forces, which in quantum mechanics are derived from electron correlation, are called “dispersive forces” or “London forces”. The two functions most frequently used to describe the Van der Waals potential are the Lennard-Jones (also called the 6-12 potential) and the Buckingham-Hill (which belongs to a more extended class of potential functions, also called the 6-exp potential) ones:

$$V_{vdw} = \varepsilon \left[\left(\frac{r^*}{r} \right)^{12} - 2 \left(\frac{r^*}{r} \right)^6 \right] \quad \text{Lennard-Jones (6-12 potential)} \quad (2.1.14)$$

$$V_{vdw} = \varepsilon \left[A \exp \left(B \frac{r^*}{r} \right) - C \left(\frac{r^*}{r} \right)^6 \right] \quad \text{Hill (6-exp potential)} \quad (2.1.15)$$

In the Hill potential, the parameters A, B and C are a function of the parameter α , which can be optimised, as are the parameters ε and r^* . In both cases, ε defines the depth of the potential hole, r^* is the minimum energy distance (associated with the Van der Waals radius of the interacting atomic pairs), while exponent 12 or parameter B defines the hardness of the potential, i.e. its depth for distances less than equilibrium, as shown in Figure 2.3.

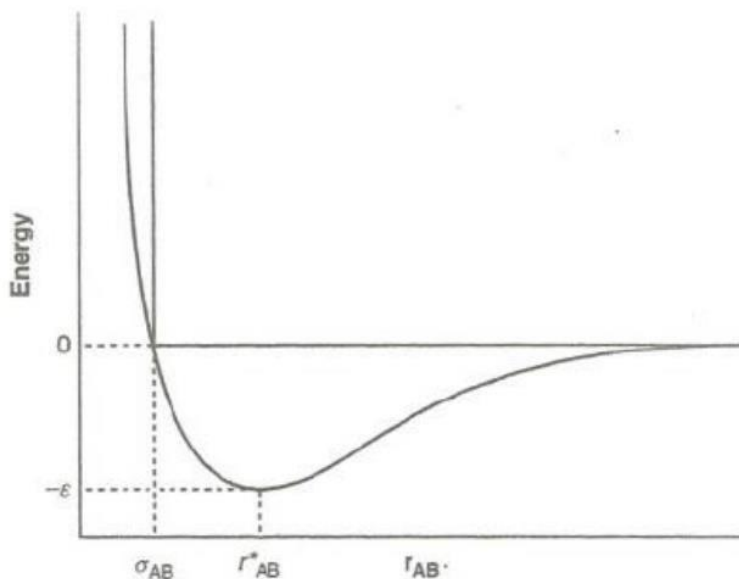


Figure 2. 3 - Lennard-Jones potential function profile. Image taken from reference¹⁴⁸

The Lennard-Jones potential is trending too strongly for short distances, whereas the Hill potential could become attractive in these regions (see Figure 2.4).

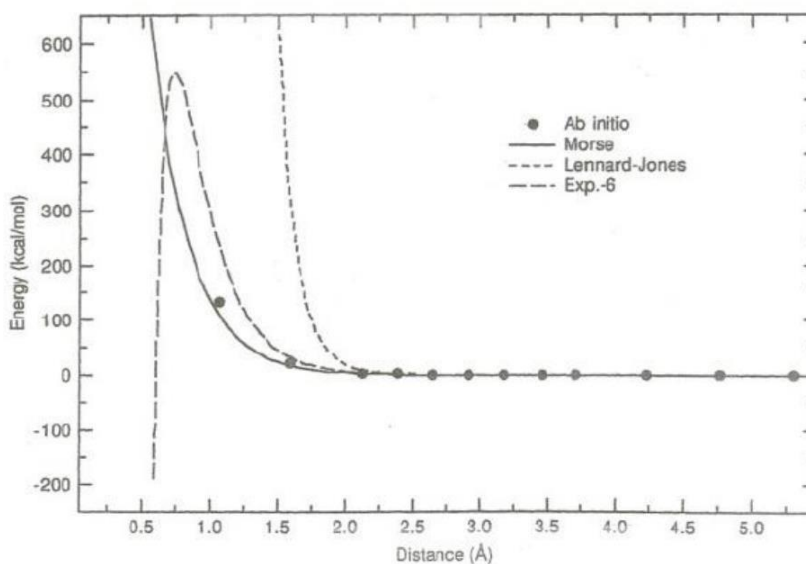


Figure 2. 4 - Comparison of different Van der Waals potential functions (solid and dashed lines) with respect to the *ab initio* results (points). Image taken from reference¹⁴⁸

The use of a pair potential means that the interaction between an atom of type A and one of type B will be described using the parameters ϵ_{AB} and r_{AB}^* obtained, by means of appropriate mixing rules, from the parameters for the interactions AA and BB. The rules generally adopted are as follows:

$$r_{AB}^* = \frac{r_{AA}^* + r_{BB}^*}{2} \quad (2.1.16)$$

$$\epsilon_{AB} = \sqrt{(\epsilon_{AA}\epsilon_{BB})} \quad (2.1.17)$$

In the MM method, the main problems for calculating non-bonding interactions arise from the fact that atoms are considered spherical in both the Lennard-Jones and Hill potential functions. Such a situation is certainly correct for isolated atoms of noble gases but may not be valid for atoms in a molecule that do not exhibit spherical symmetry. The approximation may be true for heavy atoms, but not for light ones such as oxygen or fluorine, which are essentially nonspherical. In the case of hydrogen, the situation is even more extreme because the electron density is strongly polarised along the bonding direction. Two simple methods have been introduced in some force fields to overcome this problem; in the case of hydrogen, the centre of the atoms for the calculation of the Van der Waals contribution is shifted about 10% along the bond; in the case of oxygen and nitrogen, the electron doublets are explicitly introduced as pseudo-atoms with their parameters and their non-bonding potential functions.

2.1.3.2 Electrostatic potential

In the molecular mechanics method, the electrostatic contribution is generally calculated by the interaction between the net atomic charges, using Coulomb's law:

$$V_{el} = \sum_{i,j} \frac{q_i q_j}{\epsilon r_{ij}} \quad (2.1.18)$$

where ϵ is the dielectric constant of the medium. Atomic charges in a molecule cannot be unambiguously defined or derived from experimental measurements, as they are not physical observables of the system. In molecular mechanics, atomic charges can be treated as parameters: their values can be determined as calculated properties by fitting them to experimental values. However, a generally adopted procedure involves assigning atomic charges in accordance with the values obtained from fitting the MEP calculated by ab initio methods. The value of the dielectric constant equals 1 for molecules in vacuum, while higher values can be used to represent the solvent effect. It is also possible to use a value of ϵ dependent on the distance of two interacting atoms to model the effect of electrostatic screening by solvent molecules with $\epsilon = \epsilon_0 r_{AB}$. The classical electrostatic contribution only contains the pair contributions. However, in polar species, three-body contributions can account for as much as 10-20% of the two-body contributions. These contributions can be modelled by including a polarisation term in the electrostatic potential.

Comparison of calculated results with experimental data must take into account that the latter are obtained at temperatures very different from 0 K and on systems consisting of a very large number (of the order of magnitude of Avogadro's number) of molecules, which may be mutually interacting (as in the condensed phase) or non-interacting (as in the dilute gas phase). In order to be able to properly compare the results of calculations with experimental data, it is therefore necessary to refer to a methodology that allows the calculation of thermodynamic properties of a set of macroscopic particles from the properties of individual molecules, i.e. Statistical Thermodynamics.

Statistical Thermodynamics is the bridge that connects the thermodynamic properties of a macroscopic system to the molecular properties of its microscopic constituents. In other words, Statistical Thermodynamics connects thermodynamics to quantum mechanics.

In principle, to determine the thermodynamic properties of a system, it is sufficient to solve the equations of motion of the system particles. However, this approach is unfeasible due to the large number of molecules constituting the macroscopic system. Alternatively, one can determine the probability that the molecules possess a certain energy value and, from this information, calculate the value of the macroscopic properties of the system. Statistical Thermodynamics is based on the principle that the thermodynamic observables of the system can be obtained as a weighted average of the value of the properties of individual molecules. The results obtained are very accurate for crystalline systems or in the case of ideal gases. In the case of disordered condensed phases (liquids or solutions), it is necessary to adopt different approaches that allow a real simulation of the system under investigation. These methods, known as molecular simulation methods, include the Metropolis Monte Carlo (MMC) and molecular dynamics methods.

2.2 The phase space

In general, a system consisting of many interacting particles, such as a protein system, can be described by the laws of quantum mechanics. The state of the system is described by the wavefunction $\Psi(R_N, r_m, s_m)$, which depends on the spatial and spin coordinates of m electrons and N nuclei. However, an even approximate solution of the Schrödinger wave equation (the eigenfunctions of which are $\Psi(R_N, r_m, s_m)$) is in practice very difficult to obtain for a system consisting of thousands of atoms, due to the very high computational cost of the methods for solving this equation for such a large system. It is therefore necessary to describe the system in terms of classical mechanics. Obviously, such an approach precludes the study of processes involving a significant redistribution of the electronic density, as in the case of reactive processes involving the breaking and formation of bonds, or in the case of systems at very low temperatures (0-10 K), where quantum effects become relevant.

From the classical point of view, the state of a system is completely defined when the position (r) and momentum ($p = mv$) of its individual particles are known. In principle, knowing the positions and momentum of all the particles at an initial instant, the positions and velocities of the particles at each subsequent instant can be calculated using the laws of classical mechanics. In this sense, the state of a system consisting of N particles is completely defined by $3N$ spatial coordinates r_i (which we denote by r^N) and $3N$ momentum coordinates p_i (p^N). The space defined by this set of variables is called “phase space”. The system will therefore be identified and described by the values assumed by these $6N$ coordinates. Statistical thermodynamics defines, for a system of interacting particles, the canonical partition function Q that contains all thermodynamic information about the system:

$$Q = \sum_i e^{-\beta E_i} \quad (2.2.1)$$

The sum is extended to all quantised states of the system, characterised by discrete values of energy E_i . If the system is classically described, the energy of the system is defined by:

$$E(r_1, \dots, r_N, p_1, \dots, p_N) = E(r^N, p^N) = K(p^N) + V(r^N) \quad (2.2.2)$$

The energy is then no longer quantized, but assumes a continuous spectrum of values, for each fixed set of coordinate values ($r_1, \dots, r_N, p_1, \dots, p_N$), representing a point in phase space. The total energy of the system is given by the sum of the kinetic energy $K(p^N)$, which depends on the linear momentum of the particles, and the potential energy $V(r^N)$, which depends on the position of the particles.

The canonical partition function Q for a system of indistinguishable particles takes the form:

$$Q = k \iint e^{-\beta E(r^N, p^N)} dr^N dp^N \quad (2.2.3)$$

Where k is a normalisation factor of $1/(N!h^{3N})$ and the double integral symbol is written instead of $6N$ integration signs since integration must be done with respect to $3N$ position variables and $3N$ linear momentum variables.

In the dominant configuration, the probability $P^*(r^N, p^N)$ of a state within the phase space will be given by the relation:

$$P^*(r^N, p^N) = e^{[-\beta E(r^N, p^N)]}/Q \quad (2.2.4)$$

The weighted average of property A , denoted by $\langle A \rangle$, will be the average of the values the property assumes in the various states of the system (and thus in phase space) weighted by the probability of the states in the dominant configuration:

$$\langle A \rangle = \iint A(r^N, p^N) P^*(r^N, p^N) dr^N, dp^N = \iint \frac{A(r^N, p^N) e^{[-\beta E(r^N, p^N)]}}{\iint e^{[-\beta E(r^N, p^N)]} dr^N, dp^N} dr^N, dp^N \quad (2.2.5)$$

The problem is in calculating these integrals. In general, the value of the system energy at each point in phase space should be calculated, but this procedure is unfeasible, both because the number of variables is too high and because it is not efficient. In fact, the calculation of the integral at the denominator (corresponding to Q) involves the generation of a large number of states (in principle, all of them) and thus also of all those high-energy states that consequently have a low probability and low weight in the calculation of $\langle A \rangle$.

2.3 Molecular simulation methods

Molecular simulation methods allow the calculation of energetic, structural, thermodynamic and dynamic properties of molecular systems through the generation of a set of states representative of the phase space accessible to the system. They are applied to the study of condensed phases, for the determination of properties of liquids or solutes in solution, and complex molecular systems, including proteins and used:

- to calculate structural and thermodynamic properties of systems
- as a tool for the conformational analysis of complex systems
- to refine the crystal structures of proteins and biological macromolecules
- to develop molecular models from structural information obtained from NMR spectroscopy (NOE, etc.)

Basically, there are two simulation methods: MMC and MD method. The former makes it possible to obtain information on the properties of a system at equilibrium; the latter also allows the study of its dynamic properties. The MD method finds application in the conformational analysis of complex systems, particularly biological macromolecules. This is because simulations are carried out at temperatures other than 0 K, so the system has not only potential energy, but also kinetic energy. This kinetic energy can allow potential energy barriers of the order of $k_B T$ to be overcome for each conformational degree of freedom. Note that the length of an MD simulation, which is the time for which the evolution of the system is followed, is generally 0.1-100 ns. This means that only phenomena taking place in a time range compatible with this time window can be observed.

2.3.1 Molecular Dynamics (MD)

In the molecular dynamics method, the value of a property is obtained as the time-averaged value that the property assumes in a single system that evolves for an ideally infinite time. Assuming we wish to determine the value of a macroscopic property of the system, it will depend on the position (r^N) and linear momentum (p^N) of the particles constituting the system. The instantaneous value of the property will therefore depend on the instantaneous values of these variables at time instant t , i.e.: $A(t) = A[r^N(t), p^N(t)]$.

Over time, the instantaneous value of the property will fluctuate due to interactions between the particles. The experimentally measured value will be the average over time of the instantaneous values, denoted by $\{A\}_t$. As the "observation time", τ , increases, the value of the time average will become closer and closer to the actual value of the property, denoted as A_{real} , becoming equal to A_{real} for an infinite observation time.

The average over time is given by the relation:

$$A_{\text{real}} = \{A\}_t = \lim_{\tau \rightarrow \infty} \frac{1}{\tau} \int_{t=0}^{t=\tau} A[r^N(t), p^N(t)] dt \quad (2.3.1)$$

In the ergodic hypothesis, the average over time and the average over systems coincide:

$$A_{\text{real}} = \{A\}_t = \lim_{\tau \rightarrow \infty} \frac{1}{\tau} \int_{t=0}^{t=\tau} A[r^N(t), p^N(t)] dt = \lim_{M \rightarrow \infty} \frac{1}{M} \sum_{i=1}^M A[r_i^N(t), p_i^N(t)] \quad (2.3.2)$$

This is for each set of initial condition values chosen.

To calculate the instantaneous values of $A[r^N(t), p^N(t)]$, the dynamic behaviour of the system must be simulated. To do this, the following steps must be followed:

1. Choice of a force field to describe intramolecular and intermolecular interactions of the system
2. Calculation, using the selected Force Field, of the potential energy value $V(r^N)$ for a given initial arrangement of molecules in space
3. Calculation of the forces acting on each atom of the system by differentiating the energy expression potential:

$$F_i = - \frac{\partial V(r^N)}{\partial r_i} \quad (2.3.3)$$

4. Calculation of acceleration, knowing the forces acting on each atom, from the Newton's equation of motion:

$$F_i = m a_i \quad (2.3.4)$$

5. Integration of the equations of motion for each particle, to determine changes in position, velocity and acceleration as a function of time. The integration is done by considering M time intervals, Δt , small enough to assume the particles' acceleration as constant. In this way, the trajectory is calculated from the laws of uniformly accelerated motion:

$$r(t) = r_0 + v_0 t + \frac{1}{2} a t^2 \quad (2.3.5)$$

Where r_0 and v_0 are initial position and velocity at time $t=0$.

6. Calculation of the properties of the system as the average over time of the values the property takes on in the considered M time intervals. Since the observation time is equal to the product $M \Delta t$, we can rewrite equation 2.1.24 as:

$$\{A\}_t = \frac{1}{M \Delta t} \sum_{i=1}^M A_i(r^N, p^N) \Delta t = \frac{1}{M} \sum_{i=1}^M A_i(r^N, p^N) \quad (2.3.6)$$

2.3.1.1 Integration of motion equations

The choice of interval length Δt for the integration of the motion equations depends on the phenomenon being investigated. For example, to study molecular vibrations, a suitable time scale must be defined. In general, the time interval must be chosen in such a way that it is one or two orders of magnitude smaller than the frequency of the fastest observable motion in the system. Typically, vibrations are the fastest motions: for a bond between an H atom and a heavy atom, these have a period of 10^{-14} s, corresponding to 3300 cm^{-1} .

For a system containing this type of bond, the time interval for the integration of the equations of motion should be no more than 0.1 fs. In order to increase the values of Δt , the degrees of freedom with higher frequencies can be removed from the system through the use of appropriate algorithms, such as, for example, the SHAKE algorithm¹⁵³, which allows the bond distances to be frozen. In this way, Δt can be increased by a factor of 2-3.

Different procedures for performing the integration of the equations of motion are available. An example of these procedures is the Verlet algorithm¹⁵⁴:

- acceleration is calculated at instant t
- the position at $(t + \Delta t)$ is calculated as a function of the position and acceleration at t and the position at $(t - \Delta t)$ in the following way:

$$r_i(t + \Delta t) = r_i(t) + v_i(t)\Delta t + \frac{1}{2} a_i(t)\Delta t^2 \quad (2.3.7)$$

$$r_i(t - \Delta t) = r_i(t) - v_i(t)\Delta t + \frac{1}{2} a_i(t)\Delta t^2 \quad (2.3.8)$$

Adding them up:

$$r_i(t + \Delta t) = 2r_i(t) - r_i(t - \Delta t) + a_i(t)\Delta t^2 \quad (2.3.9)$$

Subtracting them:

$$v_i(t) = \frac{r_i(t + \Delta t) - r_i(t - \Delta t)}{2\Delta t} \quad (2.3.10)$$

2.3.1.2 Initial conditions

Before starting a molecular dynamics simulation, we must define the conditions of our system. In other words, we must define the thermodynamic ensemble of the simulation, which can be different according to the quantities which have to be maintained constant during the simulation:

- Canonical ensemble (N, V, T)
- Microcanonical ensemble (N, V, E)
- Isothermal-Isobaric ensemble (N, P, T)
- Grand Canonical ensemble (μ , V, T)

Depending on the selected ensemble, the physical quantities that are to remain constant must remain controlled during the simulation. To this end, there are algorithms that allow these quantities to be scaled appropriately to ensure the predetermined conditions.

For example, the temperature is calculated at each step of the simulation from the kinetic energy of the system via the equation:

$$T = \frac{\sum_i m_i v_i^2}{(3N - N_c)k_b} \quad (2.3.11)$$

Where N is the number of atoms in the system and N_c is the number of degrees of freedom of the system that were frozen during the dynamics (typically $N_c = 3$, corresponding to the conditions to preserve the translational momentum of the entire system during the simulation). If the simulation takes place at constant T , at each step the particle velocity is scaled to keep it constant.

In order to ensure that the molecules at the boundaries of the system are not affected by the so-called “wall effect”, it is necessary to prevent the solvent molecules that form the solvation shell of the molecule of interest from “evaporating”. To do this, the molecules that form the system under investigation are placed in a “box” surrounded by a periodic series of identical boxes that exactly reproduce the original one, i.e. Periodic Boundary Conditions (PBC) are applied. In this way, molecules on or near the surface of the box interact with molecules in adjacent boxes.

Furthermore, to prevent one molecule from interacting with its copy, the minimum image convention is applied: each atom sees only a single copy of every other atom in the system, as shown in Figure 2.5.

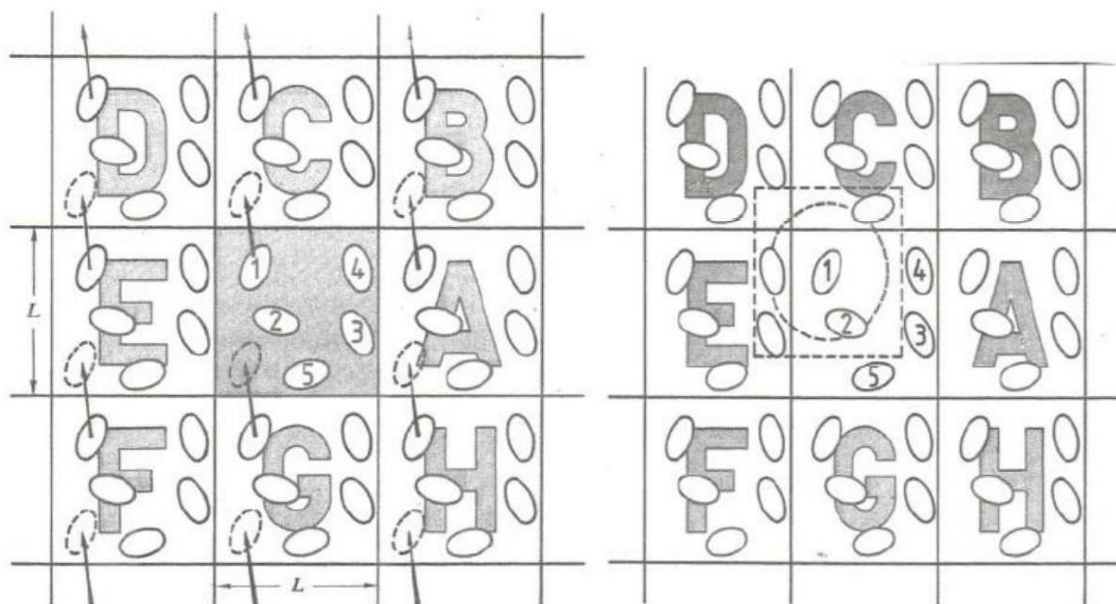


Figure 2. 5 - Schematization of how PBC and minimum image convention work. Image taken from reference¹⁴⁸

Furthermore, to increase the speed of the calculation of non-bonding interactions, a distance threshold is generally introduced above which the contribution of the non-bonding potential is considered to be absent. The threshold must be carefully chosen so that a particle does not see itself or see the same molecule twice. For this reason, it must be less than half the length of the cell. Two other things must be defined before starting a MD simulation: the initial configuration and the initial speeds. In general, it is preferable to start the simulation from an initial configuration that is as close as possible to the equilibrium situation that should be described by the simulation. For this purpose, the potential energy of the system is initially minimised so that there are no “hot points” or excessively distorted regions. Next, initial velocities are assigned to the particles, under the condition that they must be compatible with the desired temperature. They are randomly assigned to the atoms with one restriction: the total linear and angular momentum of the system must be zero, i.e. the system, as a whole, must neither rotate nor translate.

2.3.1.3 Equilibration and production phase

Starting from initial conditions, the system evolves over time and reaches equilibrium. If the simulation is carried out at constant temperature, the velocities are scaled until temperature T is reached. During equilibration, the values of various properties (E , K , V , T , P) characterising the simulation ensemble are monitored. When these quantities assume constant values, the production phase begins: the data collected in this phase is then used to calculate the properties of interest. If the simulation had sampled all points in phase space (ergodic trajectory), the results would be independent of the initial configuration. However, due to the boundless dimensions of phase space, it is not possible to obtain an ergodic trajectory from the simulation. In this sense, it is a good practice to repeat simulations from different initial configurations or by assigning different initial velocities, to improve the exploration of the phase space.

2.3.2 Analysis of simulations

In order to characterise the structural stability of the systems under investigation, different tools were used, in particular the Root Mean Square Deviation (RMSD) was assessed on the atoms of protein structures and on ligand ones, and the secondary structure of the protein was studied using the Dictionary of the Secondary Structure of Proteins (DSSP)¹⁵⁵ algorithm, as well as a cluster analysis of the trajectories, carried out using the GROMOS method. Furthermore, the Molecular Mechanics with Generalized Born and Surface Area solvation (MMGBSA) method was used to get another estimate of the binding energy of ligands.

2.3.2.1 Clustering methods

Clustering, or cluster analysis, a term introduced by Robert Tryon in 1939¹⁵⁰, is a set of multivariate data analysis techniques widely used in various fields, including data mining, pattern recognition, image analysis and bioinformatics. It consists of partitioning a dataset into subgroups with similar characteristics that are not known a priori, in order to obtain sets that are as homogeneous within themselves and as inhomogeneous between them as possible. This division is based on measures of similarity or dissimilarity that are considered in terms of distance in a multidimensional space. Clustering methods are subdivided according to different criteria, depending on the type of approach one can distinguish between aggregative and divisive methods, otherwise known as bottom-up and top-down methods. The former envisage that each element is initially considered as a cluster and, subsequently, these are joined until a predefined number of clusters are formed or until the distance between clusters exceeds a predefined reference value. In the case of top-down methods, instead, there is a progressive separation of a single starting cluster, which is divided into a predefined number of clusters or until a predefined stopping criterion is fulfilled¹⁵¹.

Another common classification of cluster analysis methods is that they are divided into partitional and hierarchical methods. Hierarchical methods find successive clusters from previously identified ones, partitional methods, on the other hand, determine all clusters simultaneously¹⁵⁶. A further possible categorisation depends on whether an element is simultaneously assigned to several clusters: if this is possible, one speaks of non-exclusive clustering methods (soft clustering), otherwise of exclusive clustering (hard clustering). In the first case, each element belongs to several clusters with different degrees of membership.

In order to analyse the trajectories obtained from the molecular dynamics simulations, it was decided to use the GROMOS hard clustering method¹⁵⁷, which evaluates the RMSD of the positions of the atoms of the system considered for each of the frames to be evaluated and proceeds to identify clusters from this. For each frame, the number of frames that are characterised by an RMSD lower than the selected threshold value is evaluated and the frame with the highest number of “neighbours” is considered the centre of a cluster. This, with its neighbours with which it forms a cluster, is removed and the remaining frames are subjected to the same process again. In this way, a series of non-overlapping clusters is obtained, as expected from hard clustering methodologies.

2.3.2.2 Define Secondary Structure of Proteins (DSSP)

For an effective understanding of the relationship between the amino acid sequence of a protein and its structure, the correct and unambiguous definition of its secondary structure is essential. To achieve this, one of the few objective assignment algorithms was used, namely the one proposed by Kabsch and Sander¹⁵⁵. This is a pattern-recognition algorithm based on the sequence of hydrogen bonds that are detected in the structure under investigation.

According to this algorithm, two types of H-bonds can be distinguished:

- "n-turns", with a hydrogen bond between the carbonyl of the i-th residue and the NH group of the (i+n)-th residue, where n is between 3 and 5.
- "bridges", with a H-bond between residues not close to each other in the amino acid sequence. This distinction makes it possible to analyse all the interactions that characterise the backbone of a protein structure. Repeated 4-turns define an α -helix, repeated bridges identify a β -helix. Other combinations are associated with 3-10 helices, rather than individual turns or bridges.

Hydrogen bonds in proteins have little overlap of wavefunctions and are well described by an electrostatic model. The electrostatic interaction energy between two H-bonded groups is calculated by placing partial charges on the carbonyl (q_1 on C, $-q_1$ on O) and NH group (q_2 on H, $-q_2$ on N).

$$E = q_1 q_2 \left(\frac{1}{r_{(O-N)}} + \frac{1}{r_{(C-H)}} - \frac{1}{r_{(O-H)}} - \frac{1}{r_{(C-N)}} \right) f \quad (2.3.12)$$

With $q_1 = 0.42e$, $q_2 = 0.20e$ and $r_{(A-B)}$ the interatomic distance between the two atoms A and B. In atomic units, r is in angstroms, the size factor f is 332 and the energy is expressed in kcal/mol. A good hydrogen bond is associated with a bond energy of about -3 kcal/mol, but in the algorithm proposed by Kabsch and Sander, a value of only -0.5 kcal/mol was chosen as the energy cutoff for recognising the presence of a hydrogen bond, so that weaker bonds could also be recognised due to inaccuracies in the coordinates. The definition of a hydrogen bond is not straightforward, as there are no boundaries between the quantum-mechanical regime, with the superposition of wavefunctions, and the electrostatic regime, with the Coulombic interaction as the protagonist, so each definition is designed specifically for a specific purpose.

For the identification of structures other than helices and sheets, other criteria were used. For instance, bends are regions of high curvature, which is quantified by considering a sequence of five amino acids and the angle between the backbone direction of the first three and the backbone direction of the last three: the cutoff value was set at 70° . Defined as C^α the position vector of C^α :

$$\text{Bend}(i) = [\text{angle} \{ (C^\alpha(i) - C^\alpha(i-2)), (C^\alpha(i+2) - C^\alpha(i)) \}] > 70^\circ \quad (2.3.13)$$

The disulphide bridges, otherwise known as SS-bonds, which in protein structures are covalent bonds between the S^γ atoms of two Cys residues are taken directly from the SSBOND data bank, since they can be considered part of the primary sequence. All potential disulphide bridges between sulphur atoms at a distance of less than $3.0 \text{ \AA}$ are added. As for chain breaks, these are identified whenever the bond distance of a peptide bond exceeds $2.5 \text{ \AA}$.

2.3.2.3 Root Mean Square Deviation (RMSD)

In order to objectively assess the general mobility/flexibility of the molecular species involved in the MD simulations, it was decided not to limit ourselves to observing the calculated trajectories, but to use the Gromacs tool for calculating the Root Mean Square Deviation (RMSD), which is defined using the structure at the beginning of the simulation as the reference structure, i.e. the structure at the instant t_2 in the following formula:

$$\text{RMSD}(t_1, t_2) = \left[\frac{1}{M} \sum_{i=1}^N m_i \|r_i(t_1) - r_i(t_2)\|^2 \right] \quad (2.3.14)$$

Where $r_i(t_1)$ is the position of the i -th atom at time t_1 , $r_i(t_2)$ the position of the atom in the structure of reference and M is the ensemble of all m_i .

2.3.2.3 Molecular Mechanics/Generalized Born Surface Area (MM/GBSA)

Protein-ligand complexes binding energies is always a daunting task to be completed, but over the years many different approaches have been proposed to get at least an estimate of the values, since these represent a crucial feature in drug design. The binding affinity can also be evaluated by docking scoring functions, but it is generally not particularly accurate, thus being unable to discriminate between binders which have the same order of magnitude in affinity¹⁶⁰. Other methods, like the alchemical perturbation (AP) methods¹⁶¹, are far more accurate, in principle, but accuracy comes together with computational cost, thus making them unfeasible in many cases. Anyway, *in medio stat virtus*, so methods like MM/PBSA (Molecular Mechanics with Poisson-Boltzmann and Surface Area solvation) and MM/GBSA are amongst the most popular ones for calculating binding energies since they provide an improved accuracy if compared to docking scoring functions without being too costly. The first development of the MM/PBSA method dates back to the late nineties and has been widely used since¹⁶². The two key equations of MM/PBSA are the one defining the free energy of a state:

$$G = E_{\text{bnd}} + E_{\text{el}} + E_{\text{vdW}} + G_{\text{pol}} + G_{\text{np}} - TS \quad (2.3.15)$$

Where E_{bnd} , E_{el} and E_{vdW} are respectively the bonded, electrostatic and van der Waals contribution to energy, G_{pol} is the polar contribution to the solvated free energy, G_{np} is the non-polar one, T is the absolute temperature and S is entropy, estimated basing on normal-mode vibrational frequencies. And:

$$\Delta G_{\text{bind}} = \langle G_{PL} \rangle_{PL} - \langle G_P \rangle_P - \langle G_L \rangle_L \quad (2.3.16)$$

Which is the formula used to calculate the binding energy, which is given by the subtraction of the average free energy for the apo-protein and the ligand alone MD simulations from the corresponding one obtained from the complex MD simulation trajectory. This means that three different simulations are supposed to be completed, one for the complex, one for the protein and one for the ligand. This is a particular implementation of MM/PBSA methodology, also known as three average MM/PBSA (3A-MM/PBSA). However, a more common approach is the one-average one, known as 1A-MM/PBSA, which exploits the complex simulation to obtain also the values for the other two. On one hand it improves precision and leads to the cancellation of errors related to the E_{bnd} contribution, on the other it does not account for changes in the ligand (nor protein) structure upon binding. The polar solvation term can be obtained with a numerical calculation with the Poisson-Boltzmann equation, but also with continuum-solvation methods like the Generalized Born one, (MM/GBSA method) which is now more popular since calculations are faster¹⁶⁰. There is no consensus with respect to what is the most accurate or reliable approach neither for the 1A or 3A implementation, nor for the polar solvation contribution, so, in this work, a one-average approach was used, with the implicit Generalized Born solvation (1A-MM/GBSA).

2.4 Docking calculations

Molecular docking is a computational tool used to predict how two molecular species interact and determine the three-dimensional structure of the resulting complex starting from the separated structure of the supposedly interacting species. Molecular docking is often used to evaluate protein-ligand interactions and find the corresponding complex structure, making it a popular tool for drug design, especially since experimental methods to get these structures are difficult to be carried out and expensive¹⁶³. All docking algorithms are based on the hypothesis that the native complex structure is near the global minimum of the PES and consist of two crucial steps: the sampling of possible poses and their evaluation and subsequent scoring¹⁶⁴. Sampling is chronologically the first step of the calculation. It consists in the generation of different possible “solutions” for the “interaction problem”. On the other hand, scoring phase is based on the evaluation of the binding affinity for each of the previously generated poses, performed with a specific scoring function, which can be either physical or empirical. In general, the score is a negative number, depending on the binding affinity in a certain way, and the more negative it is, the better the associated pose is. Sometimes, these two steps are not clearly separated and may even be divided in more sub-steps, depending on the employed methodology. No matter the algorithm, each docking program requires starting structures of the receptor and the ligand(s) to work on. If the performed calculation is a so-called “bound” docking, i.e. the final pose is already known and available, the starting structure of both ligand and receptor can be taken from the complex one, without conformational readaptation. This is the easiest case in docking, but it is usually not the goal, since the objective is to perform predictive docking, also known as “unbound”.

In this case the structures of either the ligand or the receptor can be obtained in different ways¹⁶⁴:

- *Native structures*: they consist of the experimentally obtained structures of the considered molecular species when uncomplexed, typically in solution.
- *Pseudo-native structures*: they are structures extracted from a complex, but not with the same partner which is chosen for the docking calculation.
- *Modeled structures*: these structures are obtained from previously performed calculations, either classical or quantum mechanical.

Once the starting structures are selected, they are employed to get a basic description of both the ligand and the receptor. This description is different from program to program and can be based on a simple geometric features evaluation or the atomic representation of surface residues. Predictive docking is far more complex than bound docking mainly due to the conformational changes of the ligand and the receptor, which can even be a disordered protein¹⁶⁵, an undesirable feature for docking. Conformational changes have been demonstrated to be popular in proteins, especially at the interface and around the binding sites, which are usually partially rigid and partially flexible^{166,167}. Furthermore, the adaptability of the binding sites is paired with a loss of stability in other regions of the protein¹⁶⁸. All these features point in the same direction, the necessary evaluation of conformers in docking, whenever the characteristics of the molecular species involved justify for it.

Docking procedures can be essentially threefold: rigid body approaches, semi-flexible ones and flexible docking. The former regard the interacting structures as rigid ones, without any conformational freedom, making it necessary to use more starting structures for either the receptor or the ligand if accounting for conformational changes is required. On the other hand, the latter are based on the assumption that both the receptor and the ligand structures are flexible, although the conformational freedom is necessarily limited. Semi-flexible methods, instead, do not treat the receptor and the ligand in the same way: the receptor, usually the bigger molecular species involved in the calculation, is considered as a rigid structure, while the ligand is flexible.

In general, the larger the systems involved in the calculation, the lower the grade of flexibility allowed to the structures, due to computational costs. This is why the rigid body algorithms are more popular in protein-protein docking and the flexible ones are mainly employed in protein-ligand docking, where one of the structures is smaller¹⁶⁴. Nonetheless, all of them share the previously introduced basic functioning, i.e. searching and scoring.

2.4.1 Searching

The first step of a docking calculation is searching, also known as sampling, which consists in the generation of different possible complex structures made up of the two partners whose interaction is object of analysis. This step is not focused on finding the best possible solution, but on finding many hypothetical ones.

Regardless of the grade of flexibility introduced, in almost all approaches the bigger of the interacting molecular species is called “receptor”, is fixed in space and has a reduced conformational freedom, if any, while the smaller is called “ligand” and is rotated and translated around the other. The rotation and translation of the interacting structures can be done with an exhaustive or a stochastic method¹⁶³. The latter, also called “systematic searching” is more computationally expensive, but reduces the risk to omit native structures by evaluating all the possible, and feasible, ligand reorientation and conformational changes. Stochastic methods, instead, are cheaper and more suitable for pairing with metaheuristic algorithms, i.e. approaches starting from a set of random solutions.

Many different searching methods have been proposed, belonging to two different families:

- *Geometric algorithms*: Since the interface of interacting proteins shows a prominent shape complementarity, geometry-based techniques are very popular, with different constraints depending on the situation. Generally, in protein-ligand docking, the constraints imposed are stronger, basing on the lock-key strategy, while in protein-protein docking more flexibility is conceded. A general pipeline of a geometry-based algorithm requires, first of all the reconstruction of the structures surfaces in terms of geometrical shapes. Their alignment and reciprocal fitting define the poses generation. Shape complementarity-based docking became extremely popular after the introduction of Fast Fourier Transformed (FFT) to speed up the searching process. In these methods, each structure, both the ligand and the receptor, is represented in a 3D cartesian space with discrete functions representing the surface and the internal part. Correlation functions, calculated with FFT, are used to assess the degree of overlap between the structures at each variation of the ligand position and/or orientation. Whenever proteins are involved in the calculations, small conformational freedom of the residues sidechains are taken into account by allowing a small degree of penetration between the matching surfaces. This approach is a way to introduce a small part of protein flexibility and is called soft docking. Moreover, additional terms can be introduced to refine the results, for example including electrostatic and hydrophobic matching. These approaches are popularly used due to the exhaustiveness of the rigid-body search which characterizes them, but more FFT evaluations per pose have to be done, leading to an increase in computational cost, especially if additional terms are added. Another popular geometric approach is geometric hashing, a computer vision technique which compares the similarities between two different regions after assigning each point an invariant representation. Basing on the agreement in data points representation, matching areas are found and aligned to generate a pose.

- *Population-based metaheuristic algorithms*: Each docking calculation can be seen as an optimization problem, with the energy of the complex structure as the property to be minimized, thus making it possible for metaheuristic algorithms, like the evolutionary ones, to be applied. In these algorithms the binding orientation, position and conformation of the ligand are sampled randomly. The random changes will be accepted or refused depending on different criteria, like the Boltzmann probability function for Monte Carlo (MC) methods.

$$P \sim \exp \left[\frac{E_0 - E_1}{k_b T} \right] \quad (2.4.1)$$

In equation 2.4.1 E_0 is the energy score of the ligand before the change and E_1 is the one after it, while k_b is the Boltzmann constant, and T is the temperature. Another example of population-based algorithms is the application to docking of the Genetic Algorithms (GA), proposed by John Holland in genetics¹⁶⁹, which was proposed in 2001 by Gardiner¹⁷⁰. The algorithm was designed using the set of the six degrees of freedom of a rigid ligand as a chromosome, with Niche restriction to allow the exploration of a greater part of the possible sampling space and no strict constraints on geometric complementarity to account for protein surface flexibility.

Regardless of the applied method for the generation of the poses, it can always be combined with different approaches to introduce protein flexibility whenever required. Apart from the aforementioned soft docking, other techniques can be employed, like the sidechain flexibility method, which samples the conformation of the sidechains before the searching phase, or the molecular relaxation approach, which consists of a minimization by MMC or MD simulations of the backbone and sidechains surrounding the ligand, being far more time-consuming than other approaches^{171,172}. Another option consists in using an ensemble of protein structures to describe conformational changes and it can be declined in different ways according to the method used. For example, FlexE¹⁷³ program creates an ensemble of receptor conformations and aligns them basing on the similar segments of the structure itself, exploiting the dissimilar regions to create new possible protein structures for the docking with a combinatorial approach, while Wei et al.'s algorithm¹⁷⁴ goes further by identifying the variability regions and keeping the best-fitting local conformers for each ligand placement, sticking this local structure to the rigid part of the protein.

2.4.2 Scoring

After having generated the poses, it is necessary to proceed with their analysis in order to identify the models closer to the real complex structure, which means finding the best poses amongst the variety of available options. This selection process is carried out basing on the results of scoring, the second crucial part of each docking calculation. Scoring functions can be divided into two main groups: energy-based and knowledge-based ones. The energy-based functions can then be of two different types, either force field based or empirical.

- *Force field scoring functions*¹⁷⁵: these are based on MM parameters, see Chapter 2.2, which are used to calculate the binding energy by summing up bonded and non-bonded contributions between ligand and receptor atoms. An example is the DOCK¹⁷⁶ scoring function:

$$E = \sum_i \sum_j \left(\frac{A_{ij}}{r_{ij}^{12}} + \frac{B_{ij}}{r_{ij}^6} + \frac{q_i q_j}{\epsilon(r_{ij}) r_{ij}} \right) \quad (2.4.2)$$

In order to get a more reliable result, rigorous approaches like the free energy perturbation method (FEP), where water molecules are treated explicitly, should be used, but these methods are computationally expensive.

This is the reason behind the implementation of simplified versions of those approaches and the introduction of new ones with implicit solvation, like the Poisson-Boltzmann/surface area (PBSA) and generalized-Born/surface area (GBSA) ones, which are able to give reasonable results, especially in terms of comparisons between different poses with less computational resources. However, the solvent contribution is not the only issue ff-based scoring functions have to deal with, the entropic contribution is even a bigger one. Moreover, the possibility to consider the binding free energy as the result of the sum of individual contributions for the single atoms of the involved structures is still a matter of debate.

- *Empirical scoring functions:* the same concept of energy calculation behind the ff scoring functions is at the base of the empirical scoring functions, only declined in a different way. In this case, the binding energy score of a complex is calculated as the sum of different empirical energy contributions, like VDW, electrostatic and H-bond energy, entropy, desolvation and hydrophobicity terms and others.

$$\Delta G = \sum_i W_i \Delta G_i \quad (2.4.3)$$

In equation 2.4.3, $\{\Delta G_i\}$ represents the different energy terms and $\{W_i\}$ the corresponding weights calculated by reproducing the binding affinity of a training set of already known complexes, with least squares fitting^{177,178}. Both empirical and ff scoring functions are energy-based, but the former are far quicker to be calculated, due to the simpler formulation for the binding energy. These methods are becoming more and more popular since the growth of known complexes with their associated binding energy allows for a more extended training set, thus leading to a more reliable scoring function. Some examples of these functions are GlideScore^{179,180}, ChemScore¹⁸¹ and AIScore¹⁸².

- *Knowledge-based scoring functions:* these functions are characterized by the exploitation of structural information of already known complexes whose experimental data is available¹⁸³. The key metric behind this type of scoring function is the potential of mean force¹⁸⁴, reported in equation 2.4.4¹⁸⁵.

$$\omega(r) = k_b T \ln \left[\frac{\rho(r)}{\rho^*(r)} \right] \quad (2.4.4)$$

T is the absolute temperature of the system, k_b is the Boltzmann constant, $\rho(r)$ is the number density of the protein-ligand atom pair at distance r in the training set, and $\rho^*(r)$ is the pair density in a reference state where the interatomic interactions are absent. After the computation of potential parameters for each atom pair in the complex, the total binding energy is obtained as a sum of the pair contributions, thus meaning that the calculations are simple, i.e. fast, and relatively accurate, due to the constantly growing number of complexes' experimental structures. However, even these functions have a criticality, in particular the determination of $\rho^*(r)$, due to the fact that an ideal reference state is not available for large systems¹⁸⁵.

Ideally a scoring function should be perfectly correlating with the quality of the obtained pose, but that is still not the case, despite the great improvements over the last decades. This is why, often, further analysis and refinement are required to get a more reliable solution to the interaction problem. Possible approaches include consensus scoring, which consists in combining the results obtained from different scoring functions weighing them according to a combination rule, and entropy-based scoring methods. This methodology requires clustering the poses and evaluate the clusters' entropic contribution, which is proportional to the number of the poses in each cluster¹⁶³.

2.4.3 Induced Fit Docking (IFD)

The methodology employed in this work is the Induced Fit Docking protocol (IFD)¹⁸⁶ included in Maestro Schrodinger Suite¹⁸⁷, which allows to account for both the ligand flexibility and the receptor one, by combining iteratively rigid docking, with Glide^{179,180} program, and protein structure refinement with Prime^{188,189}. IFD approach tackles the flexibility of the interacting partners exploiting Glide module for the ligand one and Prime refinement for the receptor. Essentially, IFD involves four steps:

- Rigid soft docking to reduce clashes and better sampling to generate a pool of poses
- Sampling of protein conformations for each of the best 20 generated poses, after ranking according to the GlideScore scoring function
- Redocking of the ligand into the previously obtained structures, this time without Van der Waals radius scaling, to get the final poses
- Scoring based on GlideScore and Prime energy values combined to account for both the docking energy and the receptor strain

This protocol is then repeated whenever the final structures get an almost identical final score, in order to better discriminate the best poses. The flowchart for the protocol is depicted in Figure 2.6.

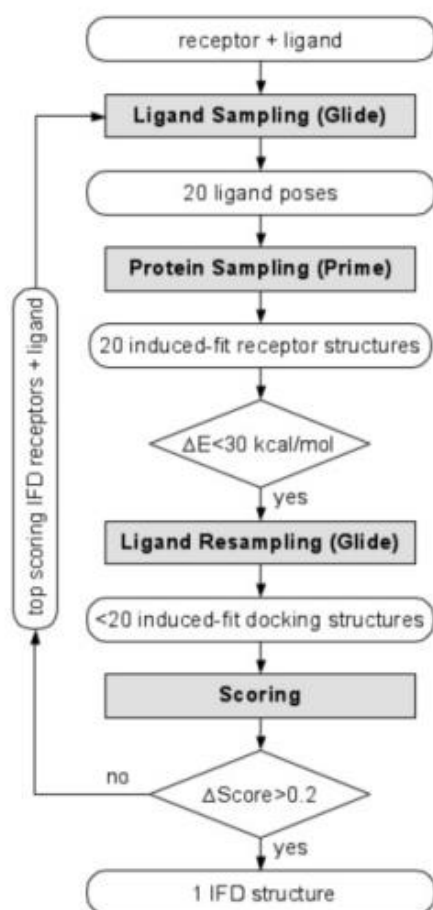


Figure 2. 6 - Induced Fit Docking (IFD) flowchart. ΔE is the energy difference between the lowest-energy structure and the considered ones, ΔScore is the gap in final IFDScores. Figure taken from reference¹⁸⁶

2.4.3.1 Initial ligand sampling

The initial ligand sampling is performed with a rigid docking approach basing on Glide program. The main challenge in this step is finding a balance between reducing the steric clashes between the ligand and the protein without changing the binding site so much that too many unfeasible conformations are brought on to the next step. This issue is addressed by first scaling the Van der Waals radii of atoms by 50% or by 70% and substituting residues predicted to be highly flexible with alanine ones. This second feature is crucial to avoid the starting protein structure to be too determinant in final poses prediction, since the mutation to alanine prevents the conformation of flexible residues to be blocking the access to the catalytic site, allowing a better sampling of it. It is indeed expected that highly flexible residues can be temporarily blocking the binding sites, but the reversibility of this process cannot be taken into account otherwise, without already considering explicitly the protein conformation changes. Glide exploits a hierarchical filtering approach to sample ligand positions within the receptor active site. First, the receptor shape and its properties are represented on a grid by different sets of fields, which allow for a progressively more accurate scoring of ligand poses. Subsequently, a set of initial ligand conformations is generated basing on the minima found in the torsion-angle space and a pre-screening process is quickly run to reduce the region of the phase space which will be considered for energy and gradient evaluations and identify promising docking regions. Once the pre-screening is completed, the best poses are then object of a minimization, in particular the ligand is minimized with the OPLS-AA forcefield including a distance-dependent dielectric model. The three to six lowest-energy poses obtained are then furtherly refined with a MC approach to orient peripheral groups and eventually adapt torsional angles. The choice of the 20 best poses to carry on for the following steps of the IFD procedure is based on GlideScore, which is an empirical scoring function derived from the ChemScore one:

$$\Delta G_{\text{bind}} = C_0 + C_{\text{lipo}} \sum f(r_{\text{lr}}) + C_{\text{hbond}} \sum g(\Delta r)h(\Delta R) + C_{\text{metal}} \sum f(r_{\text{lm}}) + C_{\text{rotl}} \quad (2.4.5)$$

where the second term is the contribution of all the ligand-atom/receptor-atom pairs defined as lipophilic, while the third is the one associated to the hydrogen-bonding pairs. f , g and h functions are employed to better weigh the contributions of pairs whose parameters are slightly out of the limits, assigning a full value (1.00) for distances or angles which are within normal limits and a partial one (from 0.00 to 1.00) for the others (inside larger threshold values). Equation 2.4.6 is modified to get to the GlideScore 2.5 scoring function:

$$\begin{aligned} \Delta G_{\text{bind}} = & C_0 + C_{\text{lipo-lipo}} \sum f(r_{\text{lr}}) + C_{\text{hbond-neut-neut}} \sum g(\Delta r)h(\Delta \alpha) + \\ & + C_{\text{hbond-neut-charged}} \sum g(\Delta r)h(\Delta \alpha) + \\ & + C_{\text{hbond-charged-charged}} \sum g(\Delta r)h(\Delta \alpha) + C_{\text{max-metal-ion}} \sum f(r_{\text{lr}}) + \\ & + C_{\text{torb}}H_{\text{roth}} + C_{\text{polar-phob}}V_{\text{polar-phob}} + C_{\text{coul}}E_{\text{coul}} + C_{\text{vdW}}E_{\text{vdW}} \\ & + \text{solvation terms} \end{aligned} \quad (2.4.6)$$

In this case the lipophilic term is defined as in equation 2.4.5, while the hydrogen-bonding term is divided in different components weighed in a different way according to their charge and the metal-ligand interaction is furtherly modified. It considers only interactions of the metal with anionic acceptor atoms, but only if the calculated net charge for the metal in the apoprotein is positive, and just the strongest one when two or more are found. Other relevant contributions to the binding energy are the Coulomb and the Van der Waals interactions, which are both dimmed. The former is reduced by modifying the charge of charged groups and setting it at the 50% of the initial value, while the other is reduced with a vdW radii diminution to roughly preserve hydrogen-bonding distances, which reflect a balance between attractive electrostatic and repulsive vdW forces. Glide docking program is also

notable for the “water-scoring” technology, which is a particular approach to introduce the solvation contribution. Charged groups, in particular, are always tricky when it comes to solvation effects, since they must be correctly assessed, avoiding their collocation in hydrophobic pockets of the enzyme, but also preventing water trapping in the same pockets. To solve this problem, Glide uses explicit water molecules docked in the binding site for each potentially correct pose, using empirical scoring terms to evaluate the interactions between them and the ligand groups.

2.4.3.2 Receptor refinement

Each of the 20 best-scoring poses from the previous step undergoes a receptor refinement process to account for protein sidechains’ flexibility, which consists in the application of the Prime refinement algorithm on the protein residues which are closer than 5 Å from the ligand. During this step, MM parameters are employed, an implicit solvation is included (Generalized Born model¹⁹⁰) and all the residues which were converted into alanines in the previous step are recovered before refining. The residues of interest are involved in a conformational analysis and in a subsequent minimization to reduce clashes. After completing this process, the low-energy conformations are then minimized again, this time including both the selected residues and the ligand in the minimization process. Once this optimization has been done, the resulting poses are then classified according to the Prime energy value and the structures within 30 kcal/mol of the minimum energy one are brought on to the next step.

2.4.3.2 Ligand redocking

The minimized structures are then object of the final step of the protocol, which is another round of Glide docking. This time, the ligand is redocked with the default settings, but in the receptor structure obtained from the previous step. At the end of the procedure, the remaining poses are reranked according to another scoring function.

2.4.3.3 Final Scoring

Final scoring is based on functions and values which have already been introduced and exploited in the previous sections of the protocol for classification of the poses in the different steps. In particular, it is a combination of Prime Energy and GlideScore. These two are combined to take into account the contributions related to the protein strain and the receptor-ligand interactions. Indeed, Prime is able to predict protein structures, but the aforementioned minimization processes are involving the ligand too and parameters for it are not as accurate as the protein ones, due to the variety of compounds which can be involved in these calculations. Glide, on the other hand, is a program designed to calculate binding affinity and, on top of that, is able to well describe the behaviour of water molecules in the active site, since it is calibrated to give an accurate estimate of their free energies. The IFDScore has been designed to try to get the best of both worlds and the equation to calculate it is the following:

$$\text{IFDScore} = \text{GlideScore} + 0.05 \text{ Prime Energy} \quad (2.4.5)$$

The Prime Energy contribution is added to the GlideScore in order to try to avoid including in the best-hitting poses structures characterized by an extreme protein strain.

2.5 Ab initio methods

In classical mechanics, the dynamics of a system, i.e. its evolution in time, is given by Newton's second law, or equation of motion:

$$F = ma \quad (2.5.1)$$

where F is the force, a is the acceleration, and m is the mass of the particle. However, since electrons exhibit both a corpuscular and wave nature, the best way to describe their behaviour is through the wavefunction Ψ . The physical meaning of Ψ is related to its square $|\Psi|^2$, known as the electron density function, which describes the probability of finding electrons in the different space-spin positions, once the geometry of the nuclei is fixed. The wavefunction is the pillar of the Schrödinger equation, which is the quantum-mechanical equivalent of Newton's second law.

$$\mathcal{H}(\mathbf{r}, t)\Psi(\mathbf{r}, t) = i\hbar \left(\frac{\partial \Psi(\mathbf{r}, t)}{\partial t} \right) \quad (2.5.2)$$

Where $\hbar = h/2\pi$ and $\mathcal{H}$ is the Hamiltonian operator for the system. *Ab initio* calculation ("from the beginning", i.e. from first principles, without empirical parameter) of observables of a many-electron system by means of QM methods require the solution of this equation, but the exact solution of it is not obtainable and the achievable ones are the result of the introduction of approximations, both on the Hamiltonian and the wavefunction.

First, the time-dependence of the Hamiltonian can be reconsidered when dealing with a system in a stationary state, which means that the dependence of the wavefunction on time is given by a phase factor:

$$\Psi(\mathbf{r}, t) = \Psi(\mathbf{r})e^{\frac{-iEt}{\hbar}} \quad (2.5.3)$$

Thus meaning that the Schrödinger equation can be rewritten as:

$$\mathcal{H}(\mathbf{r})\Psi(\mathbf{r}) = E\Psi(\mathbf{r}) \quad (2.5.4)$$

where $\mathcal{H}$ is the time-independent Hamiltonian operator of the system with which the observable energy E , eigenvalue of $\mathcal{H}$, is associated; Ψ is the wavefunction, which is $\mathcal{H}$ eigenfunction. If the wavefunction is known, it is possible to calculate the observables of any system in a particular stationary state and, since Ψ for molecular species is depending on the spatial coordinates of N nuclei ($R_A, \dots, R_N$) and on the spatial ($r_1, \dots, r_n$) and spin ($s_1, \dots, s_n$) coordinates of n electrons, equation 2.5.1 becomes:

$$\mathcal{H}\Psi(R_A, \dots, R_N, r_1, s_1, \dots, r_n, s_n) = E\Psi(R_A, \dots, R_N, r_1, s_1, \dots, r_n, s_n) \quad (2.5.5)$$

In order to solve this equation, further approximations or restrictions must be introduced to both H and Ψ .

2.5.1 Hamiltonian operator approximations

For the H operator, relativistic effects and thus the dependence of mass on velocity are neglected. The molecular Hamiltonian operator becomes, with these approximations, the non-relativistic electrostatic operator H , sum of the kinetic energy operator (T) of nuclei and electrons, and the potential energy operator (U) for all electrostatic interactions between particles:

$$H = T + U \quad (2.5.6)$$

The kinetic energy operator can be expressed as:

$$T = - \sum_{A=1}^N \frac{1}{2M_A} \nabla_A^2 - \sum_{i=1}^n \frac{1}{2} \nabla_i^2 \quad (2.5.7)$$

Where M_A is the mass of the nucleus A, ∇_A^2 is the corresponding Laplacian operator and N is the number of nuclei, while ∇_i^2 is the Laplacian operator corresponding to the i-th electron and n is the total number of electrons. While the potential energy operator in atomic units is:

$$U = - \sum_{A=1}^N \sum_{i=1}^n \frac{Z_A}{|R_A - r_i|} + \sum_{i=1}^n \sum_{j>i}^n \frac{1}{|r_i - r_j|} + \sum_{A=1}^N \sum_{B>A}^N \frac{Z_A Z_B}{|R_A - R_B|} \quad (2.5.8)$$

Where Z_A and Z_B and R_A and R_B represent the charges and coordinates of nuclei A and B, respectively, and r_i and r_j are the coordinates of electrons i and j. M_A is the mass of nucleus A, ∇_A^2 is the Laplacian operator relating to nucleus A and ∇_i^2 is the Laplacian operator of the electron i. Since Laplacian operator is:

$$\nabla_A^2 = \left(\frac{\partial^2}{\partial x_A^2} + \frac{\partial^2}{\partial y_A^2} + \frac{\partial^2}{\partial z_A^2} \right) \quad (2.5.9)$$

The Hamiltonian operator can be rewritten as:

$$H = -\frac{1}{2} \sum_A \frac{1}{M_A} \nabla_A^2 - \frac{1}{2} \sum_i \nabla_i^2 - \sum_{i,A} \frac{Z_A}{r_{Ai}} + \sum_{i,j} \frac{1}{r_{ij}} + \sum_{A,B} \frac{Z_A Z_B}{r_{AB}} \quad (2.5.10)$$

where r_{Ai} , r_{ij} and R_{AB} denote the nucleus-electron, electron-electron and nucleus-nucleus distances, respectively. The eigenfunction Ψ , which is the solution of equation 2.5.4 for the stationary state of the system, contains in principle all its electronic states, none with a particular physical meaning. These electronic states are coupled through the kinetic energy term, due to the so-called vibronic coupling. By neglecting the coupling, it is possible to associate each electronic state to a stationary state of the system and this approach, known as adiabatic approximation, is more acceptable the farther the electronic states are in terms of energy. Generally, the energy gap is sufficiently large to consider this approximation valid, but, in some cases, for example, in photochemical reactions, it is necessary to include the possibility of coupling between states.

A further, crucial, approximation to Hamiltonian operator must be introduced, the Born-Oppenheimer approximation¹⁹¹. According to this approximation, it is possible to decouple nuclear and electron motions and treat them independently. The validity of the decoupling of the motions has experimental verification. For example, the time required for the excitation of an electron from the fundamental electronic state to the first excited electronic state (lowest-energy transition in the electronic spectrum in absorption) is in the timescale of 10^{-15} s while appreciable displacements of atoms in molecules (during molecular vibrations) require at least 10^{-13} s. It can therefore be assumed that the molecular geometry does not appreciably change during the electronic transitions.

Despite all these approximations, the impossibility to obtain exact solutions for polyelectronic systems cannot be changed since there is a hurdle which cannot be overcome: the presence, in the Hamiltonian operator, of a term equivalent to the sum of the repulsion energies between electron pairs. This obstacle can only be partially circumvented by introducing appropriate approximations to the wavefunction.

2.5.2 Wavefunction approximations

To simplify the polyelectronic wavefunction, the orbital approximation is introduced, which corresponds to an independent treatment of the motion of individual electrons. If the motions of the n electrons are independent, then the polyelectronic wavefunction can be written as a product of n functions, each dependent on the coordinates of a single electron (monoelectronic function). Each electron in the polyelectronic system is then assigned a wavefunction called spin-orbital $\phi_i(r_n, s_n)$, which depends only on the spatial and spin coordinates of that particular electron.

Each spin-orbital can be expressed as the product of a spatial wavefunction $\varphi_i(r_n)$, commonly called a molecular orbital, and a spin function. The spin function can only have two analytical expressions, which are commonly denoted by the symbols α and β . Consequently, a generic spin-orbital $\phi_i(r_n, s_n)$ can be written as:

$$\Phi_i(r_n, s_n) = \varphi_i(r_n)\alpha(s_n) = \varphi_i \quad (2.5.11)$$

or:

$$\Phi_i(r_n, s_n) = \varphi_i(r_n)\beta(s_n) = \overline{\varphi_i} \quad (2.5.12)$$

The polyelectronic wavefunction can then be written as a product of n orthonormal functions called spin-orbitals (Hartree product):

$$\Psi(r_1, s_1, r_2, s_2, \dots, r_n, s_n) = \Phi_1(r_1, s_1)\Phi_2(r_2, s_2) \dots \Phi_n(r_n, s_n) \quad (2.5.13)$$

where each electron is described by the corresponding spin-orbital ϕ . Furthermore, since electrons are fermions, i.e. have half-integral spin, the wavefunction must be globally antisymmetric. This means that the polyelectronic wavefunction Ψ must remain identical except for the sign if there is an exchange of coordinates between any pair of electrons. In general, for a polyelectronic wavefunction, the condition imposed by the antisymmetry principle¹⁹² is verified if the wavefunction is expressed in determinantal form, i.e. as a Slater determinant; in fact, a determinant changes its sign if a pair of rows or columns is exchanged. The simpler Hartree product is therefore abandoned. A Slater determinant consists of n rows, as many as there are electrons, and n columns, as many as there are spin-orbitals occupied:

$$\Psi(1, \dots, n) = (n!)^{-\frac{1}{2}} \begin{vmatrix} \Phi_1(1) & \Phi_j(1) & \dots & \Phi_k(1) \\ \Phi_1(2) & \Phi_j(2) & \dots & \Phi_k(2) \\ \dots & \dots & \dots & \dots \\ \Phi_1(n) & \Phi_j(n) & \dots & \Phi_k(n) \end{vmatrix} \quad (2.5.14)$$

The single-determinant wavefunction describes a possible electronic configuration, i.e. a possible distribution of the n electrons over n spin-orbitals. If the n occupied spin-orbitals are those of lowest energy, the resulting electronic configuration Ψ_0 is called the ground electronic configuration. Any other assignment of electrons generates higher-energy configurations, the so-called excited configurations. In the Hartree-Fock approximation, the antisymmetrized polyelectronic wavefunction for the fundamental state is approximated by a single Slater determinant and, for this reason, is also referred to as the single-determinant approximation. This approximation consists in describing the fundamental electronic state by means of a single electronic configuration, the ground electronic configuration, i.e. the one constructed from the lowest-energy spin-orbitals. This is a suitable approximation to describe many molecular systems in the vicinity of their equilibrium geometries. It is no longer adequate whenever aiming at describing chemical reactivity or systems for which the fundamental state cannot be reasonably represented by a single configuration.

Since the impossibility of solving the Schrödinger equation is due to the presence of the electron-electron repulsion term, we proceed by solving single-electron equations, in which only one electron is studied in the effective field generated by the other $n-1$ electrons (described by the Fock operator $\mathcal{F}$):

$$\mathcal{F}\varphi_u = \varepsilon_u\varphi_u \quad (2.5.15)$$

Where ε_u can then be interpreted as the energy of a single electron in the molecular orbital φ_u . To obtain the optimal set of single-electron wavefunctions, called Hartree-Fock orbitals, the energy of the system is minimised by applying the variational theorem:

$$\langle E_0 \rangle = \frac{\langle \Psi_0 | \mathcal{H} | \Psi_0 \rangle}{\langle \Psi_0 | \Psi_0 \rangle} \geq E_0 \quad (2.5.16)$$

where $\langle E_0 \rangle$ represents the expectation value of the system energy, Ψ_0 is the wave function of the system, called the variational function, and E_0 is the real energy of the ground state. The variational theorem states that the mean value, or expectation value, of the energy, calculated for any variational wavefunction, is always greater than or, when the wavefunction coincides with the eigenfunction, equal to the eigenvalue. The variational theorem provides one of the most widely used guiding criteria for comparing methods and calculation options, because it allows “ranking” the approximate solutions of the Schrödinger equation, even without knowing the exact solution: the most negative energy correspond to the most accurate solution. The variational function contains variable parameters whose optimal value is calculated through the process of energy minimisation, which consists of calculating the prime derivatives of the energy with respect to the parameters, setting these derivatives equal to zero, and solving the resulting system of equations. The parameters are the coefficients of the linear combinations used to describe molecular orbitals. In the Linear Combination of Atomic Orbitals (LCAO) approach, each Molecular Orbital (MO) can be defined as the linear combination of the so-called atomic orbitals, i.e. atom-centred functions whose analytical form is known a priori:

$$\varphi_i = \sum_{j=1}^n c_{ij} \chi_j^{\text{AO}} \quad (2.5.17)$$

Where χ_j^{AO} are the atomic orbitals and c_{ij} the corresponding coefficients. This approximation is used in all ab initio methods, and the set of known functions χ_i represents the basis set $\{\chi_i\}$, i.e. the atomic distribution functions used to describe the single-electron molecular orbital φ_i through a linear combination of them. The total number of atomic orbitals employed is equal to the number of the molecular orbitals that can be described. The values of the coefficients c_{ij} are found by minimising the energy through the variational method. Since the energy depends on the density, which depends on the coefficients and the coefficients are depending on the elements of the Fock matrix, which on its side depends on the density, only an iterative approach allows to find the solution of the Roothaan equations required to get the c_{ij} values. This process is called Self-Consistent Field (SCF)¹⁹³. It seems intuitive that depending on the type and number of functions included in the basis set, different wavefunctions are obtained. Basis functions can be of two types: Slater Type Orbitals (STO) and Gaussian Type Orbitals (GTO). The two options offer different pros and cons: STOs describe well the electrons' behaviour in dependence on the nucleus-electron distance, while GTOs introduce fewer complications in determining the interactions between electrons. Once the type of function is set, it is important to define the number. Theoretically, it would be possible to use only one function for each atomic orbital (minimum basic set), but it has been demonstrated that this choice does not yield reliable results. One possibility is to select at least two basis functions for each atomic orbital (DZ, double zeta) or even three (TZ, triple zeta). Since the computational cost is proportional to the fourth power of the basis functions' number, it is preferable to keep it to a minimum. Basing on the idea that only valence electrons are

involved in the reactivity and chemical bonding of the system, it is possible to double the number of χ_i only for valence orbitals; this is the case with a split valence basis set. By introducing functions with a higher angular momentum into the basis set, a better description of molecular orbitals can be obtained, because they allow functions with a lower angular momentum to be polarised where asymmetrical shifts in the electronic distribution occur. For elements from the third period onwards, further problems arise. These atoms have a larger number of core electrons, which must be described by a sufficient number of basis functions, and in which relativistic effects, i.e. the dependence of mass on velocity, are no longer negligible. An effective core potential (ECP), commonly called pseudopotential, is therefore used, which represents core electrons as a single function and treats explicitly only valence ones.

2.6 Density Functional Theory (DFT)

“The energy E_0 of the ground state of a many-electron system is a functional of the electron density $\rho(\mathbf{r})$. This function exists and is unique.

$$E_0 = E_0[\rho] \quad (2.5.18)$$

Not only the energy, but also all other molecular properties are determined uniquely by the electron density $\rho(\mathbf{r})$.”

This is the enunciation of the Hohenberg-Kohn theorem¹⁹⁴, which was formulated in 1964, and represents the basis of DFT by stating that there is a one-to-one correspondence between the electron density of a system and its energy. The electron density function is strictly related to the coordinates of the electrons. Considering a one-electron system, the wavefunction is depending on the electron $\mathbf{x}$ coordinate, which represents the $\mathbf{r}$ spatial coordinates and the spin one (s), thus meaning that the probability of finding the electron in a specific volume element (in terms of space-spin) $d\mathbf{x}$ is:

$$\psi^*(\mathbf{x})\psi(\mathbf{x})d\mathbf{x} \quad (2.5.19)$$

This means that the probability of finding it in the element of volume $d\mathbf{r}$, centred at $\mathbf{r}$, and with a spin between s and $s+ds$ is determined by the probability function:

$$\rho(\mathbf{x}) = \psi^*(\mathbf{x})\psi(\mathbf{x})d\mathbf{x} \quad (2.5.20)$$

Generalizing to a polyelectronic system of n electrons, described by the wavefunction $\Psi(\mathbf{x}_1, \mathbf{x}_2, \dots, \mathbf{x}_n)$, the probability to find simultaneously electron 1 in $d\mathbf{x}_1$, electron 2 in $d\mathbf{x}_2$, etc. is given by:

$$\Psi^*\Psi d\mathbf{x}_1 d\mathbf{x}_2, \dots, d\mathbf{x}_n \quad (2.5.21)$$

And the corresponding probability of finding electron 1 in $d\mathbf{x}_1$ no matter the location of the others, is:

$$d\mathbf{x}_1 \int \Psi^* \Psi d\mathbf{x}_2, \dots, d\mathbf{x}_n \quad (2.5.22)$$

Since any electron has the same probability of being in a given volume element, the probability density function for a single electron in a system of n electron is:

$$\rho(\mathbf{x}_1) = n \int \Psi^* \Psi d\mathbf{x}_2, \dots, d\mathbf{x}_n \quad (2.5.23)$$

Integrating with respect to the spin variable we get the spin-independent density function, which is called electron density function, expressed as:

$$\rho(\mathbf{r}_1) = \int \rho(\mathbf{x}_1) ds_1 \quad (2.5.24)$$

The relevance of Hohenberg-Kohn theorem becomes evident when comparing the density-based approach to calculate molecular properties with the wavefunction-based one. A wavefunction for an N electron system contains 4N variables, three spatial coordinates and a spin one for each electron, while the electron density can be calculated as the square of the wavefunction, integrated over N-1 electron coordinates. While the complexity of the wavefunction scales with the dimension of the system in terms of number of electrons, the density has the same number of variables, no matter the system size. However, Hohenberg-Kohn theorem affirms only that every different density is associated with to a different ground state energy, but does not specify the functional connecting these two quantities. Different attempts for the identification of the functional have been done over the years and the great leap in the quality of the models was in 1965, with the suggestion by Kohn and Sham¹⁹⁵ that the electron kinetic energy should be calculated from the electron density, represented by an auxiliary set of orbitals. In particular, the approximations already seen on the Hamiltonian (electrostatic Hamiltonian) and the decoupling of electronic motion from nuclear motion (Born-Oppenheimer) are introduced and it can be shown that the electronic energy E_0 of the fundamental state of a closed-shell system consisting of 2n electrons is:

$$E_0 = E_T + E_V[\rho] + E_J[\rho] + E_{xc}[\rho] \quad (2.5.25)$$

Where E_T is the kinetic energy for the i-th electron:

$$E_T = 2 \sum_u^n \int \varphi_u^*(i) \left[-\frac{1}{2} \nabla_i^2 \right] \varphi_u(i) d\tau_i \quad (2.5.26)$$

$E_V[\rho]$ is the electron-nucleon potential:

$$E_V[\rho] = - \sum_A \int \frac{Z_A \rho(r_i)}{|R_A - r_i|} d\tau_i \quad (2.5.27)$$

And $E_J[\rho]$ is the electron-electron repulsive contribution:

$$E_J[\rho] = \frac{1}{2} \int \int \frac{\rho(r_i) \rho(r_j)}{|r_i - r_j|} d\tau_i d\tau_j \quad (2.5.28)$$

While $E_{xc}[\rho]$ is the exchange-correlation energy, i.e. a functional of the electron density $\rho(r)$ that takes into account both exchange effects (Fermi hole, which concerns the probability of finding two electrons with the same spin at the same point in space) and correlation effects (Coulomb hole, which concerns the probability of finding two electrons with opposite spin). It also includes the kinetic correlation energy, that is the difference between the kinetic energy of the real molecule and that of the reference system of non-interacting electrons, and the self-interaction correction (SIC), which is a non-negligible contribution due to the formulation of the traditional charge-cloud electrostatic repulsion. It incorrectly permits the interaction between the electron density of the “diffuse part” of a specific electron and the charge contributions of that same electron across space, even in the case of monoelectronic systems, while an electron does not interact with itself¹⁹⁶.

The exact ground electron density of a system of n electrons can be written as:

$$\rho(r) = \sum_{i=1}^n |\varphi_i(r)|^2 \quad (2.5.29)$$

Where φ_u are the Kohn-Sham orbitals, monoelectronic functions which are the solution of the following equation:

$$\mathcal{F}_{i\text{KS}}\varphi_i(r) = \varepsilon_{i\text{KS}}\varphi_i(r) \quad (2.5.30)$$

$\mathcal{F}_{i\text{KS}}$ is the monoelectronic Kohn-Sham operator which can be written as:

$$\mathcal{F}_{i\text{KS}} = -\frac{1}{2}\nabla_i^2 - \sum_A \frac{Z_A}{|R_A - r_i|} + \int \frac{\rho(r_j)}{|r_i - r_j|} d\tau_j + V_{i\text{xc}} \quad (2.5.31)$$

Where $V_{i\text{xc}}$ is called the exchange-correlation potential and is the functional derivative of exchange-correlation energy $E_{\text{xc}}[\rho]$:

$$V_{i\text{xc}} = \frac{\partial E_{\text{xc}}[\rho]}{\partial \rho} \quad (2.5.32)$$

Therefore, before the monoelectronic equations of KS can be solved, it is necessary to specify the mathematical form of the exchange-correlation functional $E_{\text{xc}}[\rho]$, calculate its functional derivative V_{xc} and insert it into the operator F_{KS} . Once the mathematical form of $E_{\text{xc}}[\rho]$ has been defined, the KS-SCF method is variational, since the KS orbitals generate the electronic density that minimises the value of E_0 . The mathematical forms proposed for $E_{\text{xc}}[\rho]$ are now numerous and represent an attempt to "guess" the true expression of $E_{\text{xc}}[\rho]$. Since the universal functional is not known, many approximated functionals¹⁹⁷ were developed and classified according to the "Jacob's ladder" categorization proposed by Perdew¹⁹⁸ and reported in Figure 2.7:

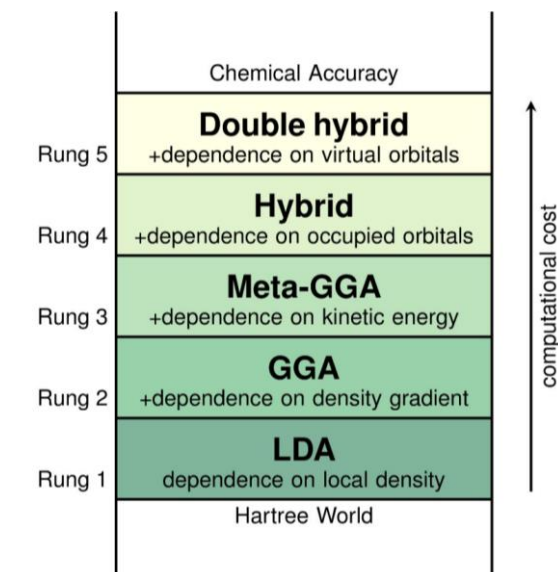


Figure 2.7 - Perdew's Jacob's ladder: each rung corresponds to a different approximation and class of functionals. Image taken from reference¹⁹⁹

The choice of the approximate functional is crucial for the DFT calculation. In general, the E_{xc} functional is divided into two contributions: the exchange functional (E_x), which refers to the exact exchange contribution calculated with the HF method, and the correlation functional (E_c). The first generation of correlation and exchange functionals, termed local functionals, used the local density approximation (LDA), whereby only the value of the electron density at a given point in space determines the contribution of this point in the energy E_{xc} and thus all non-local effects are neglected:

$$E_{\text{xc}}^{\text{LDA}}[\rho] = \int V_{\text{xc}}[\rho(r)]\rho(r)dr \quad (2.5.33)$$

No importance is given to the trend of $\rho(r)$ around the point, thus meaning that this approximation treats the molecules as if they contained a homogeneous electron gas in a uniformly distributed positive charge field. This is an unrealistic representation for molecules and more appropriate for metals or materials with strongly delocalized electrons, nonetheless the LDA functionals provide a relatively accurate structure, even underestimating the exchange energy and overestimating the correlation energy. In molecular structures, electronic density is reflecting the nuclei distribution which is often irregular. The inclusion of contributions depending on density variations over short distances can give better results. The generalised gradient approximation (GGA)²⁰⁰, typical of non-local functionals²⁰¹, adds a term that takes into account the trend of $\rho(r)$ around r' and not just its value at the r' point.

$$E_{xc}^{GGA}[\rho] = E_{xc}[\rho(r)|_{r=r'}, \nabla\rho(r)|_{r=r'}, \dots] \quad (2.5.34)$$

In addition to these pure density functionals, LDA and GGA, there are hybrid functionals, which combine the correlation and exchange functional DFT, usually approximated by GGA, and the Hartree-Fock exact exchange functional.

$$E_{xc}^{Hyb}[\rho] = E_{xc}^{LDA} + a_1(E_x^{HF} - E_x^{LDA}) + a_2\Delta E_x^{GGA} + a_3E_c^{GGA} \quad (2.5.35)$$

Where a_1 , a_2 and a_3 are different according to the selected functional. These hybrid functionals turn out to be the most widely used for calculations involving molecular species, while pure functionals have, over time, been abandoned. Hybrid functionals^{202,203} include the B3LYP^{204,205} and PBE0²⁰³. B3LYP exchange correlation functional contains three parameters, specifying the contributions of exact HF exchange, local exchange, and gradient-corrected exchange and correlation. On the other hand, the PBE0 functional mixes the HF exchange with the GGA-type exchange and correlation contribution. Of these two functionals, depending on the system to which they are applied, the former is more reliable than the latter or vice versa. The main challenge for global hybrid functionals, such as the aforementioned two, is the treatment of charge transfer states. This is why range separated hybrids (RSH)²⁰⁶ like CAM-B3LYP²⁰⁷ were designed. These functionals treat differently the long-range and the short-range exchange contributions, with an exchange-correlation functional whose general formula is:

$$E_{xc}^{RSH}[\rho] = E_{xc}^{DFA} + E_{xc}^{LR-HF} + E_{xc}^{SR-DFA} \quad (2.5.36)$$

Where DFA stands for Density Functional Approximation. Global hybrids and their range separated counterparts are not the latest designed functionals, since double hybrids have been introduced too²⁰⁸⁻²¹⁰, allowing to take into account also the effects of non-occupied orbitals. Over the years, and through this continuous development and refinement, DFT has demonstrated to be a computational technique that allows satisfactory results to be obtained even for relatively large systems, as it is less sensitive to the number of atoms and the size of the basis set than methods based on wavefunction search.

2.6.1 Time-Dependent Density Functional Theory (TD-DFT)

DFT approach is widely popular not only when dealing with ground state molecular structures, but also whenever the studied system is interacting with a time-dependent external potential, for example the electromagnetic field involved in light-matter interactions. A time-dependent external potential forces to consider the Schrödinger equation in its time-dependent formulation:

$$\mathcal{H}(\mathbf{r}, t)\Psi(\mathbf{r}, t) = i\hbar\left(\frac{\partial\Psi(\mathbf{r}, t)}{\partial t}\right) \quad (2.5.37)$$

Which means that the electrostatic Hamiltonian operator in atomic units must be recast as:

$$H_{el} = -\frac{1}{2} \sum_i \nabla_i^2 + \sum_{i,j} \frac{1}{r_{ij}} - \sum_i v(\mathbf{r}_i, t) \quad (2.5.38)$$

Where $v(\mathbf{r}_i, t)$ is the time-dependent potential. In order to define a time-dependent formalism corresponding to static DFT, first it is necessary to verify that, as it was for static DFT, for a given initial wavefunction there is a unique corresponding density, functional of the time-dependent potential up to a constant, and vice versa. This means having a Hohenberg-Kohn theorem analogous, which is the Runge-Gross theorem²¹¹, dating back to 1984. Basing on this, it is possible to consider every observable of the system as a functional of the density. The calculation of time-dependent density, however, is not trivial and the Kohn-Sham formalism used for ground state DFT cannot be directly used for time-dependent calculations. The Van-Leeuwen theorem²¹², enunciated in 1999, represents the link between the KS formalism and the time-dependent case:

“If we consider a system described by the wavefunction Ψ_0 and subject to an external potential $v(\mathbf{r}, t)$, there is another system described by another wavefunction Ψ_0' subject to an external potential differing from the first by a time dependent function, such that the electron density $\rho(\mathbf{r}, t)$ of the two systems is identical”.

This means that a fictitious system, appropriately designed, can be used to find the density of the real one. The direct consequence of the Runge-Gross theorem and the Van-Leeuwen one is the possibility of using Kohn-Sham formalism to carry out time-dependent calculations. Since the external potential can be written as:

$$v(\mathbf{r}, t) = v_0(\mathbf{r}) + v_1(\mathbf{r}, t)\theta(t - t_0) \quad (2.5.39)$$

at time $t=t_0$, the effective potential is:

$$v_{eff}[\rho_0](\mathbf{r}) = v_0(\mathbf{r}) + v_{xc}[\rho_0](\mathbf{r}) + \int \frac{\rho_0(\mathbf{r}')}{|\mathbf{r} - \mathbf{r}'|} d^3\mathbf{r}' \quad (2.5.40)$$

and, keeping in mind the Van-Leeuwen theorem, when $t > t_0$ it can be recast as:

$$v_{eff}[\rho_0](\mathbf{r}, t) = v(\mathbf{r}, t) + v_{xc}[\rho_0](\mathbf{r}, t) + \int \frac{\rho_0(\mathbf{r}', t)}{|\mathbf{r} - \mathbf{r}'|} d^3\mathbf{r}' \quad (2.5.41)$$

Once the effective potential is defined, the time-dependent density can be obtained as a sum of the densities associated with each Kohn-Sham orbital resulting from the following equation:

$$\left[-\frac{\nabla^2}{2} + v_{eff}[\rho](\mathbf{r}, t) \right] \varphi_j(\mathbf{r}, t) = i \left(\frac{\partial \varphi_j(\mathbf{r}, t)}{\partial t} \right) \quad (2.5.42)$$

This approach, in principle, provides an exact value of the time-dependent density. As in the static DFT, also in the TD extension approximations on the exchange-correlation potential must be introduced. In this case, the approximation is not limited to the design of the exchange-correlation functional as described for the static case, since the time-dependent exchange-correlation potential must be somehow depending on the time-dependent electron density. According to the adiabatic approximation, the time-dependent density is included in the static exchange-correlation potential v_{XC}^0 :

$$v_{XC}^A(\mathbf{r}, t) = v_{XC}^0[\rho_0](\mathbf{r})|_{\rho_0(\mathbf{r}) \rightarrow \rho(\mathbf{r}, t)} \quad (2.5.43)$$

But this approximation is valid only in the case of a “weak” perturbation, since it is valid only for systems remaining in their instantaneous eigenstates and whenever any memory dependence of the exchange-correlation potential is not relevant. For example, it does not work for modeling of resonant pulse-driven electron dynamics, since, as soon as the system leaves the ground state, it becomes described by a superposition of non-stationary ones.

Ideally, the resolution of the time-dependent Kohn-Sham equations is required whenever the object of interest is a molecular system under the action of a time-dependent potential, no matter the strength of the perturbation, but, since in many cases the desired observables are related to processes which do not require a strong deviation from the equilibrium state, e.g. absorption, it is more practical to calculate the deviation caused by the perturbation rather than solve all the equations. In Linear-Response framework, this is exactly how the observables are obtained. The difference between the electronic density affected by the perturbation and the static one is defined as “response” and it can be written in terms of a Taylor’s series expansion:

$$\rho(\mathbf{r}, t) - \rho_0(\mathbf{r}) = \rho_1(\mathbf{r}, t) + \rho_2(\mathbf{r}, t) + \rho_3(\mathbf{r}, t) + \dots \quad (2.5.44)$$

Where ρ_1 is the linear response to the perturbation and the following ones are the higher order responses. In case of a “weak” perturbation, i.e. electromagnetic field, only the first order is relevant and can be calculated with the density-density linear response function χ_{eff} :

$$\rho_1(\mathbf{r}, t) = \int dt' \int d^3\mathbf{r}' \chi_{\text{eff}}(\mathbf{r}, t, \mathbf{r}', t') v_{\text{eff}_1}(\mathbf{r}', t') \quad (2.5.45)$$

Where χ_{eff} is:

$$\chi_{\text{eff}}(\mathbf{r}, t, \mathbf{r}', t') = \left. \frac{\partial \rho[v_{\text{eff}}](\mathbf{r}, t)}{\partial v_{\text{eff}}(\mathbf{r}', t')} \right|_{v_{\text{eff}}[\rho_0](\mathbf{r})} \quad (2.5.46)$$

Which means that χ_{eff} is a density functional of the ground state. v_{eff_1} , instead, can be expressed as:

$$v_{\text{eff}_1}[\rho](\mathbf{r}, t) = v_1(\mathbf{r}, t) + \int \frac{\rho_1(\mathbf{r}', t)}{|\mathbf{r} - \mathbf{r}'|} d^3\mathbf{r}' + v_{\text{xc}1}(\mathbf{r}, t) \quad (2.5.47)$$

And $v_{\text{xc}1}(\mathbf{r}, t)$ as:

$$v_{\text{xc}1}(\mathbf{r}, t) = \int dt' \int d^3\mathbf{r}' f_{\text{xc}}(\mathbf{r}, t, \mathbf{r}', t') \rho_1(\mathbf{r}', t') \quad (2.5.48)$$

Where f_{xc} is the time-dependent exchange-correlation kernel, a ground state density functional. By placing equality between the linear response equation calculated for the non-interacting system and the exact one for the interacting system, which is true only when integrated over the Kohn-Sham potential v_{eff_1} , the resulting equation is the Dyson-type equation:

$$\begin{aligned} \chi(\mathbf{r}, t, \mathbf{r}', t') &= \chi_{\text{eff}}(\mathbf{r}, t, \mathbf{r}', t') + \int d\tau \int d^3\mathbf{x} \int d\tau' \int d^3\mathbf{x}' \chi_{\text{eff}}(\mathbf{r}, t, \mathbf{r}', t') \left\{ \frac{\delta(\tau - \tau')}{|\mathbf{r} - \mathbf{r}'|} + \right. \\ &\quad \left. + f_{\text{xc}}(\mathbf{r}, t, \mathbf{r}', t') \right\} \chi(\mathbf{r}, t, \mathbf{r}', t') \end{aligned} \quad (2.5.49)$$

Which contains all the relevant information about the system interaction with the perturbation.

Since in this work TD-DFT was employed to get absorption and emission spectra, it is necessary to introduce the Fourier-transform of equation 2.5.49:

$$\chi(\mathbf{r}, \mathbf{r}', \omega) = \chi_{\text{eff}}(\mathbf{r}, \mathbf{r}', \omega) + \int d^3\mathbf{x} \int d^3\mathbf{x}' \chi_{\text{eff}}(\mathbf{r}, \mathbf{x}, \omega) \left[\frac{1}{|\mathbf{x} - \mathbf{x}'|} + \right] f_{\text{xc}}(\mathbf{x}, \mathbf{x}', \omega) \chi(\mathbf{x}', \mathbf{r}', \omega) \quad (2.5.50)$$

Where:

$$f_{\text{xc}}(\mathbf{r}, \mathbf{r}', \omega) = \chi_{\text{eff}}^{-1}(\mathbf{r}, \mathbf{r}', \omega) - \chi^{-1}(\mathbf{r}, \mathbf{r}', \omega) - \frac{1}{|\mathbf{r} - \mathbf{r}'|} \quad (2.5.51)$$

$$\chi_{\text{eff}}(\mathbf{r}, \mathbf{r}', \omega) = \sum_{j,k=1}^{\infty} (f_k - f_j) \frac{\varphi_j^0(\mathbf{r}) \varphi_k^{0*}(\mathbf{r}) \varphi_j^{0*}(\mathbf{r}') \varphi_k^0(\mathbf{r}')}{\omega - \omega_{jk} + i\eta} \quad (2.5.52)$$

$\chi(\mathbf{r}, \mathbf{r}', \omega)$ is the frequency-dependent response dependent on the Kohn-Sham orbitals ($\varphi_{j,k}$), their energies and occupation numbers ($f_{j,k}$). Since the non-null contributions are only the ones with $f_k - f_j \neq 0$, which means only the occupied/unoccupied pairs contribute to the response function and that the absolute values of $\omega_{jk} = \varepsilon_j - \varepsilon_k$ are the excitation energies of the system. The existence of the poles means that the system can absorb or emit energy only at those frequencies and the corresponding oscillator strengths, i.e. the absorption intensity, is depending on the strength of the poles for these excitations. Obviously, those poles are not the exact ones, but the correction performed by the exchange-correlation kernel f_{xc} makes it up for it and the correct values are obtained.

Equation 2.5.50 is a focal point in the Casida dissertation²¹³ and reformulation of the problem as an eigenvalue one expressed in matrix terms in the form:

$$\begin{pmatrix} \mathbb{A} & \mathbb{B} \\ \mathbb{B} & \mathbb{A} \end{pmatrix} \begin{pmatrix} \mathbf{X} \\ \mathbf{Y} \end{pmatrix} = \Omega \begin{pmatrix} -1 & 0 \\ 0 & 1 \end{pmatrix} \begin{pmatrix} \mathbf{X} \\ \mathbf{Y} \end{pmatrix} \quad (2.5.53)$$

In brief, the prediction of excitation energies is done in three steps:

1. Calculation of the ground state Kohn-Sham orbitals and corresponding energies
2. Recasting of the problem in Casida's equation
3. Application of the exchange-correlation kernel correction

The correctness of the results is strongly dependent on the exchange-correlation functional and kernel, but, in principle, if they were exact, the exact excitation energies would be obtained. TD-DFT can also be used to predict fluorescence emission spectra since it is possible to optimize excited state geometries and then predict the vertical excitation energies at those geometries.

2.7 Hybrid approaches - Quantum Mechanics/Molecular Mechanics (QM/MM)

Computational chemistry has a wide range of tools to investigate the variety of possible chemical systems and there are two main features to take into account when choosing the appropriate approach to treat the system: size and kind of process of interest²¹⁴. Whenever it does not involve bond breaking/formation or photochemical phenomena, size becomes the only criterion to follow, opting for QM approaches when feasible, and for Molecular Mechanics if QM is unnecessary. When the process of interest requires the explicit treatment of electrons, things are not that simple, because QM is required, but, in some cases, may not be feasible. Common examples of this dilemma are enzyme-ligand pairs for mechanistic studies of the catalytic cycles²¹⁵ or the absorption/emission of protein-ligand complexes^{216,217}. The former situation is the one which was faced by Karplus²¹⁸ first, and Warshel and Levitt²¹⁹ later, in the early 70s and then tackled in different ways by the three, who were all awarded a Nobel prize in chemistry in 2013 for their contributions to the development of QM/MM methods²²⁰. In 1976 Warshel and Levitt introduced the multiscaling Quantum Mechanics/Molecular Mechanics (QM/MM) approach to investigate the lysozyme reaction and their novel technique became increasingly popular over the decades, becoming the method of choice not only for enzymology studies, but also for the analysis of solid-state systems^{221,222}, biomolecular complexes and for the inclusion of explicit solvent²²³ in processes. In these cases, the electronic structure of the involved chemical species can change, thus meaning that electronic motion must be taken into account, but QM calculations are feasible only for systems including up to several hundred atoms. For larger systems, there is no alternative than a multiscaling approach, where a section of the system is considered at QM level and the remaining part is treated at MM level. In principle, this methodology is simple, but its implementation is not trivial, since the region to be treated at QM level must be chosen wisely, different schemes can be used to deal with the interaction between MM and QM parts and the boundary between the two requires specific precautions²²⁴.

Partitioning the system into the QM subsystem and the MM one, ideally, would happen without cutting covalent bonds at the boundary and this is a typical situation for protein-ligand pairs, but sometimes it is simply unavoidable, for example in the case of polymers²²⁵ and proteins²²⁶. However, the completeness of the QM subsystem must be maintained, and this is guaranteed by the addition of artificial atoms, which are the heart of Link Atom (LA) approach²¹⁹. These artificial atoms are generally hydrogens, since they are the lightest available ones and do not have a strong electron-withdrawing, or electron-donating, behaviour. The introduction of these atoms leads to the necessity to correct the final energy evaluation, which will be discussed more thoroughly later on.

Once the boundary has been set, the level of interaction between the two subsystems has to be defined. There are three main options: mechanical embedding, electrostatic embedding and polarizable embedding²²⁷. The former consists of calculating the electrostatic effects of the MM part on the QM one by considering the QM atoms as point charges or dipoles, without including the MM effects in the Hamiltonian operator for the QM calculations. This approach has been proven to be computationally cheaper than the electrostatic embedding one, but is not always reliable, as Hagiwara and coworkers demonstrated. In electrostatic embedding, the QM subsystem directly “feels” the presence of the MM region in terms of its own polarization since the MM charges are included in the Hamiltonian operator with the addition of this contribution:

$$H_{\text{QM/MM}} = - \sum_{i,m} \frac{q_m}{r_m} + \sum_{A,m} \frac{Z_A q_m}{R_{Am}} \quad (2.5.54)$$

Where q_m are the MM atoms charges, Z_A the atomic number of QM atoms, i runs over all the electrons of the QM region, A over all QM atoms (LA included) and m over all MM atoms. This way, the QM part is polarized by the MM one, but the MM part will not be affected by the electrostatic behaviour of the QM one. Accounting for MM polarization can lead to improvements in the results, especially for systems including metal ions directly interacting with the MM part, for example water (a strongly polarizable solvent) around a sodium or potassium ion²²⁸.

Finally, after setting the treatment of electrostatic interactions between the QM and the MM regions, the total energy of the system can be computed, exploiting two different schemes: the additive approach and the subtractive one²²⁴. The former considers the QM system as embedded in the MM one and the total energy is given by the sum of the energy related to the QM subsystem, the one corresponding to the MM subsystem and the contribution of the interactions between the two:

$$E_{QM/MM} = E_{QM}(QMS) + E_{MM}(MMS) + E_{QM-MM}(QMS + MMS) \quad (2.5.55)$$

Where the subscripts indicate the level of theory for calculation and QMS and MMS are the QM System and the MM Subsystem, respectively. In the subtractive scheme, instead, the value is obtained by subtracting the MM contribution for the QM part from the whole system MM energy and then adding the QM energy for the QM subsystem.

$$E_{QM/MM} = E_{MM}(QM + MM) - E_{MM}(QM) + E_{QM}(QM) \quad (2.5.56)$$

This equation is the basis for the ONIOM (Our own N-layered Integrated molecular Orbital and Molecular mechanics) scheme²²⁹, one of the most popular QM/MM approaches.

2.7.1 ONIOM scheme

The roots of ONIOM method lay in the QM/MM scheme proposed by Maseras and coworkers in 1995²³⁰, the Integrated Molecular Orbital + Molecular Mechanics scheme (IMOMM) which is an example of a subtractive scheme where the total energy of the system is calculated with the 2.5.56 equation. Since the method is defined this way, it can be considered as a size extrapolation from the accurate QM calculation of the model system, i.e. the QM region, to the large one, by introducing, in a cheap way, the effects of the environment. Comparing this approach with the additive schemes, IMOMM requires the calculation of the MM energy for the QM part, which means that good MM force field parameters must be available, especially for the bonded parameters for the QM/MM boundary. Nevertheless, only those parameters, the non-bonded ones and the atomic charges are required, since there is a cancellation of additive MM terms. Only a year later, a new approach was proposed, which shared the same basilar concept of fragmentation of the system into differently treated parts, this time combining not QM and MM calculations, but two different QM methods. This method, the Integrated Molecular Orbital + Molecular Orbital scheme (IMOMO)²³¹, as the IMOMM one, represents an example of the two-layer ONIOM2 methods, which is a class of methods based on the partition of the system in two subsystems, treated at a different level of theory. The ONIOM scheme anyway, can be ideally extended to the desired number of levels, without limitations. Svensson and coworkers developed a three-layer version, which was effectively named ONIOM, that can also be simplified to the ONIOM2 version whenever two of the layers are treated at the same level. Ideally, the fragmentation can be extended as much as desired, but the most popular approaches are either ONIOM2 and ONIOM3 and treatment of the higher level, i.e. the one where a more accurate QM calculation is performed, is done with DFT approaches.

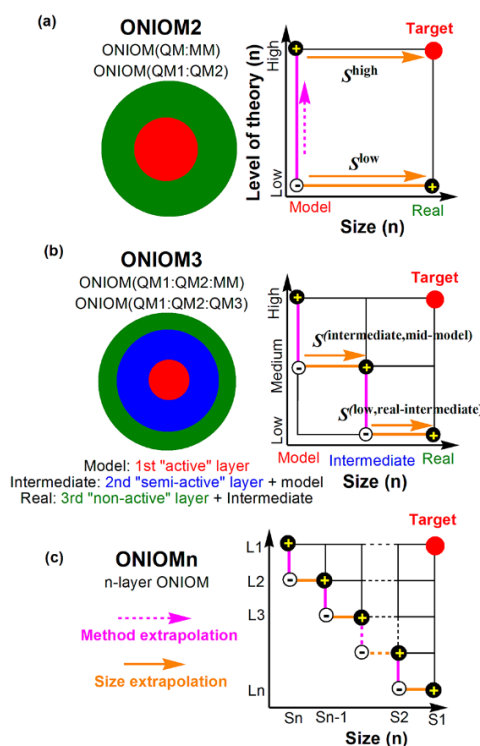


Figure 2. 8 - Representation of system partitioning and treatment in 2-layer ONIOM (a), 3-layer ONIOM (b) and n-layer ONIOM (c) schemes. Figure taken from reference²²⁹

2.8 Applied protocols

This thesis work is made up of three main core elements, which are the study of the furimazine-NanoLuc complex and the subsequent furimamide-NanoLuc emitting pair, the prediction of the furimazine characterizing spectra (^{13}C and ^1H NMR and UV/Vis absorption and emission) in dimethyl sulfoxide (DMSO) and the analysis of the reaction mechanism of furimazine oxidation in DMSO.

2.8.1 Modeling of NanoLuc-furimazine and NanoLuc-furimamide complexes

The first part of the study has been carried out with a multiscale approach, as, despite the size of NanoLuc luciferase (one of the smallest on the market), the protein structure did not allow for a full-QM approach in each calculation that was including it.

Since no deposited crystal structure of the whole NanoLuc-furimazine system is available in the protein data bank (PDB), a starting structure for the complex had to be built up by docking calculations. The NanoLuc input structure for the docking protocol was obtained with the following procedure:

1. the crystal structure of NanoLuc was taken from the Brookhaven Protein Data Bank (<https://www.rcsb.org/structure/5IBO>)
2. the Protein Preparation Wizard²³² tool of Maestro¹⁸⁷, part of the Schrodinger Suite, was used to add eventual missing sidechains, check for missing loops and determine the protonation state of residues with PropKA²³³
3. a MD simulation in water solution was carried out on the structure (see details in Appendix 2)
4. a cluster analysis with the GROMOS¹⁵⁷ algorithm using a cutoff of 0.2 nm on the resulting trajectory was then performed on the MD trajectory frames
5. the centre of the main cluster, the structure with the smallest average RMSD from all other structures of the cluster, was considered for further analysis

The substrate structures for the docking, i.e. furimazine in the keto and enol forms (furimazine_K and furimazine_E), were obtained by ground state geometry optimization in water solution (implicit solvation with CPCM scheme^{234,235}) at CAM-B3LYP²⁰⁷/6-31+G(d)^{236–239} level of theory. The docking calculations were carried out with the induced fit docking (IFD) module^{186,240,241} to account for the flexibility of both the ligand and the protein residues' sidechains for their rotatable bonds (parameters of IFD calculations have been further detailed in Appendix 2). The scoring parameter adopted for best-poses identification and selection is the Glide Docking Score^{179,180}. The best-scoring poses for the keto and enol forms were object of a MD simulation in water followed by cluster analysis on the resulting trajectory.

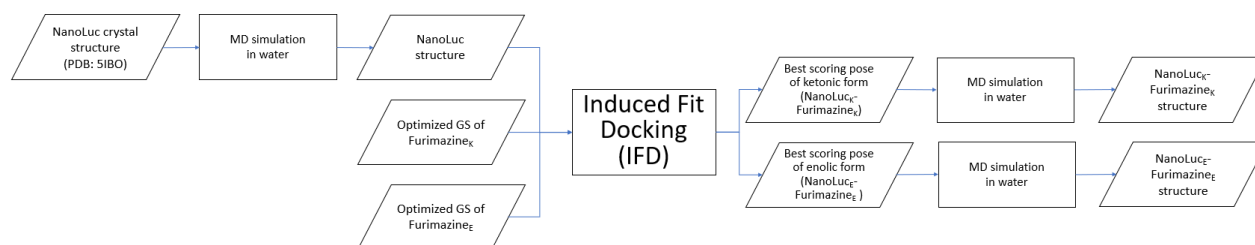


Figure 2. 9 - Workflow for NanoLuc-Furimazine_K and NanoLuc-Furimazine_E complexes construction

The resulting most representative frames were used to get two NanoLuc structures, namely NanoLuc_K and NanoLuc_E, respectively the structure of NanoLuc from the centre of the main cluster derived from the MD simulation of NanoLuc-furimazine_K complex and the analogous for NanoLuc-furimazine_E. These two luciferase conformations were used as receptor starting structures for another round of IFD, this time the ligands being three different forms of furimamide (a neutral one, FurA, a zwitterionic one, FurC and an anionic one, FurD), i.e. furimazine corresponding oxyluciferin. The ten highest ranking poses for

each starting structure were carefully inspected in order to find the best-scoring one, also presenting an arrangement of the furimamide core inside the NanoLuc enzymatic pocket analogous to the one found for furimamide. The six different furimamide-NanoLuc complex structures were subjected to a MD simulation in H₂O and a subsequent cluster analysis to get a representative frame for each of them, which was used as the starting point for a QM/MM optimization of geometry for the ligand excited state in order to get the final structure to predict the emission spectra for each pair. Final spectra calculations were carried out with a PBE0 single point calculation to correct the blue shift characterizing CAM-B3LYP results, as shown in Appendix 2.

2.8.1 Prediction of UV/Vis and NMR spectra of furimazine in DMSO

The prediction of the characterizing spectra of furimazine in DMSO has been done by quantum mechanical calculations in the DFT and TDDFT framework. CAM-B3LYP/6-31+G(d) was the level of theory chosen for the geometry optimizations, while the final UV/Vis spectra prediction has been performed with a single point PBE0 calculation with the same basis set. The NMR spectra, instead, have been obtained at B3LYP/6-311+G(d,2p) level of theory. Each predicted spectrum has been compared to the corresponding experimental result.

2.8.2 Modeling of furimazine oxidation mechanism in DMSO

The study of the reaction mechanism in DMSO has been done in the DFT and TDDFT framework, at the CAM-B3LYP/6-31+G(d,p) level of theory. Furimazine oxidation is an example of imidazopyrazinone (ImPy) oxidation and, as shown in previous studies⁴⁵, is a multistep process which can be divided into two different parts: oxygenation of ImPy to give a dioxetanone intermediate and its decomposition to form the final products, CO₂ and the oxidized compound. In this work, two different paths for the oxidation were evaluated, the so-called C2 and C3 oxygenation paths, and three different options for the decomposition were considered, one for each form of furimamide which has been taken into account in the previously described activity. The oxygenation process has been investigated on the ground state potential energy surface for both the C2 and the C3 pathways, considering as the starting point a biradical configuration obtained as a consequence of the SET which is supposed to be the first step. To model this particular situation, Broken Symmetry (BS) approach was used. For the dioxetanone decomposition, the investigation was performed by first optimizing the ground state of three of its different protonation states, since three different final products were object of the analysis and then proceeding with the optimization of the excited state final product starting from a modified geometry of the dioxetanone. In particular, the C-C and O-O bonds which are supposed to be broken during the reaction process were elongated enough to lead to the formation of the final products. Once the products have been found, the transition states (TS) were optimized and a single point excited state calculation was performed on all the geometries given by the IRC calculations, in order to evaluate the energy difference between the ground state and the excited state PESs to roughly evaluate the feasibility of the chemiexcitation required to get to the final emissive product.

3. Results

The main results obtained during the computational study of the NanoLuc-Furimazine system are collected and discussed in the present chapter, which is constituted by three main sections. In particular, the analysis of the enzyme-ligand complex is described first; the reactants and products of the catalytic process were investigated, giving a perspective on which molecular interactions are crucial for the enzymatic reaction.

The second part focuses on a further verification of the possible enol form involvement as a reactant by predicting the ^{13}C and ^1H NMR spectra and absorption and emission spectra of furimazine in DMSO, a solvent which is commonly used for imidazopyrazinonic compounds' experimental studies.

The third section reports findings from the investigation of the oxidation mechanism in DMSO and vacuum, judging how different solvents may change the course of the process and compare furimazine oxidation process with the one of a coelenterazine analogous, i.e. coelenterazine disulfonate.

3.1 Modeling of NanoLuc-furimazine and NanoLuc-furimamide complexes

Predicting oxyluciferin fluorescence inside the luciferase pocket is a specific example of the more general problem of modelling systems whose optical behaviour is influenced by the environment. The effects of the environment on the optical properties of the luminescent unit can be threefold:

1. a conformational constraint, imposed by the enzymatic pocket
2. a polarization effect on the electronic response
3. the direct participation of luciferase residues in the optical response

To address these different aspects, it is necessary to get a reliable molecular model of the protein-ligand pair and to employ a reliable QM level of theory for the simulation of the spectroscopic behaviour of the involved species.

3.1.1 Identifying the substrate: furimazine enol and keto forms

Compounds with an imidazopyrazinone core may exhibit keto-enol tautomerization. This implies the need to verify the difference in stability of the two forms in water solution, given the prevalence of this solvent in the cellular medium and the uncertainty on the form involved in systems similar to NanoLuc-furimazine²⁴². To do that, ground state (GS) geometry optimizations of two furimazine tautomeric forms, an enol and a keto one (reported in Figure 3.1), in aqueous solution (CPCM implicit solvent) were performed.

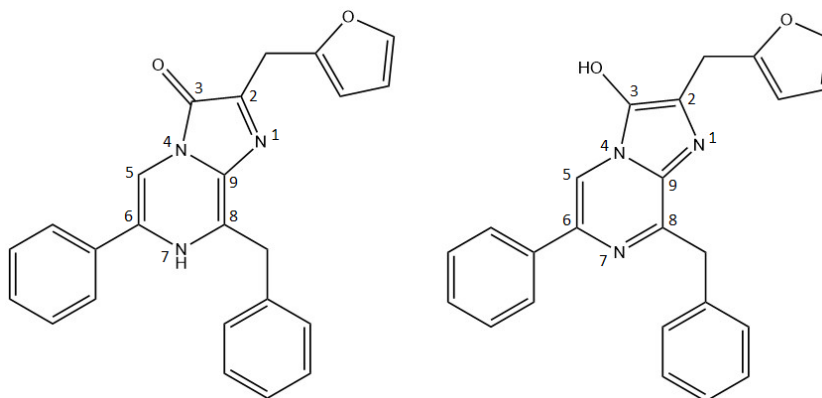


Figure 3. 1 - Sketch of the keto (left) and the enol (right) form of furimazine

The resulting energy difference, 2.5 kcal/mol in favour of the keto form. However, since solvent molecules could potentially interact strongly with the solute in ways that are not captured by a simple implicit solvent model, a mixed cluster-continuum approach has been considered. In this model, four explicit water molecules were added, one for each of the main hydrogen bonding sites, as reported in Figure 3.2.

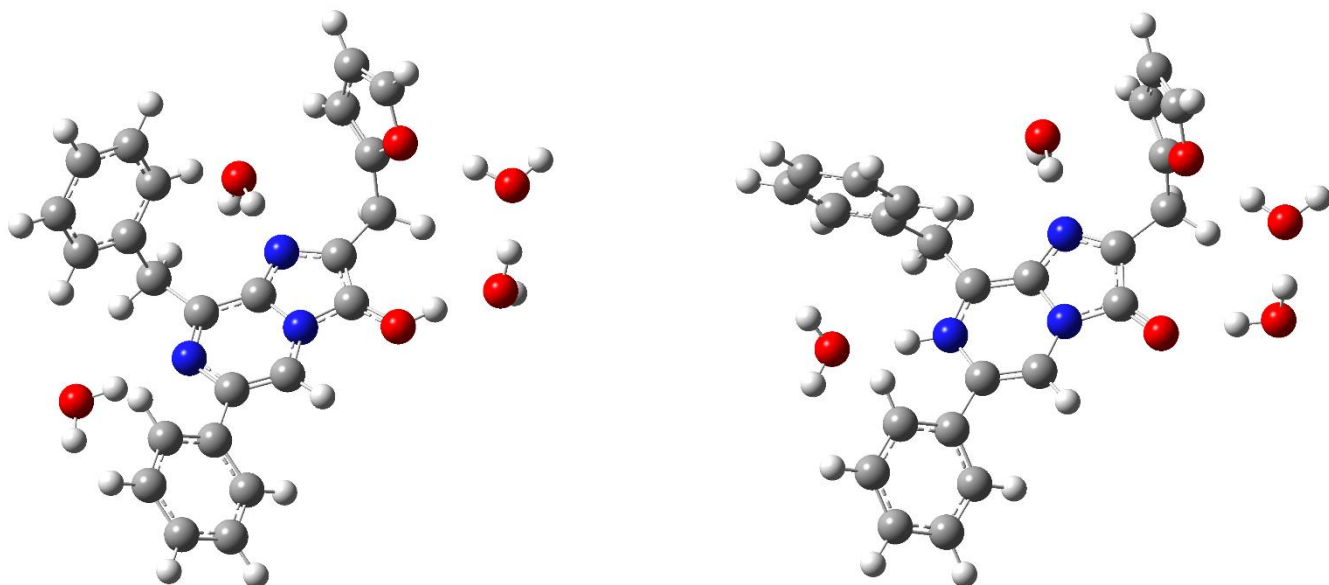


Figure 3. 2 - Optimized structures (Level of Theory: CAM-B3LYP/6-31+G(d), implicit solvation with CPCM, solvent=water) for enolic and ketonic forms of furimazine in presence of 4 explicit H₂O molecules

The result was a reduction of the energy gap, which was lowered to 0.04 kcal/mol, indicating that both tautomers can be simultaneously present in solution and eventually involved in the subsequent reaction. For this reason, both forms have been considered in the following steps. It should be stressed that only neutral furimazine has been considered for these calculations. Recently, it has been proposed that in Renilla-type luciferases, coelenterazine enters the binding pocket already with a deprotonated core⁶¹. This does not seem to be the case for furimazine, since NanoLuc peak activity occurs at a pH around 8 and still displays a residual 20% activity at pH=5, even in extracellular medium^{62,63}. Given that the pK_a of coelenterazine is supposed to be around 7.55 and assuming a similar one for furimazine, the peak activity range aligns well for both systems, yet the residual of luminescent activity at lower pH indicates that the deprotonation of furimazine before entering the enzyme pocket is not required for the enzyme activity.

3.1.2 Building the NanoLuc-furimazine complexes

Since, as remarked in subsection 2.8.1, no crystal structure for the NanoLuc-furimazine pair is available in the PDB, molecular docking was used to find out the placement of the ligand in the protein. In order to apply the IFD protocol, a starting structure for the protein was required; therefore, a MD simulation in water was performed on the PDB 5IBO entry, which corresponds to the NanoLuc crystal structure. After performing the trajectory cluster analysis, the centre of the main cluster (collecting 84% of the structures) was chosen as the enzyme reference structure to be employed for docking calculations.

In Figure 3.3 and Table 3.1 the first four clusters' centre structures and the results of the cluster analysis are reported.

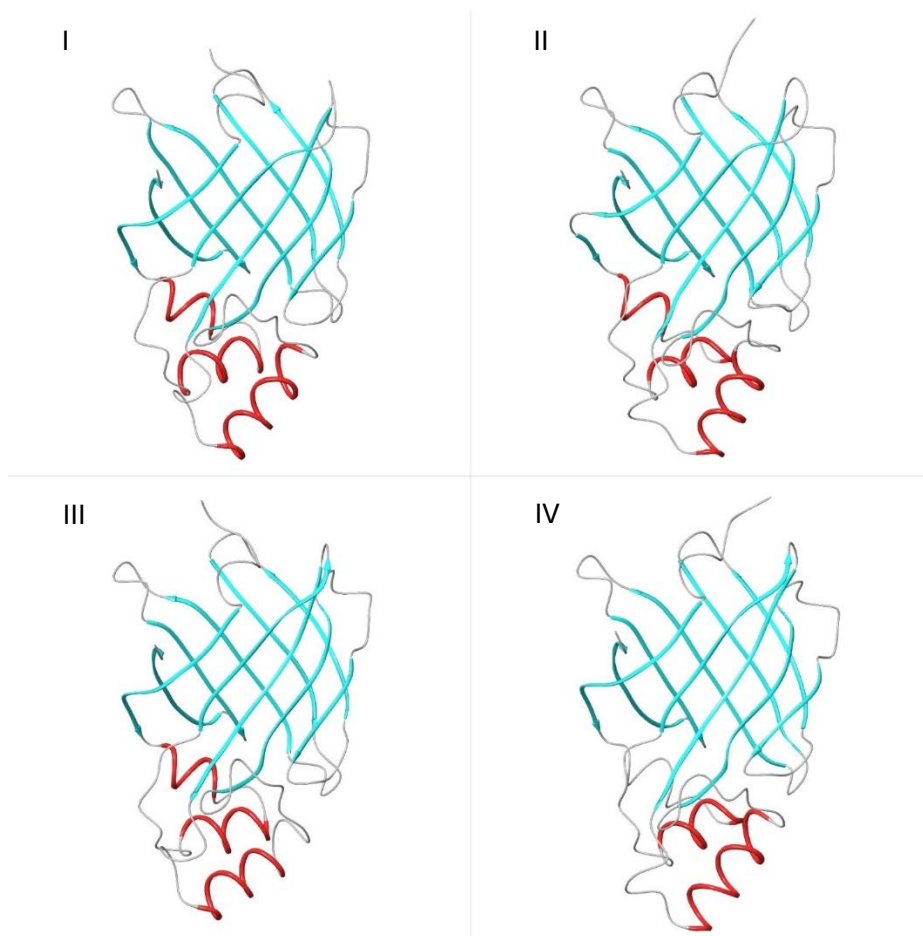


Figure 3. 3 - Centres of the four main clusters from the NanoLuc MD trajectory. Protein structure coloured according to secondary structure: alpha-helices in red and beta-strands in sky blue

<i>NLuc</i>				
% conf / cluster				
N° Clusters	I	I-II	I-III	I-IV
8	84	94	97	98

Table 3. 1 - Results of the cluster analysis on the *NLuc* MD trajectory

The stability of the protein structure has been furtherly verified with the DSSP algorithm; the resulting graphical representation being reported in Figure 3.4.

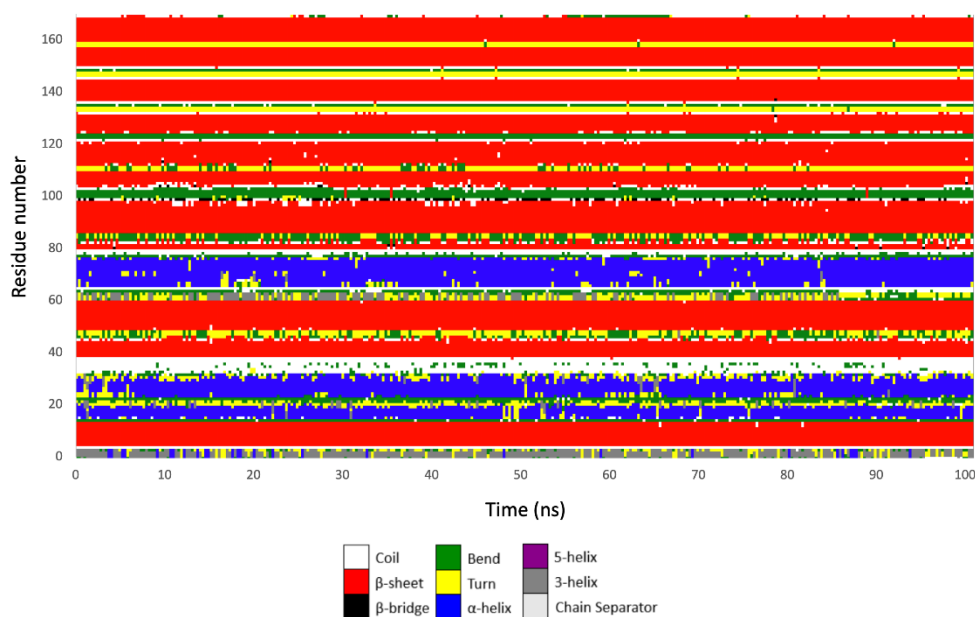


Figure 3. 4 - Graphical representation of DSSP algorithm result for NanoLuc MD simulation. Each secondary structure type is associated with a different colour, legend being reported under the plot

As it can clearly be seen from Figure 3.4, the protein backbone is found to be pretty stable, since the enzyme maintains its initial fold throughout the duration of the entire simulation, especially with regard to the β -sheets, which are almost never altered during the 100 ns of dynamics. After having selected the centre of the first cluster as the most representative structure for it, a first identification of NanoLuc most probable binding site was carried out with the SiteMap^{243,244} tool of the Schrodinger suite. The central cavity of the β -barrel of the enzyme was found to be the most probable location for the substrate interaction, as shown in Figure 3.5.

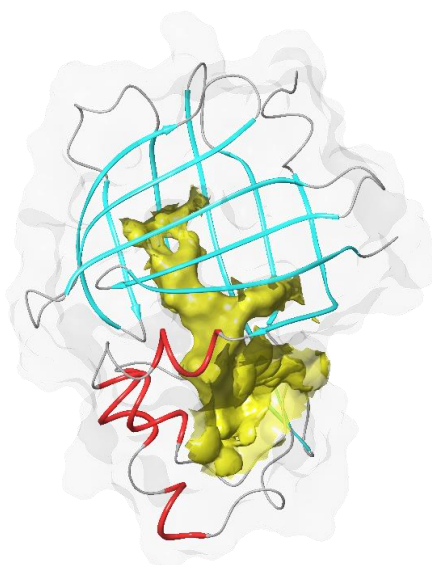


Figure 3. 5 - NanoLuc most probable binding site, in yellow, according to the SiteMap tool

To get to a higher precision identification of the binding site and build up a starting structure for the complexes, the previously described docking procedure was applied to both furimazine forms and the results, reported in Figure 3.6, highlight that both the keto and the enol form of furimazine are placed inside the β -barrel of NanoLuc, but with different orientations.

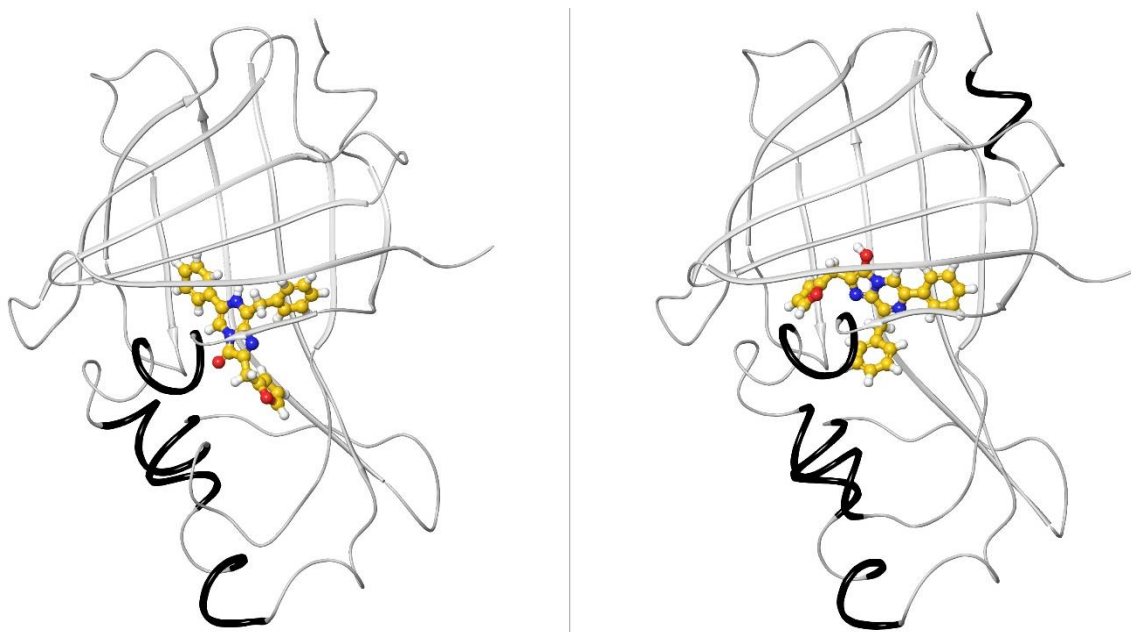


Figure 3. 6 - Comparison between the best-scoring docking pose of ketonic (left) and enolic (right) forms of furimazine

The docking scores for the best poses of the two forms suggest that the enol form is somewhat preferred with respect to the keto one, the docking score being -12.4 for the former and -11.2 for the latter, values that indicate very plausible docking poses¹⁸⁰. However, the specific orientation of the two forms is totally different as well as the network of their interactions with NanoLuc.

The keto form best-scoring pose is mainly characterized by vdW interactions with sidechains of non-polar residues like Phe151 and Ile56, as well as by a hydrogen bond involving Arg162 and N1 (Figure 3.7a).

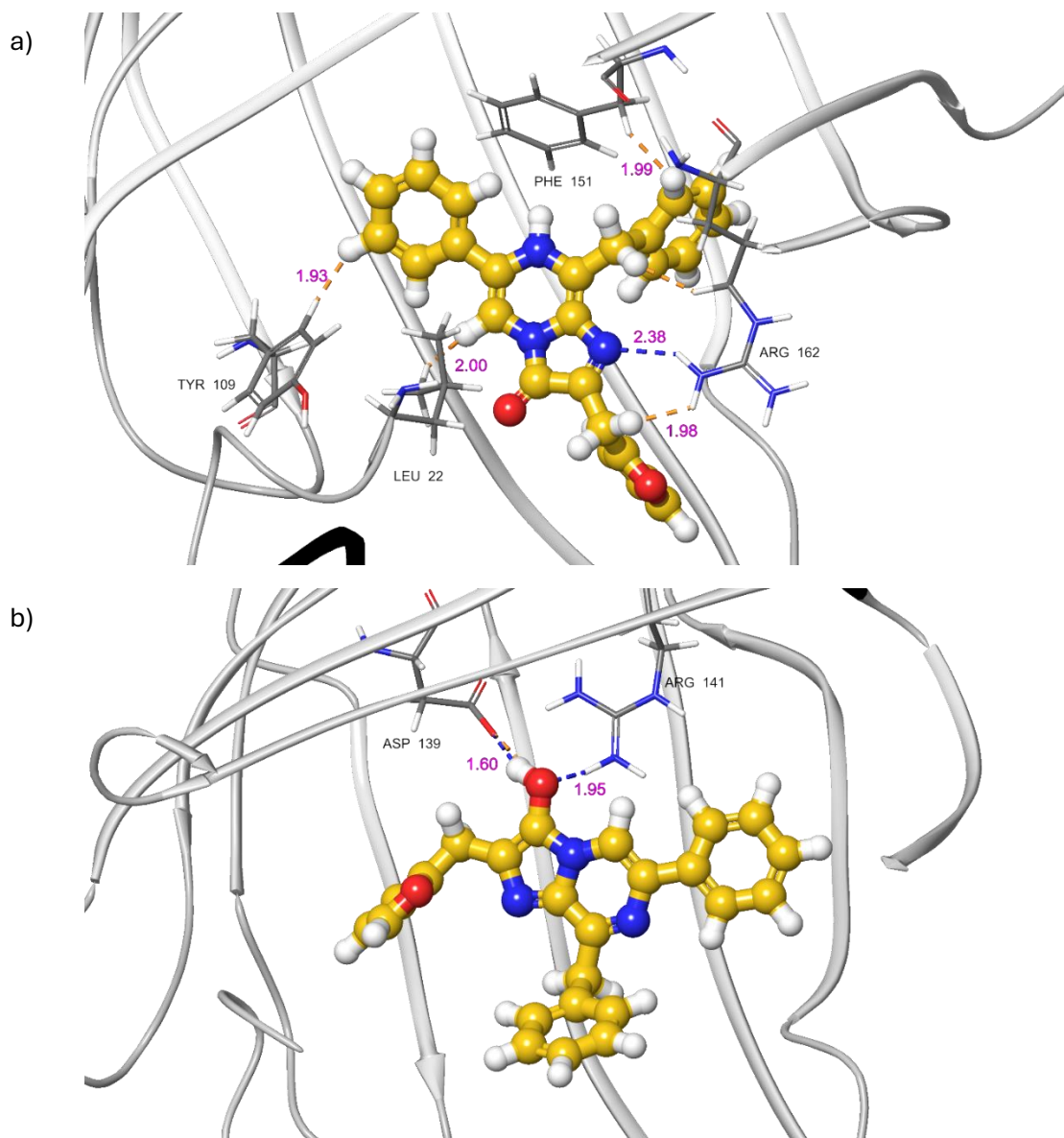


Figure 3. 7 - Best scoring pose of the keto (a) and enol (b) forms of furimazine, close-up on the ligand. Residues with one atom within 2 Å from the ligand represented as tubes, other residues hidden. Hydrogen bonds represented as dashed blue lines, close contacts as orange dashed lines and distances of interest reported in fuchsia

On the other hand, the enol pose features a pair of H-bonds involving the enol group, one with the aspartate Asp139, and the second one with an arginine, Arg141, as shown in Figure 3.7b, on top of additional vdW interactions like those involving residues Ile56 or Ile131. Interestingly, in the case of the keto form best-scoring pose, certain contacts between furimazine and the sidechains of some residues (Pro40, Phe151, Leu22, Arg162, Val38) correspond to clashes, whereas the best-scoring pose for the enol form presents just a single occurrence of such unfavourable interaction, i.e. the one involving the thiol group of Cys164.

To assess the stability of the substrate into the NanoLuc enzymatic pocket a MD simulation starting from each of the best-scoring pose obtained for the two forms of furimazine was performed. In both cases, furimazine remained inside the binding pocket without significant changes in position along the trajectory as shown by tracking the distance between furimazine and NanoLuc centres of mass reported in Figure 3.8.

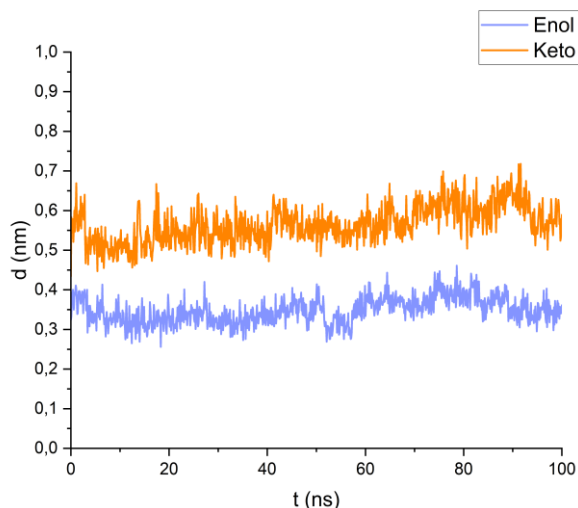


Figure 3. 8 - Distance between centre of mass of NanoLuc and centre of mass of furimazine enolic form (blue) and ketonic form (orange) in the corresponding MD simulations

Furthermore Figure 3.8 shows that the enol form of furimazine is a little bit deeper into the β -barrel than the corresponding keto tautomer, thus having a smaller distance between the protein centre of gravity and its own. MD simulations also allowed the investigation of the stability of the two H-bonds present in the best scoring pose of the enol form. The stability of the H-bond involving Asp139 is evident when analysing the distance between the hydrogen of the enol group and the oxygens of the carboxylate, shown in Figure 3.9 and at the cluster analysis results (Figure 3.10 and Table 3.2): the H-bond is present in the centres of the four main clusters (collecting 89% of the total structures).

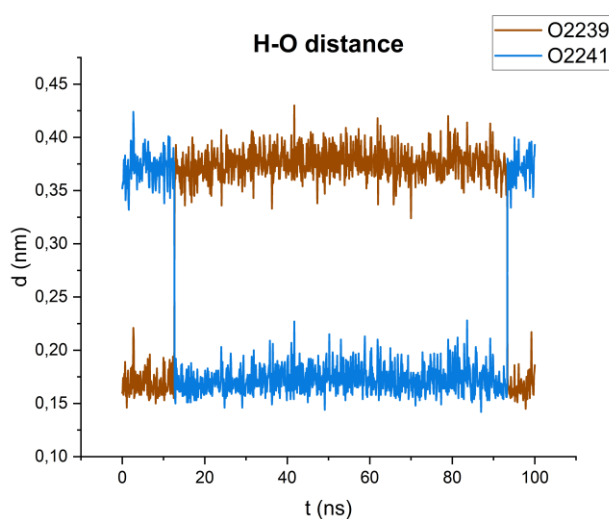


Figure 3. 9 - Distances between the enolic group hydrogen and the Asp139 carboxylate oxygens

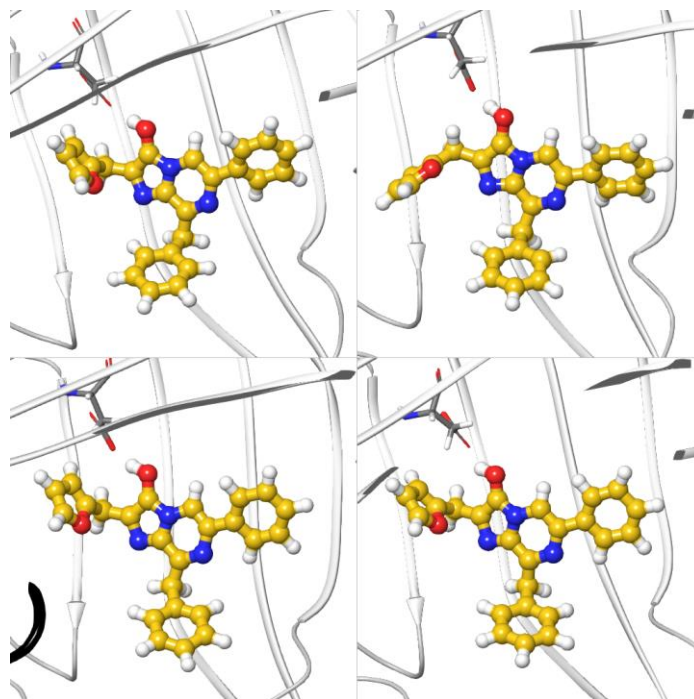


Figure 3. 10 - Centres of the four main clusters from the NanoLuc-furimazine (enol form) MD trajectory, zoom on furimazine. All residues hidden, only Asp139 shown in tubes

NLuc-furimazine (Enol form)				
% conf / cluster				
N° Clusters	I	I-II	I-III	I-IV
15	60	71	81	89

Table 3. 2 - Results of the cluster analysis on the NLuc-furimazine (enolic form) MD trajectory

On the other hand, the H-bond involving Arg141 is lost during the equilibration phase. The stabilizing effect of the Asp139 H-Bond is also evident looking at the interaction energy between the protein and the ligand at each frame of the MD simulation, calculated by the *gmx energy* tool of Gromacs. The trends in interaction energy values (Figure 3.11 and Tables 3.3 and 3.4) show little variability, with results from molecular dynamics conducted with enol furimazine featuring more favourable energy values throughout the whole duration of the simulation.

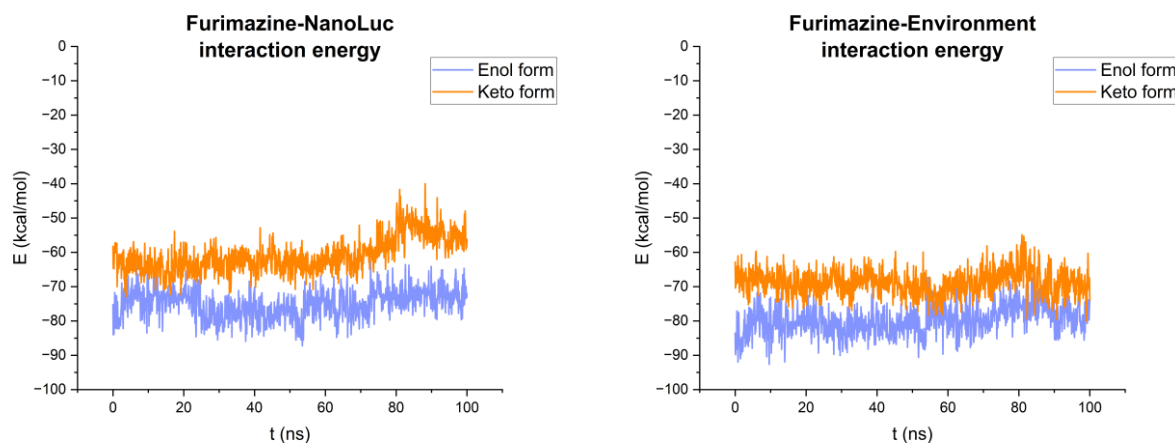


Figure 3. 11 - Interaction energies trends from the MD simulations for the NanoLuc-furimazine (enol and keto) pairs without (left) and including (right) H₂O contributions. Energy reported in kcal/mol

NanoLuc - Furimazine avg. Interaction Energy (H ₂ O excluded)		
	NanoLuc - <i>Furimazine_E</i>	NanoLuc - <i>Furimazine_K</i>
Energy (kcal/mol)	-74.94	-60.85

Table 3. 3 - Average interaction energies from the MD simulations for the NanoLuc - furimazine pairs (H₂O contributions excluded)

NanoLuc - Furimazine avg. Interaction Energy (H ₂ O included)		
	NanoLuc - <i>Furimazine_E</i>	NanoLuc - <i>Furimazine_K</i>
Energy (kcal/mol)	-79.96	-68.74

Table 3. 4 - Average interaction energies from the MD simulations for the NanoLuc - furimazine pairs (H₂O contributions included)

Indeed, the average interaction energy is -75.0 kcal/mol for the enol form and -61.8 kcal/mol for the keto tautomer. Since the MD simulations were performed in water, during the solvation step some solvent molecules were added in the protein cavities, including the catalytic site. In this case, the interaction energies between the ligands and the surrounding environment were recomputed considering also the solvent contributions and the average values fell to -80.0 and -68.7 kcal/mol for the enol and the keto tautomer, respectively. This difference of about 11 kcal/mol, makes the enol tautomer the most plausible candidate to take part in the reactive process. A similar gap in the interaction energies is obtained when performing MMGBSA calculations. Indeed, in this case, the estimated binding free energies are -53.3 and -45.5 kcal/mol for the enol and the keto tautomer, respectively (Table 3.5). While these values are still likely to overestimate the energy values, since they are affected by the errors of the underlying models, including solvent and entropy effects^{64,65}, the relative difference (-7.8 kcal/mol) reported for the two similar molecules can be considered meaningful and still supports the indication given by the bare interaction energies, thus further validating the enol tautomer as the most plausible candidate to participate in the reaction.

NanoLuc - Furimazine avg. MMGBSA Binding Energy		
	NanoLuc - <i>Furimazine_E</i>	NanoLuc - <i>Furimazine_K</i>
Energy (kcal/mol)	-53.28	-45.48

Table 3. 5 - Binding free energies from the MD simulations for the NanoLuc-furimazine pairs as computed according to MMGBSA model

The relevance of the H-bond involving the aspartate Asp139, becomes apparent when considering the hypothesized reaction mechanism, which requires the deprotonation of the substrate. Indeed, if furimazine must be deprotonated in order to react with the oxygen molecule, then the presence of a negatively charged sidechain residue in close proximity to the ligand could clearly play a key role; this appears to be especially the case here, as Asp139 takes part to an H-bond involving the furimazine proton that is supposed to be transferred.

3.1.3 Furimamide possible protonation states

Due to the lack of experimental data, the protonation state of the emissive product is not clear. In a previous theoretical study, eight different forms of Fur were considered and the neutral and the zwitterionic forms were found to be “dominant in the whole pH range in the aqueous solution in GS and ES, respectively”. In addition to these two, the corresponding amide anion should be also considered as a possible emissive product, since in other luciferin-luciferase systems^{217,245} a deprotonated form of the product is involved in the emission. Furthermore, the hypothesized catalytic mechanism is thought to start with a deprotonation of the substrate, thus making it possible to generate an anionic species as product. These are the reasons why three different protonation states of furimamide (the neutral form, FurA, the zwitterionic form, FurC, and the anionic one, FurD reported in Figure 3.12) have been considered in this study.

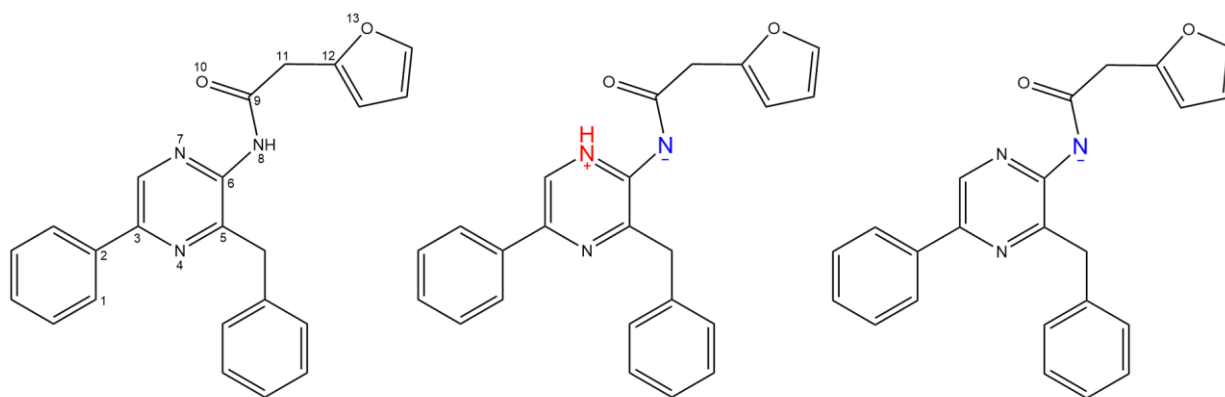


Figure 3. 12 - Sketches of the considered forms of furimamide: FurA (left, neutral), FurC (middle, zwitterion) and FurD (right, anion). Nomenclature taken from reference²⁴⁶.

Aiming at a more comprehensive analysis, a fourth form was later added to the pool, the iminol tautomer of FurA, with the proton moving from N8 to O10, reported in Figure 3.13.

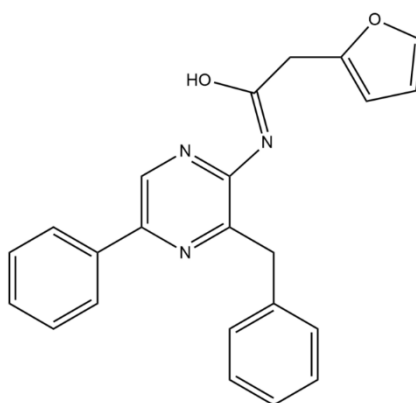


Figure 3. 13 - Sketch of furimamide iminol tautomer

The GS of each form was optimized in vacuo and in aqueous solution at the CAM-B3LYP/6-31+G(d) level. On the optimized geometries, TD-DFT single-point calculations have been performed to find the brightest ES and subsequently optimize it.

These calculations led to the early exclusion of the iminol form from the pool of candidates, since it was found to be 2.3 kcal/mol higher in energy at the GS and 8.4 kcal/mol higher in the ES (in aqueous solution) if compared to FurC, which is almost identical in terms of geometry, as shown in Figure 3.14 except for the proton location.

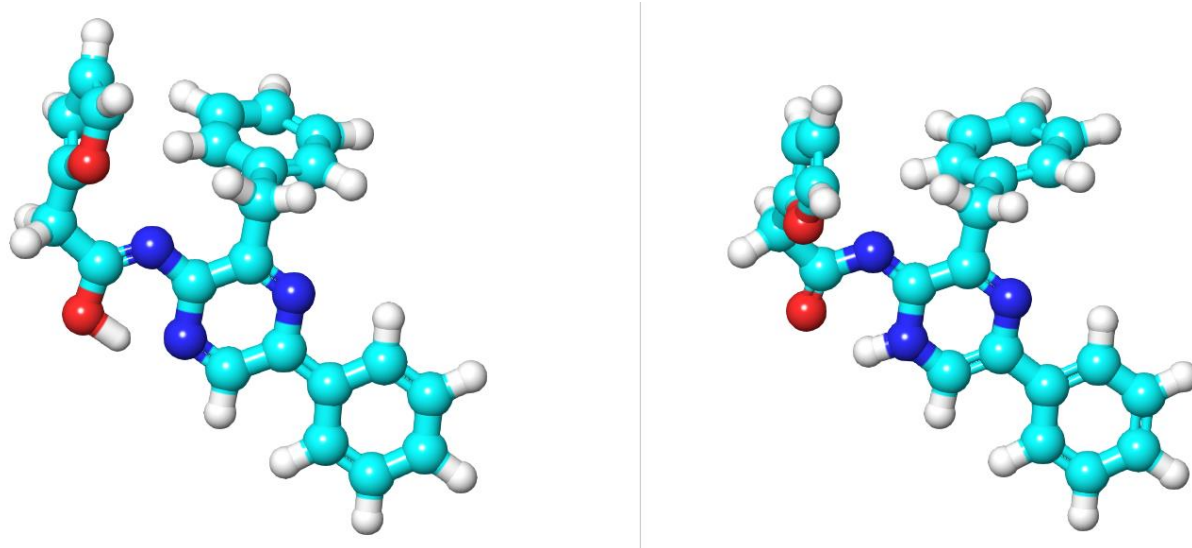


Figure 3. 14 - Comparison of the aqueous solution optimized ES structures of iminol form of furimamide (left) and FurC (right). Level of theory: CAM-B3LYP/6-31+G(d), implicit solvation with CPCM, solvent=water

Furthermore, a flexible scan on the O10-H (see Figure 3.12 for numbering) distance in the ES was performed, in order to get a rough estimate of the barrier for the conversion of this species to FurC, i.e. the proton transfer, and it was found to be only 1.3 kcal/mol. A similar result was obtained in the GS, where the research of the transition state connecting the two minima was done: the energy barrier is higher, 3.2 kcal/mol, but still not prohibitive. Results of the scan for the analysis of the ES proton transfer and the IRC calculation for the GS one are reported in Figure 3.15. These numbers suggest that this form is strongly unfavourable and, therefore, has not been furtherly considered.

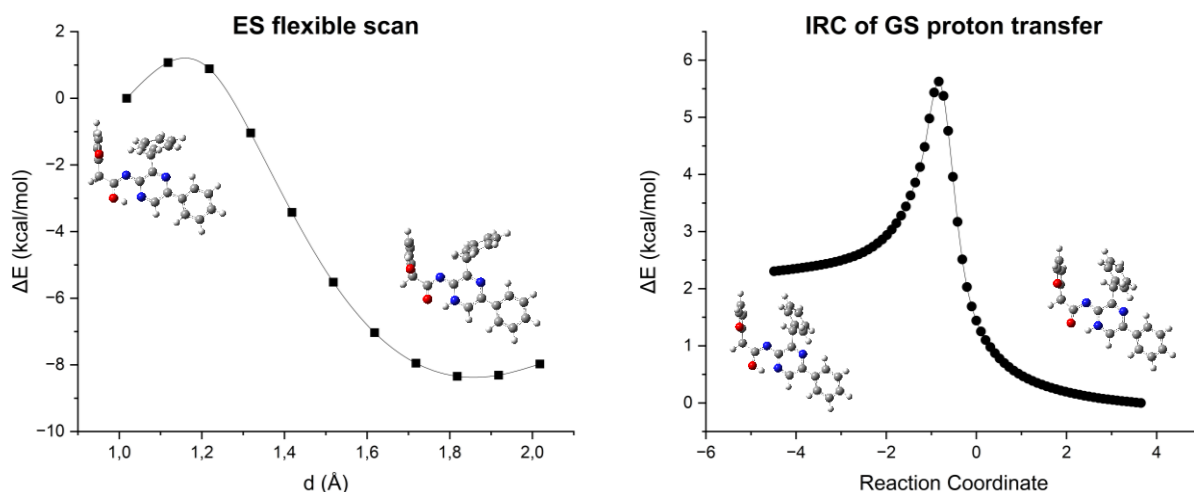


Figure 3. 15 - Results of calculations performed to evaluate the energy barrier for the O10-N7 proton transfer: ES flexible scan of O10-H distance (left) and IRC path for proton transfer in GS (right). In both plots, FurC is on the right, while iminol form of furimamide is on the left. Level of theory: CAM-B3LYP/6-31+G(d), implicit solvation with CPCM, solvent=water

For the remaining three furimamide forms, Table 3.6 (in vacuo results) and Table 3.7 (in water solution results) report for the first five excited states the relative vertical absorption energies, calculated at the GS geometry, and the corresponding oscillator strengths, paired with the vertical transition energy for the emission calculated at the optimized ES geometries.

FurA							
Excitation				Emission			
Excited state	ΔE (eV)	λ (nm)	f	ΔE (eV)	λ (nm)	f	Contribution of LUMO -> HOMO
S1	4.35	285	0.20	3.54 ^a	350	0.49	82%
S2	4.53	273	0.41				
S3	4.96	250	0.15				
S4	5.05	245	0.03				
S5	5.20	239	0.04				
FurC							
Excitation				Emission			
Excited state	ΔE (eV)	λ (nm)	f	ΔE (eV)	λ (nm)	f	Contribution of LUMO -> HOMO
S1	3.83	324	0.28	3.05	406	0.23	92%
S2	3.85	322	0.10				
S3	4.49	276	0.31				
S4	4.52	274	0.21				
S5	4.85	255	0.00				
FurD							
Excitation				Emission			
Excited state	ΔE (eV)	λ (nm)	f	ΔE (eV)	λ (nm)	f	Contribution of LUMO -> HOMO
S1	3.65	340	0.51	3.14	395	0.29	75%
S2	3.94	314	0.26				
S3	4.15	299	0.04				
S4	4.48	277	0.02				
S5	4.52	274	0.02				

Table 3. 6 - In vacuo vertical excitations corresponding to the First Five Singlet ESs at the GS optimized geometries and emissions of the Optimized ES at the ES optimized geometries, and corresponding oscillator strengths (f), calculated at the CAM-B3LYP/6-31+G(d) level.

^a The emitting ES is the lowest in energy after the optimization. It is reported as S2 because it was the second lowest at the ground state geometry, which was the starting point for the geometry optimization of the ES.

<i>FurA</i>							
Excitation				Emission			
Excited state	ΔE (eV)	λ (nm)	f	ΔE (eV)	λ (nm)	f	Contribution of LUMO -> HOMO
S1	4.30	289	0.58	3.21	386	0.94	95%
S2	4.47	278	0.33				
S3	4.97	249	0.25				
S4	5.09	244	0.08				
S5	5.20	238	0.07				

<i>FurC</i>							
Excitation				Emission			
Excited state	ΔE (eV)	λ (nm)	f	ΔE (eV)	λ (nm)	f	Contribution of LUMO -> HOMO
S1	3.77	329	0.54	2.98	417	0.56	95%
S2	3.99	311	0.00				
S3	4.53	274	0.52				
S4	4.62	269	0.03				
S5	4.99	249	0.00				

<i>FurD</i>							
Excitation				Emission			
Excited state	ΔE (eV)	λ (nm)	f	ΔE (eV)	λ (nm)	f	Contribution of LUMO -> HOMO
S1	3.97	312	0.32	3.03	409	0.88	94%
S2	4.31	288	0.18				
S3	4.56	272	0.37				
S4	4.81	258	0.10				
S5	4.99	249	0.00				

Table 3. 7 - Aqueous solution vertical excitations corresponding to the first five singlet ESs at the GS optimized geometries and emissions of the optimized ES at the ES optimized geometries, and corresponding oscillator strengths (f), calculated at the CAM-B3LYP/6-31+G(d) level. Solvent model: CPCM

Concerning absorption in vacuum, there is a red shift when going from FurA to FurD, with the first intense transition going from 285 nm (4.4 eV) to 340 nm (3.7 eV). Interestingly, the absorption of FurC is similar to the one of the anionic form (FurD), showing a peak at 324 nm (3.8 eV). Situation is slightly different in water, since FurA and FurC absorptions are almost not shifted if compared to the corresponding values in vacuum, while FurD is blue-shifted to 4.0 eV. Emission bands for the three species appear at higher wavelengths in both vacuum and aqueous solution, the Stokes shift decreasing across the forms in vacuum, from -0.9 eV of FurA to -0.5 eV of FurD. In water, FurA shift increases to -1.1 eV and FurD one is almost doubled, -0.9 eV, with FurC maintaining the -0.8 shift characterising this form in vacuum. These large shifts can be attributed to the significant structural changes. In FurA, relevant differences are observed between the ground state and the excited state, mostly concerning the conformation of the central ring. In the ground state, the pyrazinic core is nearly planar, whereas, upon excitation, the molecule undergoes a noticeable conformational change resulting in a flexed central ring. This change is characterized by an increase of the N7-C5-N4-C3 dihedral angle (from 0° to 9°, see Figure 3.12 for atom labelling), indicating a departure from planarity, with a partial pyramidalization of C3 and C6. The phenyl group is also involved in this rearrangement, with a strong reduction of its tilting: the dihedral angle defined by the moiety itself and the central core goes from -23° to 2°, suggesting an enhanced conjugation at the excited state.

In the case of FurC, the structural changes are limited to the rotation of the benzyl group, more coplanar to the central ring in the ES, and to the phenyl group, which is less tilted in the ES optimized geometry, with the dihedral defined by the central ring and the phenyl itself reduced from -24° to -12° . The bond distances of atoms belonging to the pyrazinic core of the molecule are the most interesting values, with alternating elongations and contractions of the six bonds going from the GS to the ES, a common feature of π - π^* transitions in conjugated systems. FurD main structural changes between GS and ES optimized structures concern almost only the phenyl group. It shows a similar behaviour to FurC corresponding moiety: in the GS the dihedral angle between the phenyl and the central core measures -19° , while in the ES is of just 2° . This structural modification is paired with some changes in the bond lengths for the central core, but their average magnitude is smaller than the one characteristic of FurC pyrazinic ring bonds and the pattern highlighted for the zwitterionic form is less evident here.

To have a better understanding of how these structural changes could influence the Stokes shift, an orbital analysis was performed. Since in all furimamide forms, both *in vacuo* and in aqueous solution, the emission is almost fully due to a LUMO $\rightarrow$ HOMO transition, *in vacuo* results were employed to analyse the HOMO and LUMO characteristics, with focus on the contribution of different moieties to such orbitals. The analysis of Figure 3.16 reveals that the HOMO is mainly distributed over the pyrazine and the phenyl for all forms and, in the case of FurC only, on the furan moiety. The LUMO is largely localized on the amide group and the pyrazine ring in all cases, with a significant contribution of the phenyl moiety only in the case of FurA and FurD. These findings are coherent with the analysis of GS and ES structures already discussed, thus further highlighting the key role of the central ring in explaining FurA and FurC behaviour and the relevance of the phenyl orientation for FurA and FurD.

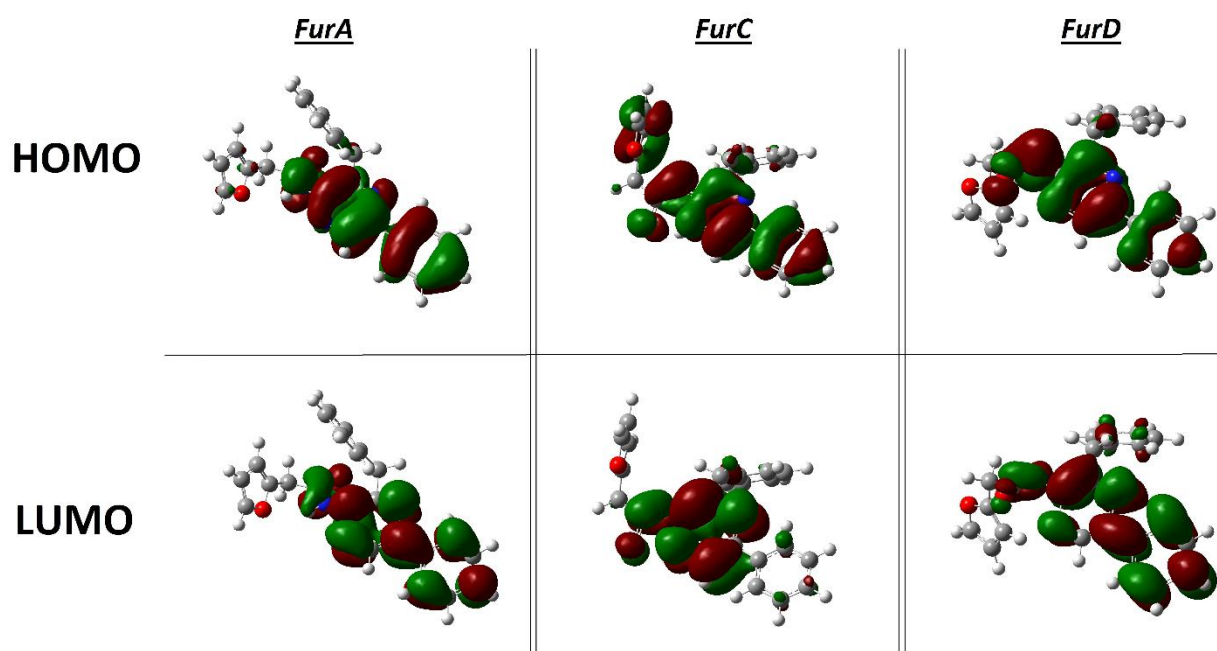


Figure 3. 16 - HOMO and LUMO of FurA, FurC and FurD species at the ES optimized geometries *in vacuo*

As far as the relative stability of the two different neutral species *in vacuo* is concerned, optimization results show that FurA is 6.2 kcal/mol more stable than FurC in the GS, while in ES it is 5.4 kcal/mol higher. The situation is similar in aqueous solution: FurA is 5.6 kcal/mol lower than FurC in the GS, whereas in ES it is 5.2 kcal/mol higher, confirming what was found by Sahihi et al.²⁴⁶.

3.1.4 Building the NanoLuc-furimamide complexes

Once the ES optimized structures for furimamide had been obtained, the following step was the construction of the NanoLuc-furimamide complexes. This step has been completed with another round of IFD calculations, this time considering the aqueous solution ES optimized geometries of FurA, FurC, and FurD as ligand starting structures. For the enzyme one, instead, two different structures have been employed, namely NanoLuc_K and NanoLuc_E. The former is obtained from the centre of the main cluster derived from the MD simulation of the keto form of furimazine-NanoLuc complex, whereas the latter is obtained from the MD simulation of the enol form of furimazine bound to the enzyme. Both structures feature the same fold of the apoprotein, and they just differ for the residues' sidechains orientation and displacement, as a consequence of the interactions with the two forms of furimazine considered. The ten highest ranking poses for each starting structure were then carefully inspected in order to find the best scoring one, also presenting an arrangement of the furimamide core inside the NanoLuc enzymatic pocket analogous to the one found for furimazine. In the case of FurD in the Enol form, Asp139 was converted into Ash139, i.e. the protonated form, since this is the residue which is supposed to deprotonate furimazine and the anion form may exist only if the proton has not been retransferred on the ligand.

The selected structures were subject to a MD simulation in water and the trajectories were analysed by means of direct observation and cluster analysis. These simulations revealed that the output geometries of the docking procedure are quite stable along the trajectories as indicated by the relatively small RMSD values for both the protein and the ligand (shown in Figures 3.17 and 3.18).

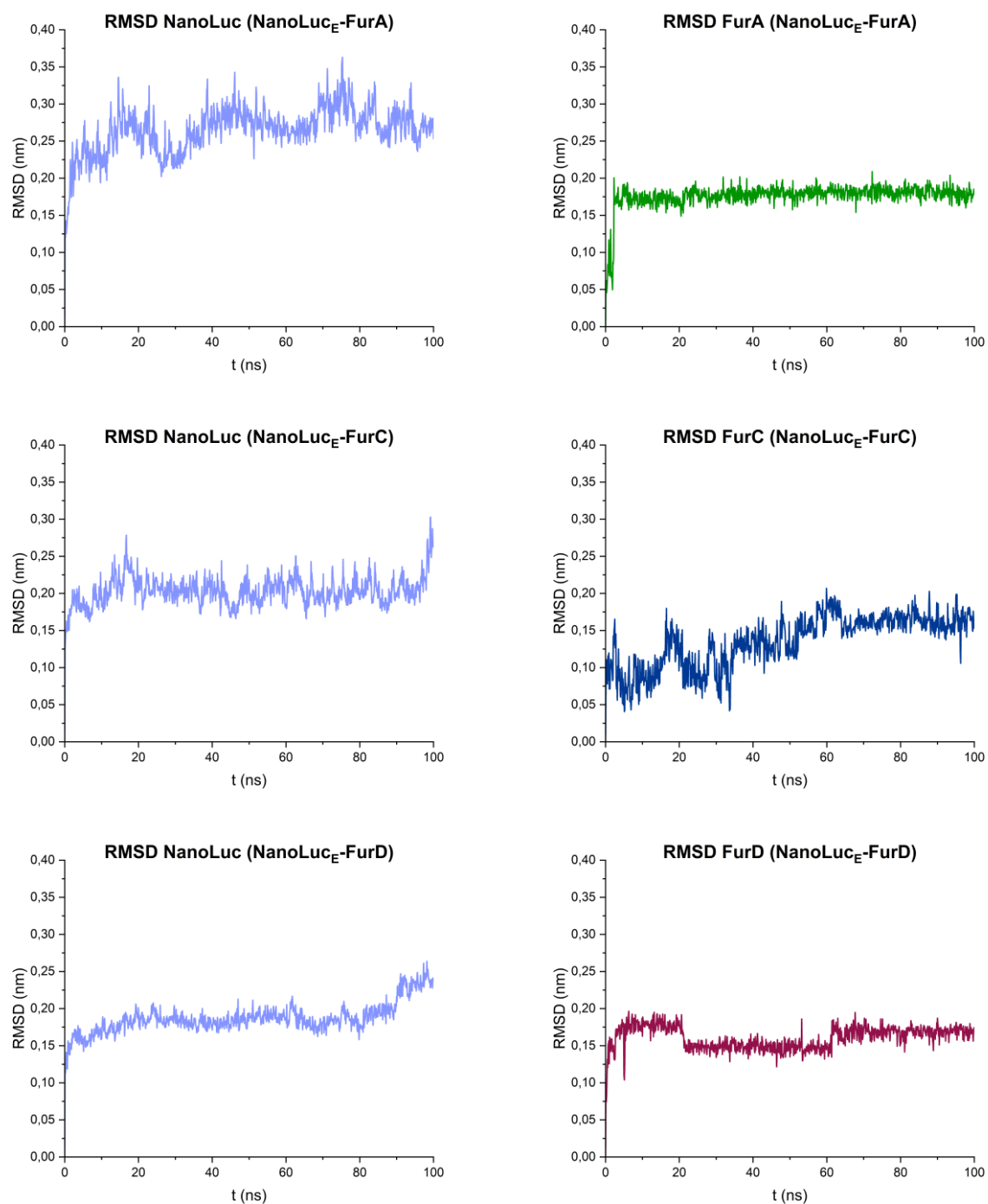


Figure 3. 17 - NanoLuc and Fur* RMSD for Enol pose MD simulations. The RMSD trends for the protein structure are reported in the plots on the left, while the corresponding ones for the ligand are on the right.

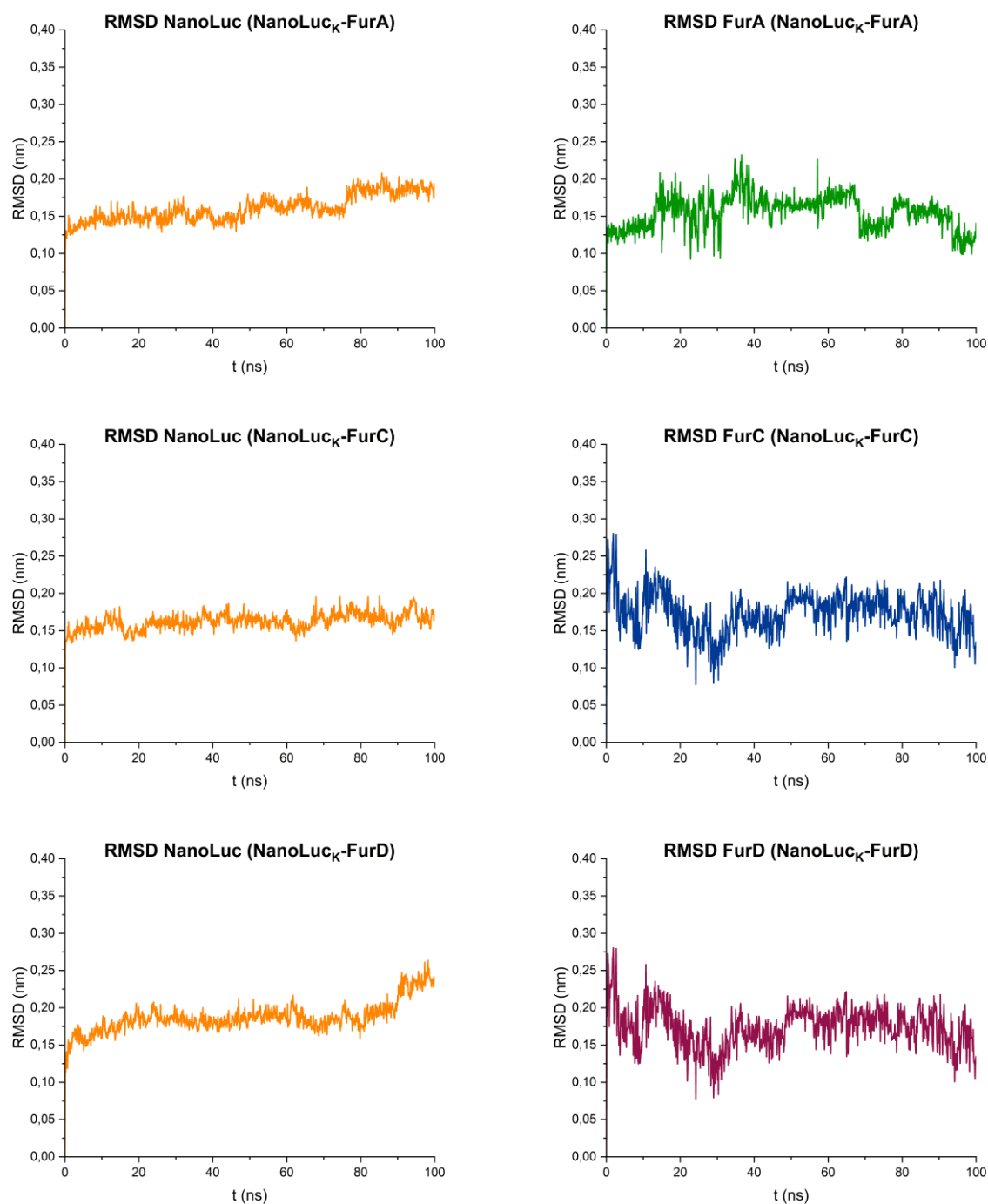


Figure 3. 18 - NanoLuc and Fur* RMSD for Keto pose MD simulations. The RMSD trends for the protein structure are reported in the plots on the left, while the corresponding ones for the ligand are on the right.

Interestingly, strong specific interactions between the protein and the molecule are found only for NanoLuc_K-FurC and NanoLuc_E-FurD pairs: the first presents vdW interactions, like the other pairs, but also a hydrogen bond involving the oxygen of the amide group as acceptor and the positively charged termination of the Arg162 sidechain as donor. NanoLuc_E-FurD shows a different hydrogen bonding network instead (see Figure 3.19), which still involves an arginine sidechain (Arg141) as donor, while the acceptor is a nitrogen atom of furimamide (N7), but also the oxygen of the furan ring (O13). The hydrogen bonds characterizing NanoLuc_E-FurA and NanoLuc_E-FurC pairs involve, in both cases, water molecules, but they deserve nonetheless a mention, since one of these water molecules lies exactly between the ligands and Asp139 as shown in Figure 3.19.

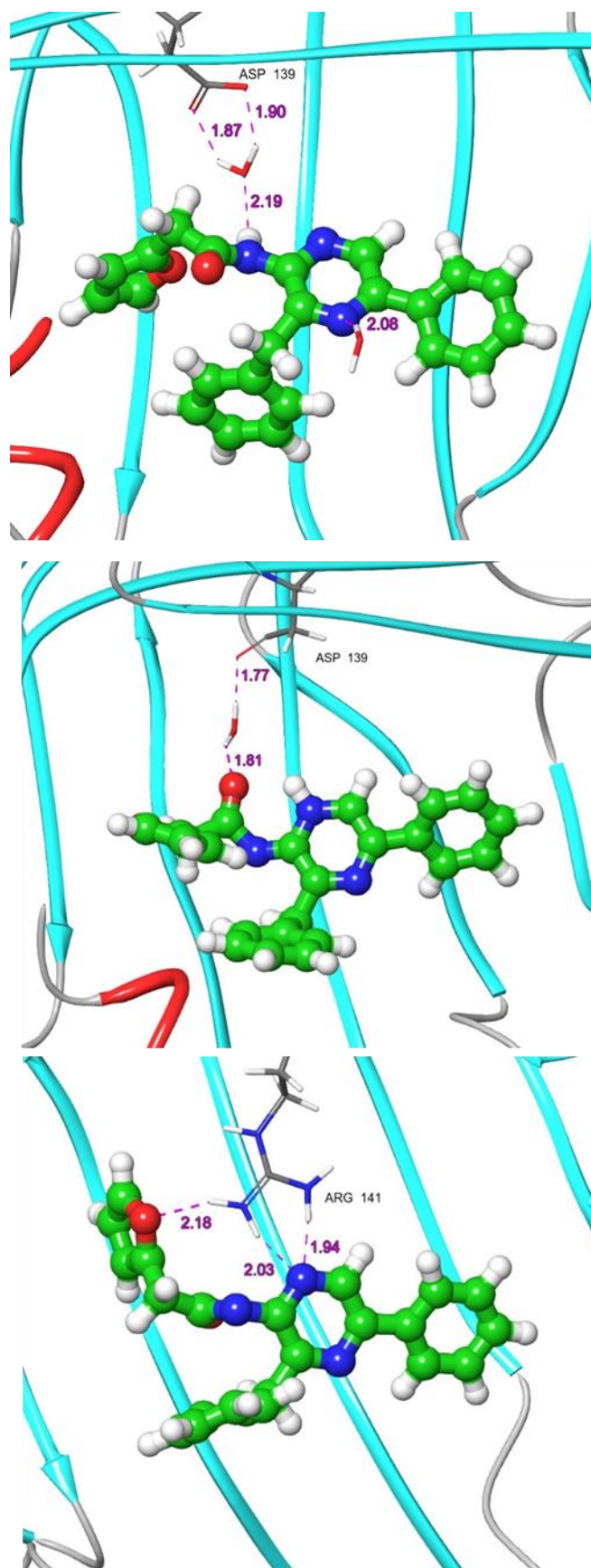


Figure 3. 19 - Centre of the main cluster of MD simulations on FurA, FurC and FurD in NanoLuc_E, close-up on the ligand. Residues and water molecules directly involved in H-bonds with the ligand represented as tubes, other residues hidden and distances of interest reported in fuchsia

Interestingly the NanoLuc_E-FurC and the NanoLuc_K-FurA pairs do not show any specific interaction with the closest protein residues for the entire duration of the simulations. Indeed, these two systems present less favourable values of the average protein-ligand interaction energies (reported in Figure 3.20) with respect to the other pairs (-49.4 and -48.0 kcal/mol, respectively; values reported in Table 3.8).

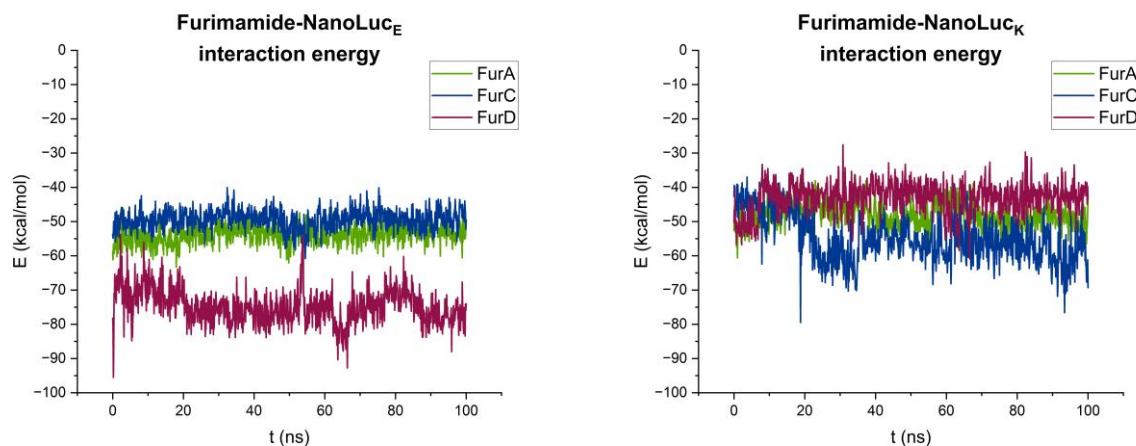


Figure 3. 20 - Interaction energies trends from the MD simulations for the NanoLuc-furimamide pairs without H₂O contributions; results for NanoLuc_E protein environment on the left, NanoLuc_K on the right. Energy reported in kcal/mol

NanoLuc - Fur* avg. Interaction Energy (H ₂ O excluded)			
	<i>FurA</i>	<i>FurC</i>	<i>FurD</i>
NanoLuc_E	-53.82	-49.42	-74.80
NanoLuc_K	-47.99	-54.91	-43.04

Table 3. 8 - Average interaction energies from the MD simulations for the NanoLuc-furimamide pairs (H₂O contributions excluded)

Also the NanoLuc_K-FurD pair presents a quite small interaction energy (-43.0 kcal/mol), but in this case the ligand only partially enters into the cavity and thus it has larger solvent exposure. If the interaction energies are calculated considering also the contribution of the ligand interactions with the surrounding water molecules (Figure 3.21 and Table 3.9), NanoLuc_K-FurD becomes the most stable pair (-100.2 kcal/mol) while NanoLuc_E-FurC and NanoLuc_K-FurA continue to show the weakest interaction energy (-66.5 and -63.6 kcal/mol, respectively).

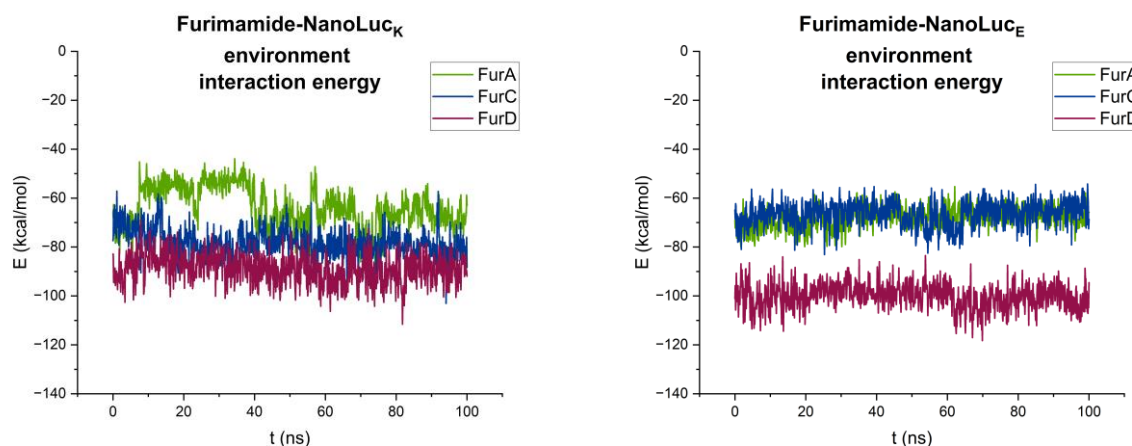


Figure 3. 21 - Interaction energies trends from the MD simulations for the NanoLuc-furimamide pairs with H₂O contributions; results for NanoLuc_E environment on the left, NanoLuc_K on the right. Energy reported in kcal/mol

NanoLuc - Fur* avg. Interaction Energy (H ₂ O excluded)			
	<i>FurA</i>	<i>FurC</i>	<i>FurD</i>
NanoLuc_E	-68.10	-66.48	-100.14
NanoLuc_K	-63.58	-78.14	-100.19

Table 3. 9 - Average interaction energies from the MD simulations for the NanoLuc- furimamide pairs (H₂O contributions included)

A smaller value for this quantity, especially if compared to the one computed for the reactant, could explain the high light intensity of the NanoLuc-furimazine pair^{247,248} as it reduces the risk of product inhibition and favours the release of the enzyme following the catalysis. For the possible products, the picture emerging from the MMGBSA analysis, whose results are reported in Table 3.10 is slightly different from both quantitative and qualitative aspects.

NanoLuc - Fur* avg. MMGBSA Binding Energy			
	<i>FurA</i>	<i>FurC</i>	<i>FurD</i>
NanoLuc_E	-40.23	-41.35	-48.57
NanoLuc_K	-34.73	-40.57	-27.13

Table 3. 10 - Binding free energies from the MD simulations for the NanoLuc-furimazine pairs as computed according to MMGBSA model

Here, the binding free energies range between -48.6 (NanoLuc_E-FurD) to -27.1 (NanoLuc_K-FurD) kcal/mol, all numbers, as expected, significantly lower in absolute value with respect to the simple interaction energies, but unfortunately being less discriminating. Indeed, the weakest interactions, that should, as mentioned, favour the release of the product, are found for NanoLuc_K-FurD (-27.1 kcal/mol), followed by NanoLuc_K-FurA (-34.7 kcal/mol) and all the others being close to each other (within 8 kcal/mol).

3.1.5 Prediction of the emission spectra

The last step is the prediction of the fluorescent properties of the furimamide forms within the enzyme. To this end, QM/MM geometry optimizations of the brightest ES were performed on the most representative structure of each trajectory. Normalized emission spectra for all NanoLuc-Fur pairs are shown in Figure 3.22 and associated maximum emission wavelengths are reported in Table 3.11. These spectra do not include NanoLuc_K-FurC pair, for which it was not possible to localize an ES structure.

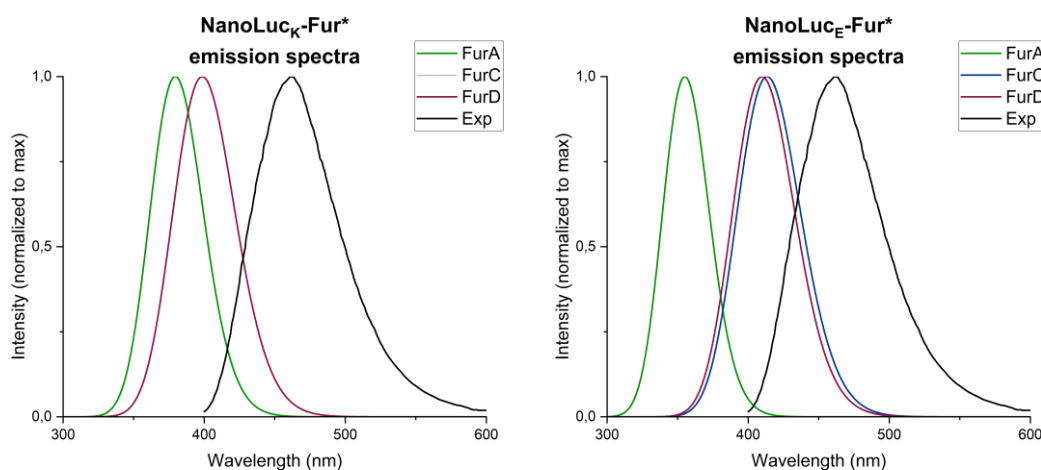


Figure 3. 22 - Normalized emission spectra of NanoLuc-Fur* systems deriving from the best-scoring pose of keto furimazine (left) and from the best-scoring pose of enol furimazine (right). Experimental emission spectrum is reported for comparison

Maximum emission wavelength ($\lambda_{\max}$)				
	NanoLuc _E	<i>f</i>	NanoLuc _K	<i>f</i>
FurA	355	0.48	380	0.03
FurC	413	0.10	-	-
FurD	408	0.28	409	0.20
Exp	460			

Table 3. 11 - Computed maximum emission wavelength ($\lambda_{\max}$, nm) and associated oscillator strengths (*f*) for the NanoLuc-Fur* pairs and experimental value from reference¹³⁴.

The NanoLuc_E-FurC (413 nm), NanoLuc_E-FurD (408 nm), and NanoLuc_K-FurD (409 nm) pairs feature maxima very close to each other, blue-shifted by about 50 nm with respect to experimental value (460 nm)⁵⁴. Instead FurA gives largely blue-shifted values (by 105 and 80 nm, respectively for NanoLuc_E and NanoLuc_K). These latter amply justify the exclusion of FurA from the pool of product candidates.

To analyse the effect of the protein environment on emission, Table 3.12 reports the maximum emission wavelengths calculated on the ES geometries optimized in the protein environment but *in vacuo*.

Maximum emission wavelength (nm)		
	<i>In vacuo</i> @ NanoLuc _E -Fur*	<i>In vacuo</i> @ NanoLuc _K -Fur* <i>In vacuo</i>
FurA	339	375
FurC	430	-
FurD	416	426

Table 3. 12 - Maximum emission wavelength (nm) calculated *in vacuo* (PBE0/6-31G**+) for the different forms of Fur* at the geometries adopted in the enzymatic pockets. For comparison, values calculated *in vacuo* at the *in vacuo* ES optimized geometries (PBE0/6-31G**+/CAM-B3LYP/6-31G**+) are also reported.

A slight shift (within 16 nm) is observed for FurA and a small red-shift (within 17 nm) is observed for FurC and FurD with respect to the value obtained in the protein environment. These results suggest that the polarization effects of the enzymatic pocket are quite small. On the other hand, the maximum emission wavelengths calculated *in vacuo* on the ES geometry optimized *in vacuo* (see Table 3.6) show a significant red-shift for FurC (38 nm) with respect to the values calculated for the structures optimized in the enzyme pocket. Since the *in vacuo* optimized geometries display quite different conformations with respect to those calculated in the enzymatic pocket, the conformational constraints enacted by the enzymatic scaffold could constitute the dominant factor modulating the maximum emission wavelength. To assess this and identify the relevant moieties involved, the influence of the conformational constraints imposed by the enzymatic site on FurA, FurC and FurD was analysed. In the following analysis, only the NanoLuc_E-Fur* pairs, since data suggest that the enol form of the furimazine is likely the most probable substrate.

The structures of the three forms of furimamide in the enzymatic pocket and *in vacuo* are reported in Figure 3.23.

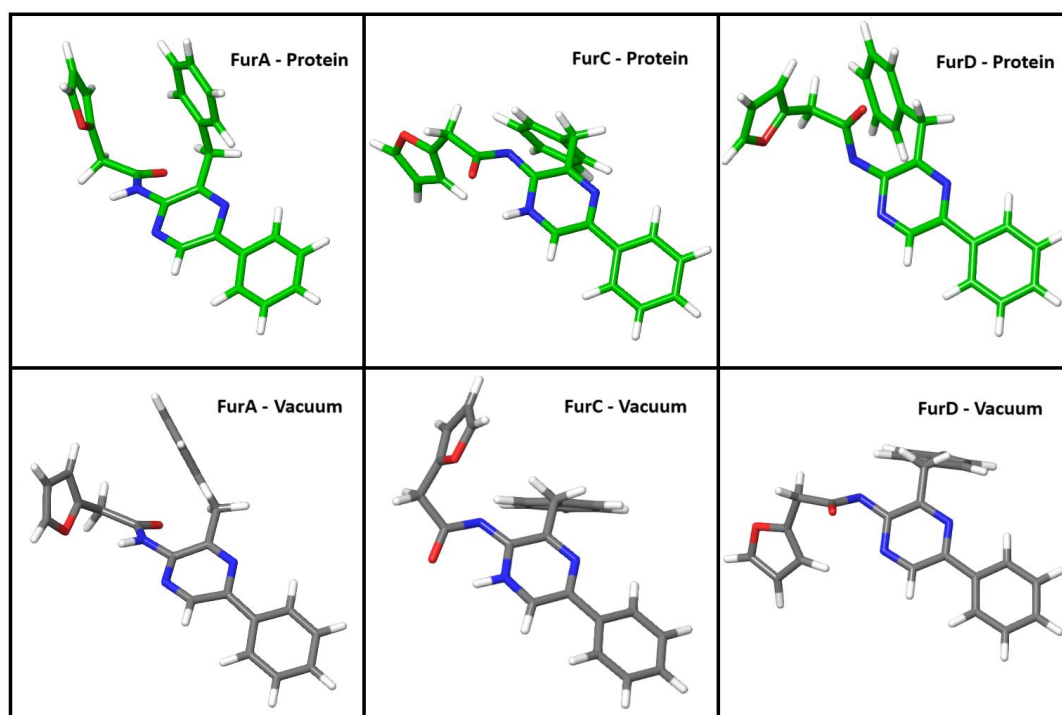


Figure 3. 23 - Comparison of the Fur* ES geometry in NanoLuc_E environment (green carbons) and *in vacuo* (dark grey carbons)

In FurA, the central ring planarizes going from the gas phase to enzyme environment, with the corresponding N7-C5-N4-C3 and N4-C5-N7-C6 (see Figure 3.12 for numbering) dihedral angles changing from 9° to -1° and from -167° to 174°, as shown in Table 3.13.

<i>Torsional angles of FurA optimized geometries</i>				
	1-2-3-4 (°)	7-6-9-10 (°)	8-9-11-12 (°)	9-11-12-13 (°)
Vacuum	6.0	-140.7	178.5	-72.1
Protein	12.9	-128.8	76.2	-69.7

Table 3. 13 - Comparison of torsional angles of furimamide neutral form excited state optimized geometries in vacuum and in NanoLuc_E protein environment

This conformation, less flexed, is similar to the ones of the other furimamide forms of in both vacuum and NanoLuc_E environment. The furan ring also undergoes structural variations, but they do not affect the emission spectrum, as showed by the emission energies computed as function of the furan torsion angle (Table 3.14).

<i>Scan of the FurA C9-C11-C12-O13 dihedral angle</i>	
α (°)	λ (nm)
-69.7	339
-59.7	339
-49.7	339
-39.7	339
-29.7	339
-19.7	339
-9.7	339
0.3	340
10.3	340
20.3	340
30.3	340
40.3	341

Table 3. 14 - Emission wavelength of FurA at different values of the C9-C11-C12-O13 dihedral angle. Scan performed on the optimized geometry obtained in NanoLuc_E protein environment. (Level of Theory: PBE0/6-31+G(d))

Thus, the blue shift which characterizes the emission spectrum obtained at the structure optimized in the protein environment is mainly due to the conformational changes specifically involving the phenyl moiety and the central ring. Since there was no internal coordinate which could be easily associated with the shift, the contribution of those groups was evaluated with single-point calculations on different FurA structures obtained by modifying the one optimized in NanoLuc_E environment. First, each of the two moieties was modified individually, imposing the corresponding in vacuo geometrical parameters values and maintaining the rest of the molecular structure unchanged; then, the combined effect of these same modifications was evaluated.

The results obtained are reported in Table 3.15 and highlight that both moieties are relevant in explaining the shifted emission.

<i>FurA predicted maximum emission depending on single group modification</i>					
	<i>In vacuo @ NanoLuc_E</i>	Core	Phenyl	Core+Phenyl	<i>In vacuo @ in vacuo</i>
λ (nm)	339	345	356	367	367
ΔE (eV)	3.53	3.58	3.48	3.38	3.38

Table 3. 15 - Predicted emission wavelength and corresponding ΔE of FurA NanoLuc_E optimized geometry in vacuum with moiety geometrical parameters modifications. Each column represents a specific modification to the geometry of the system. These modifications are based on the gas-phase optimized structure. “Core” is the pyrazinic ring, “Phenyl” is the ring directly linked to the central core. Values calculated in vacuo at the in vacuo ES optimized geometries reported for comparison. (Level of Theory: PBE0/6-31+G(d))

Concerning FurC, the most evident difference (see Figure 3.23 and Table 3.16) concerns the furan ring, which is almost parallel to the benzyl one when the molecule is in the protein environment, while it is more perpendicular to it in the vacuum geometry.

<i>Torsional angles of FurC optimized geometries</i>				
	1-2-3-4 (°)	7-6-9-10 (°)	8-9-11-12 (°)	9-11-12-13 (°)
Vacuum	-8.7	-14.9	-37.6	-66.6
Protein	18.0	29.2	79.1	175.8

Table 3. 16 - Comparison of torsional angles of furimamide zwitterionic form excited state optimized geometries in vacuum and in NanoLuc_E protein environment

In this case, the furan moiety is a stronger candidate to explain the shift, since MOs analysis in vacuum highlighted the main difference between FurC and the other furimamide forms: the LUMO is located only on the pyrazine ring, while the HOMO involves also the furan group, even though it does not appear to provide a major contribution. Maintaining fixed the structure optimized in the protein environment and modifying the geometrical arrangement of the furan group like in the geometry optimized in led to a strong reduction of the shift from a 0.13 eV blue-shift to a red-shift of just 0.03 eV (see Table 3.17), supporting the idea that furan position is crucial in modifying the emission spectrum.

<i>FurC predicted maximum emission depending on single group modification</i>			
	<i>In vacuo @ NanoLuc_E</i>	Furan	<i>In vacuo @ in vacuo</i>
λ (nm)	430	456	451
ΔE (eV)	2.88	2.72	2.75

Table 3. 17 - Comparison of predicted emission wavelength and corresponding ΔE of FurC (NanoLuc_E) optimized geometry in vacuum and after modification of furan moiety geometrical parameters based on the gas-phase optimized structure. Values calculated in vacuo at the in vacuo ES optimized geometries reported for comparison. (Level of Theory: PBE0/6-31+G(d))

To be sure that the main modulating factor had been correctly identified, a rigid scan on the amide group rotation was performed, since it displayed a more pronouncedly out of plane position with respect to the *in vacuo* optimized geometry.

The scan results, reported in Table 3.18, highlighted the reduced, although still present, effect played by the small conformational changes of this group.

Scan of the FurC N7-C6-N8-C9 dihedral angle	
α (°)	λ (nm)
-17.7	433
-12.7	427
-7.7	425
-2.7	425
2.3	427
7.3	428
12.3	429
17.3	431
22.3	433
27.3	434
32.3	436

Table 3. 18 - Emission wavelength of FurC at different values of the N7-C6-N8-C9 dihedral angle. Scan performed on the optimized geometry obtained in NanoLuc_E protein environment. (Level of Theory: PBE0/6-31+G(d))

Finally, FurD does not show any relevant shift in the emission spectrum (see Table 3.11), although being characterized by different geometries depending on the environment in which the structure is optimized, especially for what concerns the furan and the benzyl moiety positions, as clearly shown in Figure 3.24. This suggests that the conformational constraints imposed by the different protein environments do not play a significant role in the definition of the maximum emission wavelength.

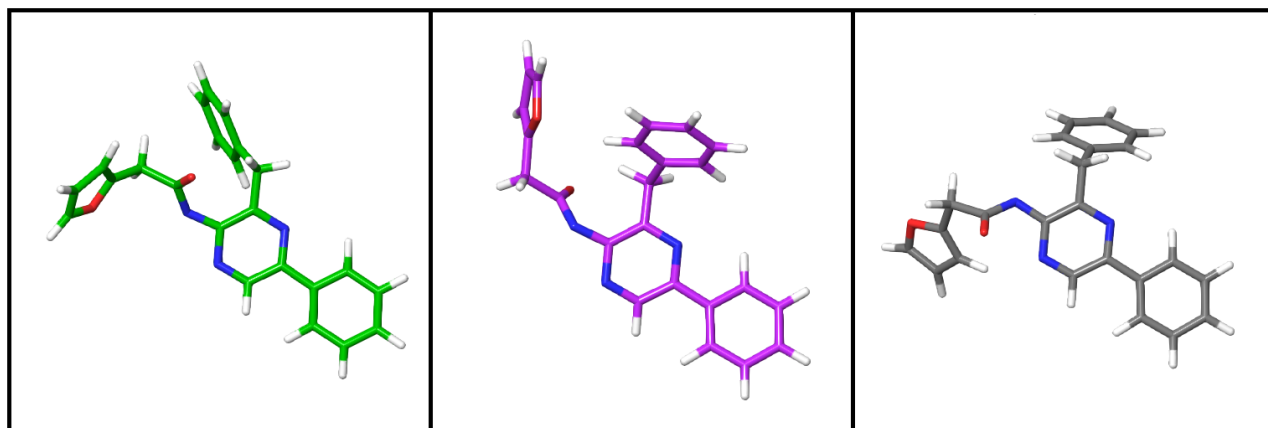


Figure 3. 24 - ES geometries of FurD optimized in NanoLuc_E (left), NanoLuc_K (middle) and vacuum (right)

Analysis of the molecular orbitals involved in the transition on the protein-ligand optimized geometries allowed the possibility to highlight the contribution of the moieties in the modulation of the emission. General results that hold across the small set reveal that the benzyl group is not included, neither in the HOMO, nor in the LUMO.

Given this, it is possible to hypothesize that the substitution of the benzyl with a group conjugated to the pyrazinic ring could possibly have a large effect on the emission wavelength. This has been recently demonstrated by Xiong et al.²⁴⁹ who designed a coelenterazine analogue without the two hydroxyl groups on the phenyl rings and a 4-quinolinyl substitution at the C8 position of the imidazopyrazinone core (Figure 3.25), obtaining a peak emission at 585 nm. However, substrate changes must be evaluated carefully as the emission is dependent on the access of the substrate to the active site, which is - in turn - associated with the size and flexibility of the moieties.

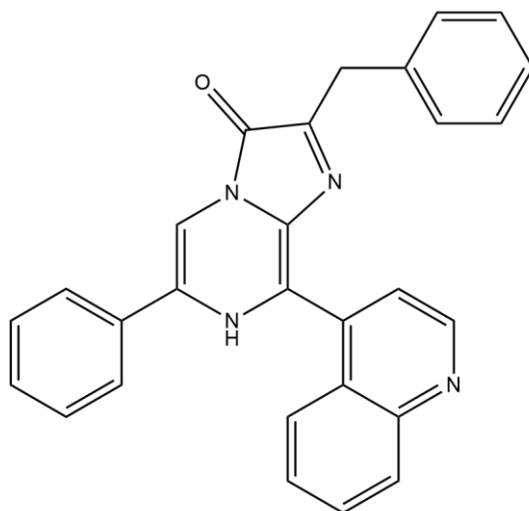


Figure 3. 25 - QTZ, the coelenterazine analogue characterized by a 4-quinolinyl substituent, designed by Xiong et al.²⁴⁹

3.1.6 Conclusions

This part of the thesis work was focused on two specific aspects of the mechanism behind the bioluminescence of the NanoLuc-furimazine system. The first concerns a possible keto-enol tautomerization of the substrate (furimazine) and the second one concerning the most probable protonation state of the emissive final product (furimamide).

Concerning the first point, results provide hints that indicate the enol form to be preferable: this tautomer appears as stable as the keto form in aqueous solution and the results from docking and the analysis of the MD interaction energy in the protein pocket appear to favour it. On top of that, it is a more plausible reagent candidate for the chemiluminescent reaction catalysed by NanoLuc. Indeed, the docked structures for the enol tautomer reveal the presence of a stable H-bond with a residue of the β -barrel (Asp139), that suggests a direct involvement of this interaction in the initial deprotonation leading to furimazine oxidation. The presence of this H-bond makes for a more straightforward initial step compared to the keto case and further supports the role of the enol form.

Concerning the second point, fluorescence emissions have been simulated for three furimamide forms (Fur A, C and D) corresponding to different protonation states of the product. QM/MM calculations have provided emission spectra indicating the zwitterionic form, FurC, to have the best adherence to the experimental results (λ_{max} of 413 nm and 460 nm, respectively). This is observed for the NanoLuc_E-FurC complex, the complex deriving from the docking of FurC into the NanoLuc pocket that hosted the furimazine enol form. This further supports the involvement of the furimazine enol form in the reaction mechanism.

Moreover, the absence of specific interactions involving FurC and protein residues in the NanoLuc_E-FurC complex is another argument in favour of its involvement as the final product since it would more easily leave the protein pocket, thus reducing the risk of product inhibition^{247,248}. The negatively charged form FurD shows a similar maximum emission wavelength in both NanoLuc_E and NanoLuc_K. However, the values for FurA and FurC pK_a^{*} have been previously estimated to be very high (16.9 and 11.3, respectively)¹⁸, indicating that the protonation of FurD is highly probable in the biological medium, so that also this species could be reasonably excluded from the candidates.

The situation is different for FurA, the neutral form of furimamide, whose emission is strongly blue-shifted with respect to the experimental data, no matter the kind of protein environment, so that this species can be easily discharged as probable luminescent product. This is further supported by the fact that FurA is also less stable in the excited state compared to the zwitterionic form FurC, hinting at the more favourable formation of the latter during the oxidative process.

Future developments, currently underway, are aiming at the unambiguous identification of the emitter through an in-depth reconstruction and modelling of the full catalytic mechanism, to finally understand the photoactivity of this important luciferase-luciferin pair and drive and inspire its further optimization.

3.2 Furimazine and furimamide characterization

3.2.1 Furimazine

All the results presented in Section 3.1 are based on the assumption that the enolic form is either dominant, or present in a sufficient amount, in the cellular environment, i.e. in aqueous cytosol, a polar solvent. Considering the limited water solubility of furimazine¹⁴⁵, and with a view to a possible future experimental study of the mechanism of furimazine oxidation in solution proposed in Section 3.3, it was decided to proceed with the study of the spectroscopic properties of furimazine in dimethyl sulphoxide (DMSO) and dichloromethane (DCM), in order to evaluate the keto-enolic equilibrium and provide further support for the previously described hypothesis. The choice of DMSO as the main solvent for the spectroscopic characterization was based on the fact that it is a polar solvent, like water, it is commonly used for experimental studies on imidazopyrazinone core systems²⁵⁰ and considering the assumption that the single electron transfer which is supposed to be the spark which ignites the reaction is strongly favoured in polar solvents⁴⁵. The experimental characterization of furimazine, whose spectra were compared to the corresponding computational results, was performed by professor Beverina and his collaborators at University of Milano-Bicocca.

A molecule undergoing tautomerization exists as a mixture of different tautomers, with their relative abundance being dependent on their stability in the selected solvent. Tautomerization leads to a reorganization of the electron density of the system, thus the two (or more) possible forms are different chemical species, therefore having different physical properties. A popular strategy to evaluate the tautomeric equilibrium constant and determine the abundance of each form in a specific solvent, is based on measuring physical properties of the solution and comparing them to the corresponding values for each pure tautomer, which are required to perform this kind of evaluation²⁵¹. One approach involves “blocked” tautomers, where the labile proton has been replaced by a methyl group to prevent tautomerization, while the main alternative not requiring it is the exploitation of spectroscopy combined with quantum chemical calculations. Since the tautomers are characterized by non-identical electron densities, they must be considered as distinct chemical species, thus meaning that their spectroscopic properties can be used as fingerprints for their identification, but this does not resolve the issue of requiring the knowledge of isolated tautomers properties and this is where QM calculations come into the picture. When predicting absorption or emission spectra with this kind of calculations, no tautomerization is taken into account, thus meaning that spectra for the two different forms can be easily predicted and compared with the experimental results. The spectral information commonly used for tautomerism studies are the UV/Vis absorption and emission spectra and the NMR ones. For the UV/Vis results, since the excitations are of the order of fs, the conformational effects of rotatable bonds with low energy barriers to rotation have a reduced effect, while for the NMR ones they become far more relevant, thus making it harder for those spectra to be reproduced.

3.2.1.1 UV/Vis spectroscopy

First, the UV/Vis absorption spectra of both furimazine keto and enol forms in DMSO were calculated. The obtained results are reported in Figure 3.26.

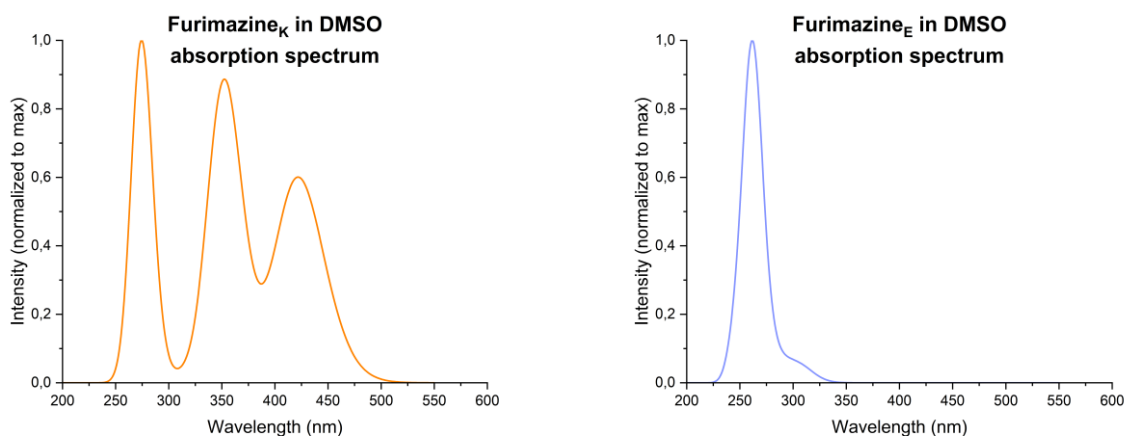


Figure 3. 26 - Absorption spectra of furimazine keto (left) and enol (right) forms normalized to their own maxima. Spectra obtained simulated by broadening the vertical excitation energies for the lowest ten singlet excited states with a gaussian function of HWHM = 0.19 eV. Level of theory: PBE0/6-31+G(d)//CAM-B3LYP/6-31+G(d)

The calculated spectra differ significantly, with the enol form seemingly absorbing only in the UV range, with a notable peak at 270 nm characterized by a shoulder around 300 nm. The keto form, instead, is the only absorbing species in the wavelength range between 400 and 500 nm. The great difference between the two species absorption spectra would have allowed to exclude the presence of the ketonic form, if it was not existent in DMSO, since the absence of absorbance in the aforementioned range would have clearly been associated with the exclusive existence of the enol form. However, the experimental UV/Vis spectrum reports absorption in that range and it is closely matching the keto form calculated spectrum, as shown in Figure 3.27.

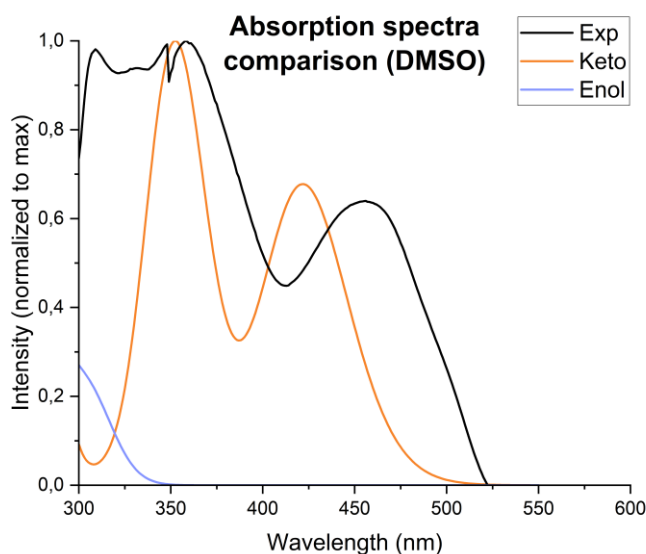


Figure 3. 27 - Experimental absorption spectra of furimazine in DMSO compared to the predicted ones for its enol and keto forms. Artifact at 350 nm caused by the switch from the deuterium arc lamp to the mercury arc one. Normalization of the experimental spectrum done with respect to the maximum peak at $\lambda > 320$ nm, normalization for enol and keto forms done with respect to the keto form maximum peak at 357 nm. Band broadening with gaussian function of HWHM = 0.19 eV for both calculated spectra. Level of theory: PBE0/6-31+G(d)//CAM-B3LYP/6-31+G(d)

The comparison between the experimental data and the calculated spectra highlights the certain existence of at least a fraction of the keto form of furimazine. Anyway, the enol form presence cannot be excluded or evaluated with certainty, since it does not absorb in the evaluated range of the electromagnetic spectrum, but the absorption spectrum cannot be collected at lower values of λ since the DMSO absorption may cause artifacts or forbid data collection due to its cutoff, which is at 270 nm.

To further confirm the quality of the protocol for the UV/Vis spectra prediction, the absorption spectra of both enol and keto forms of furimazine in DCM were calculated and compared with the experimental data, also collected by Beverina's group. The results, reported in Figure 3.28 are strikingly similar to the ones obtained in DMSO, except for the absorption around 300 nm which in this case is absent.

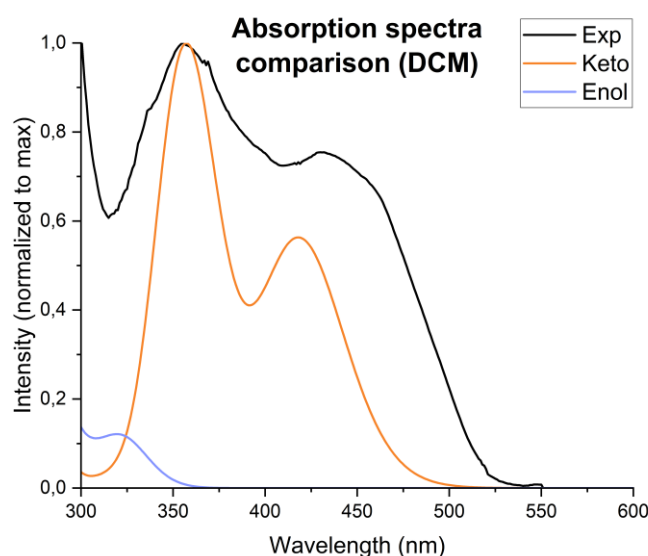


Figure 3. 28 - Experimental absorption spectra of furimazine in DCM compared to the predicted ones for its enol and keto forms. Normalization of the experimental spectrum done with respect to the maximum peak at $\lambda > 320$ nm, normalization for enol and keto forms done with respect to the keto form maximum peak at 357 nm. Band broadening with gaussian function of $HWHM = 0.19$ eV for both calculated spectra. Level of theory: PBE0/6-31+G(d)//CAM-B3LYP/6-31+G(d)

A possible explanation for the absence of this peak is related to the different solvent cutoff. Since DCM has a lower one (233 nm), artifacts around 300 nm are less probable, but it can also be related to the different quantitative of the enol form in the two solvents: in DMSO, which is a polar solvent able to form weak hydrogen bonds in presence of an hydrogen bond donor, the enol form could be favoured, while in DCM the keto form is more probably dominant. To further discriminate between the two forms in DMSO, NMR spectra were also collected.

3.2.1.2 NMR spectroscopy

The collected NMR data includes both the ^{13}C and the ^1H spectra, respectively reported in Figures 3.29 and 3.30, in order to get a more detailed picture of the tautomeric equilibrium between the keto and the enol form in DMSO and be able to verify the hypothesis that the enol form could exist in a polar solvent.

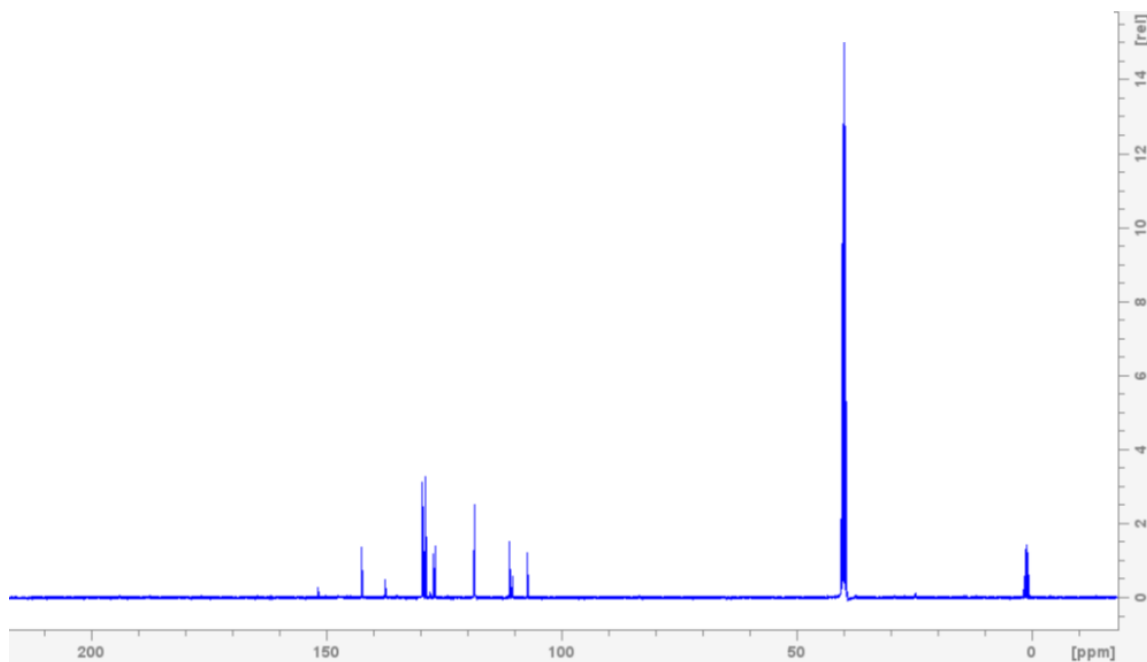


Figure 3. 29 - ^{13}C NMR spectrum of furimazine in DMSO

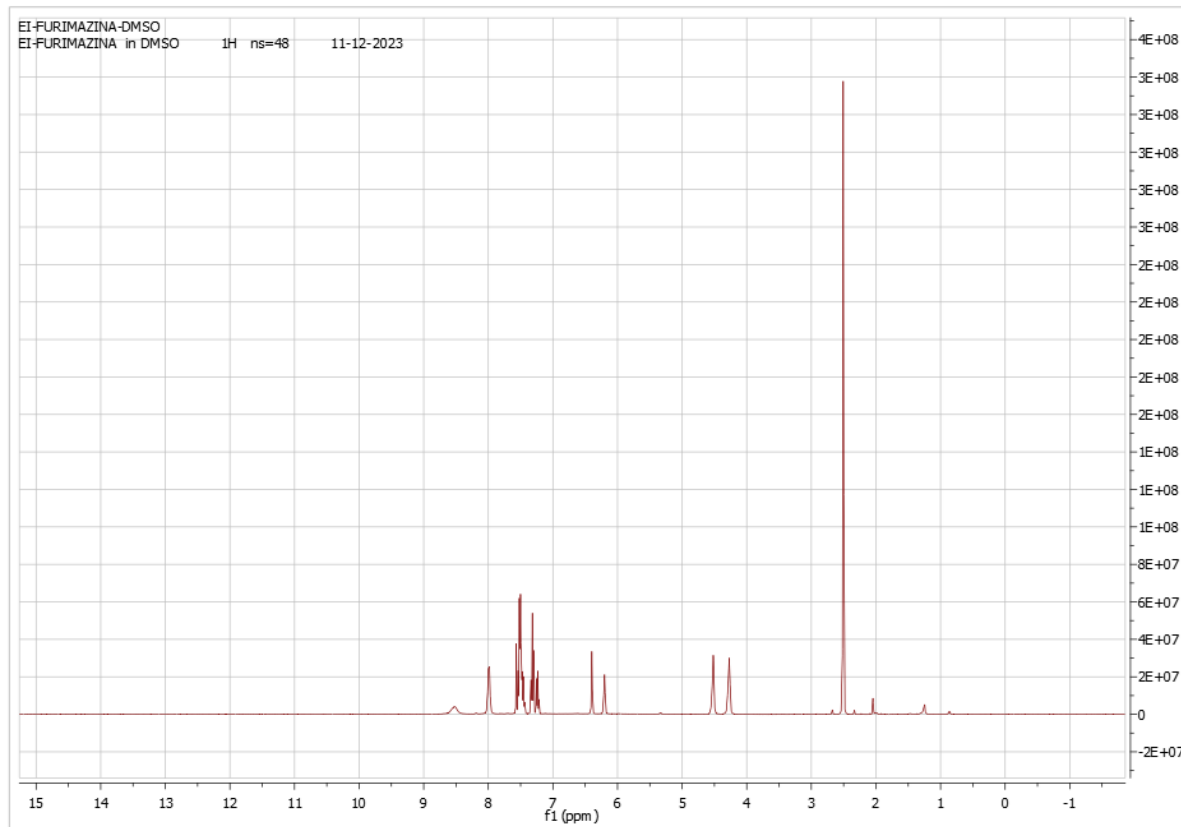


Figure 3. 30 - ^1H NMR spectrum of furimazine in DMSO

A first glance at the ^1H spectrum makes it clear that there is a dominant form in the solution, since there is a single set of peaks. However, the profile does not correspond to the expected one for the keto form of furimazine, which is the only form confirmed to be present in the DMSO solution, according to the UV/Vis absorption spectrum. The computed spectra for this species, reported in Figure 3.32, indeed, exhibit peak locations that differ from the experimental data, and it becomes even more evident if compared to the assignation of the experimental signals carried out by Beverina's group (Figure 3.31).

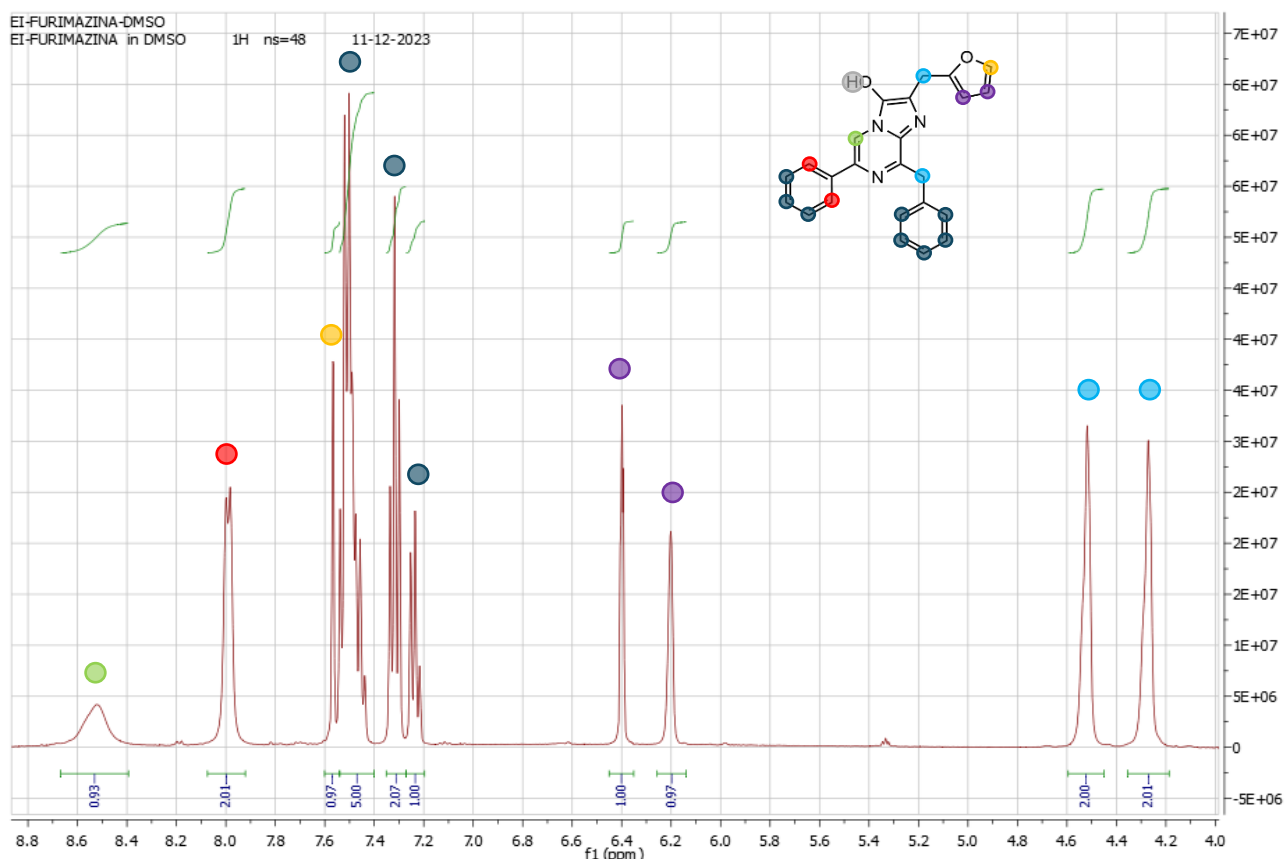


Figure 3. 31 - Zoom on the region of interest of the NMR ^1H spectrum of furimazine in DMSO. Assignment of peaks to corresponding protons highlighted with different coloured circles, tautomeric proton reported in grey, but absent in the spectrum due to exchange with residual D_2O molecules in solution. The result of the peaks integration is reported under the spectrum profile.

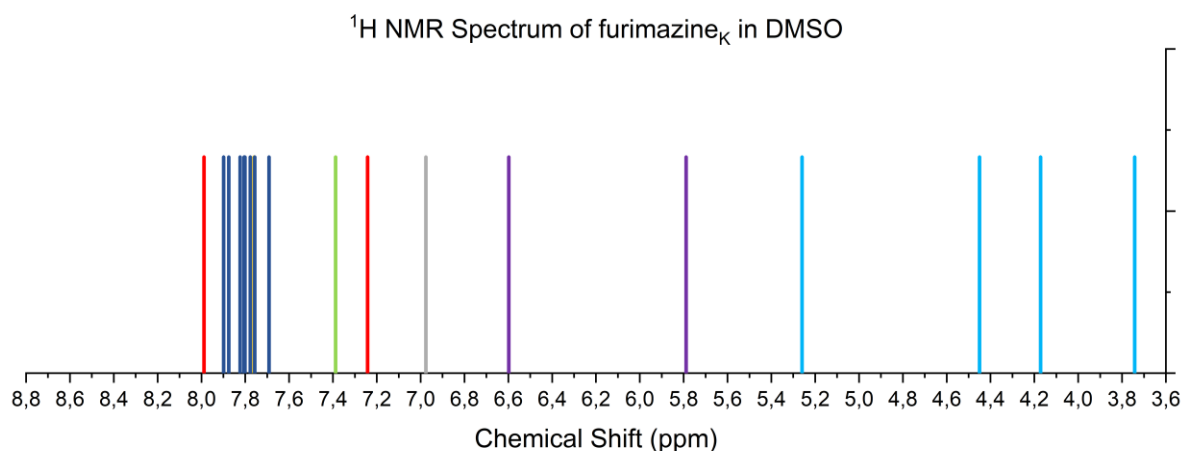


Figure 3. 32 - Computed ^1H NMR spectrum for keto form of furimazine. Each proton corresponds to a different line, coloured according to the colour-code employed for the experimental peaks assignation.

Comparing the calculated spectra for the keto form of furimazine and the experimental one, it is important to remember that the computed signals are not taking into account conformational effects, which can almost be neglected in the UV/Vis spectra due to the quickness of the analysis, but are far more relevant in the case of NMR results since data collection takes longer. This is the reason why the 4 sky blue protons signals, for example, do not coalesce in two peaks, but are more spread over the spectrum. Anyway, some similarities between the two spectral profiles are evident, such as the fact that those 4 protons are the ones characterized by the lowest chemical shifts, followed by the two “purple” ones of the furan ring, and the presence of the group of signals associated with the aromatic protons around 7.5-8.0 ppm, which include also the furan ring proton reported in yellow. On the other hand, some differences stand out, in particular the red protons seem to be less shifted than expected and, more interestingly, the green proton signal, the only one of the central ring apart from the tautomeric one, is far less shielded in the calculated spectrum, leaving no calculated signal with a chemical shift higher than 8.0 ppm. As regards the tautomeric ^1H , reported in grey, it is not present in the experimental spectrum due to proton exchange with D_2O molecules still present in the solvent, thus meaning that it cannot be used as signature to clearly identify the dominant tautomer. A different picture comes out if the enolic form is taken into account as the most relevant one, since the computed spectrum, reported in Figure 3.33 seems to be more fitting with the experimental one.

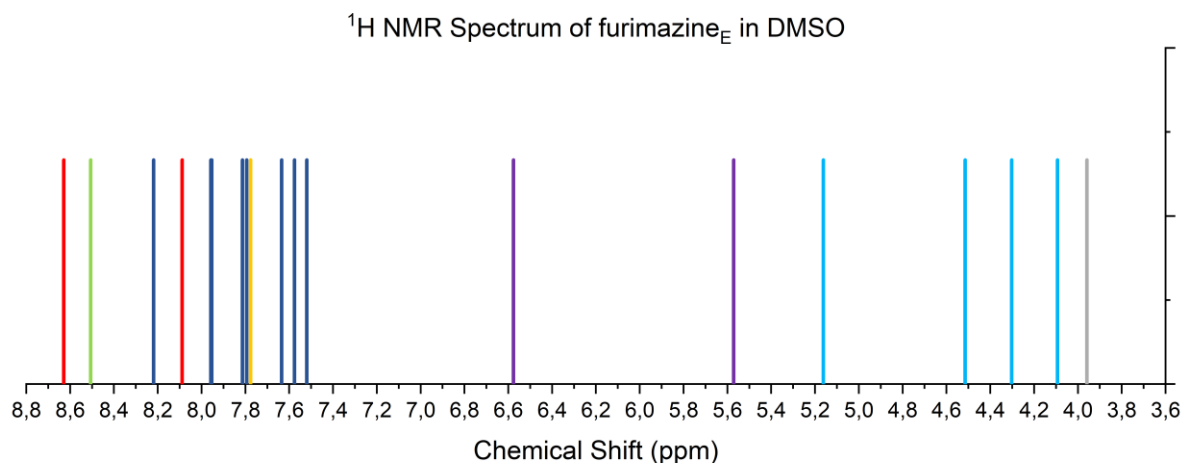


Figure 3. 33 - Computed ^1H NMR spectrum for enol form of furimazine. Each proton corresponds to a different line, coloured according to the colour-code employed for the experimental peaks assignment.

The enol form ^1H spectrum, indeed, shows the same correct collocation of the sky blue and purple protons, and a more spread distribution of the aromatic signals, but the red protons are more shifted, with an average of 8.3-8.4 ppm and, more importantly, the core proton, i.e. the “green” one, is at 8.5 ppm which is exactly the same shift observed in the experimental results. The hypothesis that the NMR proton spectrum is hinting at a dominance of the enolic form in the polar solvent DMSO gains even more credibility considering the results obtained by Nakai and his group in 2003²⁵². In the paper, reporting NMR spectra of different imidazopyrazinone core derivatives, including their alkylated derivatives, only the forced to be enol forms displayed a signal with a comparable chemical shift (> 8.0 ppm), while the keto ones did not display any.

Similar conclusions can be obtained by focusing on the ^{13}C results (see Figure 3.34).

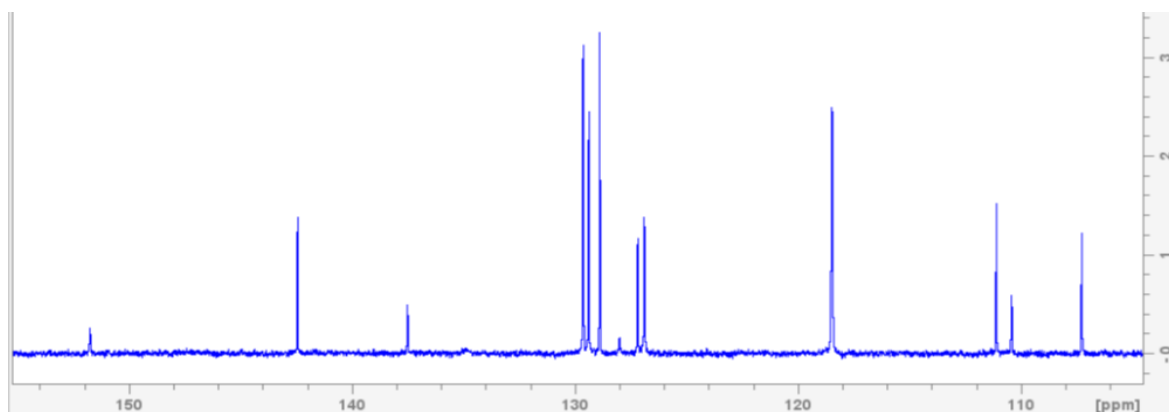


Figure 3. 34 - Zoom on the region of interest of the NMR ^{13}C spectrum of furimazine in DMSO.

The experimental spectrum does not show any strong signal with a chemical shift higher than 143 ppm, while in the keto forms studied by Nakai there is always one or even two, and a signal at 137, where in the literature there is none between 130 and 140. On the other hand, according to Nakai et al.'s paper the forced to be enol forms display a signal at 141-142 ppm and one at 136 ppm, both perfectly fitting with the two relevant peaks present in the collected data.

A strongly discordant note is the computed relative stability of the two forms, since in DMSO the keto form is supposed to be dominant, with an electronic energy difference in its favour of ~ 5 kcal/mol. However, performed calculations did not take into account explicit solvation and possible hydrogen bonds. This is why the GS optimization of geometry was repeated for the system, this time including an explicit DMSO molecule, located in front of the tautomeric proton in both structures. Still, the keto form remained favoured, but with a gap reduction to ~ 1.8 kcal/mol. The fact that the proton of interest was not visible in the ^1H NMR, which is the main result in support of the enolic presence in a polar solvent, due to exchange with remaining water molecules in the solvent led us to consider the hypothesis of a water microsolvation that could alter the equilibrium between the two forms. This is why the previously optimized structures of furimazine microsolvated in water (see Figure 3.2) were reoptimized, this time including DMSO as the implicit solvent. These calculations revealed a reversal in stability with the enol form being slightly favoured according to the electronic energy values (~ -0.1 kcal/mol). Since the basis set employed for those calculations is not including polarization functions on hydrogens, which are supposed to be used whenever dealing with hydrogen bonding, the results for both the furimazine-water cluster in implicit water and in implicit DMSO were recomputed with the 6-31+G(d,p) basis set, leading to a great change in the electronic energy difference, which became strongly favouring the enol form (-1.8 kcal/mol in DMSO and -1.67 kcal/mol in water), as expected. All results are reported in Table 3.19.

<i>Electronic energy difference</i>			
Basis set	Explicit solvent molecule(s)	Implicit solvation	ΔE keto-enol (kcal/mol)
6-31+G(d)	None	DMSO	5.0
6-31+G(d)	DMSO	DMSO	1.8
6-31+G(d)	H ₂ O	DMSO	-0.1
6-31+G(d,p)	H ₂ O	DMSO	-1.8
6-31+G(d,p)	H ₂ O	H ₂ O	-1.67

Table 3. 19 - Relative stability of keto and enol forms of furimazine with respect to explicit and implicit solvation and basis set employed for calculations

3.2.1 Furimamide

Another crucial aspect of the results reported in Section 3.1 is the identification of the emitting form of furimamide involved in the bioluminescent process. The emission spectra analysis for the keto and enol derived NanoLuc-furimamide complexes led to the reduction of the pool of candidates to just two different forms of the product, FurC and FurD, but the latter was predicted to be less likely to exist due to the strong tendency to acquire a proton and convert into either FurA or FurC. To verify this hypothesis, the neutral form of furimamide was synthesized by Beverina's group and UV/Vis characterization in acetonitrile (ACN) was performed.

First, the absorption and emission spectra were collected and compared to the predicted ones for FurC and FurA. The result of the comparison is reported in Figure 3.35.

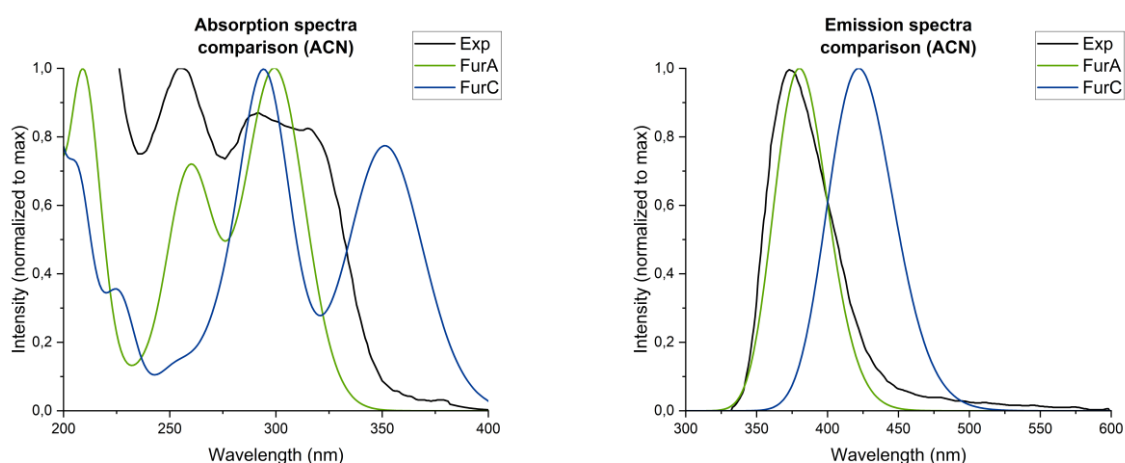


Figure 3. 35 - Experimental absorption and emission spectra of furimamide in ACN compared to the predicted ones for FurA and FurC. Normalization of spectra done with respect to each own maxima at $\lambda > 250$ nm. Band broadening with gaussian function of $HWHM = 0.19$ eV for both calculated spectra. Level of theory: PBE0/6-31+G(d)//CAM-B3LYP/6-31+G(d)

The interpretation of the results is easier than it was for furimazine enol and keto forms, since the experimental absorption spectrum is matching with FurA calculated one and the neutral form of furimamide has a lower computed GS electronic energy than its zwitterionic counterpart ($\Delta E = 7.1$ kcal/mol). The emission spectrum fits even better with FurA predicted emission, despite the predicted inversion of stability between this form and FurC in the ES ($\Delta E = -4.1$ kcal/mol). This is probably due to the fact that, after the excitation, the radiative relaxation process is too fast to allow the tautomerization from FurA to FurC.

Once the furimamide form in solution had been identified, the absorption spectrum collection was repeated after the addition of sodium ethoxide, to check whether the deprotonation of furimamide was easily achievable or not and the results are reported in Figure 3.36.

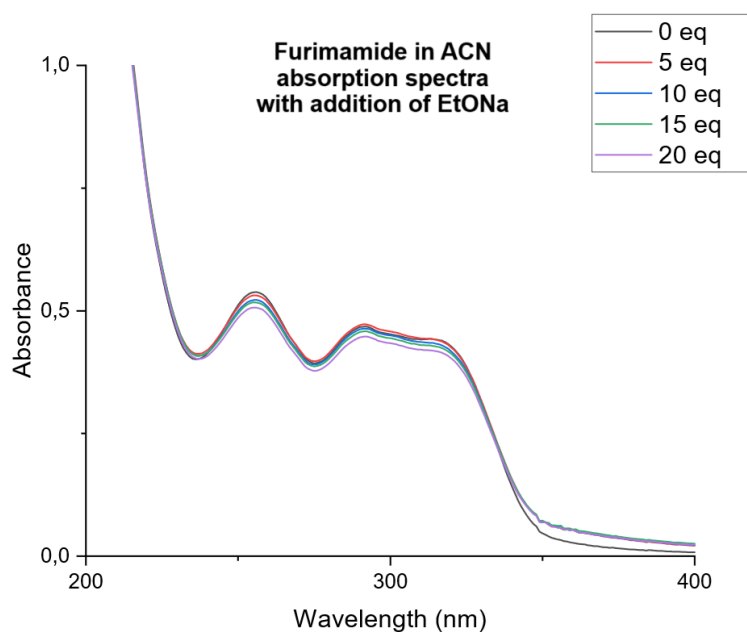


Figure 3. 36 - Experimental absorption spectra of furimazine in ACN after addition of different amounts of EtONa solution. Details about solutions molarity and employed spectrometer reported in Appendix 2, section A2.2

As expected, the deprotonation of furimamide is everything but easy to achieve, thus meaning that either the protein environment plays a crucial role in the stabilization of the anionic form of furimamide, or the emitting species of the NanoLuc-furimazine bioluminescent pair is not negatively charged, which seems to be the most probable option.

3.2.3 Conclusions

Section 3.2 was focused on the computational and experimental characterization of furimazine and furimamide, the two main chemical species in the bioluminescence mechanism of NanoLuc. First of all, the keto-enol tautomerism of furimazine was studied in polar solvents like DMSO. UV/Vis spectroscopic data confirmed the existence of at least a fraction of the keto form, while NMR results suggest that one of the two forms is strongly dominant on the other and that the prevailing form could be the enol one. The dominance of the enol form and a reduced amount of keto tautomer would easily account for the apparently contrasting pictures given by the UV/Vis and NMR spectra, since the NMR displays a single set of peaks, pointing at a strongly prevalent form, i.e. the enol one, while the UV/Vis absorption spectrum could be determined by a small amount of keto form as the counterpart is transparent in the studied wavelength range. Electronic energy calculations, especially those that accounted for the microsolvation effects, supported this hypothesis, with the enol tautomer being energetically favourable when explicit solvent interactions were included and basis set extended with the addition of polarization functions on H atoms. These findings are therefore indicative of coexistence of both tautomers, with the solvent conditions playing an important role in shifting this equilibrium.

The analysis reported in section 3.1 narrowed down the emission candidates to two forms of furimamide, the zwitterionic FurC and the anionic FurD, the latter being predicted to be a strong base, easily reprotonated whenever formed and thus being a less probable final product of the bioluminescent process. To assess the viability of FurD formation, the UV/Vis characterization of a neutral form of furimamide was performed and its deprotonation attempted with the addition of a strong base like EtONa. The experimental absorption and emission spectra are matching with the predicted data for FurA, results that are consistent with the electronic energy difference between the GS of FurA and FurC, favouring the neutral form ($\Delta E_{\text{GS}} = 7.1$ kcal/mol), and the rapidity of the radiative decay of furimamide excited state, not allowing the tautomerization to give the more stable zwitterionic form after excitation ($\Delta E_{\text{ES}} = -4.1$ kcal/mol). The addition of EtONa proved the scarce tendency to deprotonation of neutral furimamide, with an absorption spectrum that remained unaffected even by the addition of 25 eq of base. These findings suggest that either the protein environment provides great stabilization to the anionic form or, most likely, the anionic form cannot be considered a candidate for the final emission of the NanoLuc-furimazine pair.

3.3 Modeling of furimazine oxidation mechanism in DMSO

Once the reactants and the product of the catalytic cycle have been identified, the natural prosecution of an enzymology study aiming at optimizing the enzyme properties via rational design is the evaluation of the catalytic mechanism, which requires a deep knowledge of the different reaction steps, in order to model it in the most suitable way. This section aims at gaining this kind of insight into the furimazine oxidation mechanism, trying to identify the intermediates of this process and finding out whether there are steps with an activation energy barrier so high that an enzymatic intervention is required to lower it to a more reasonable value. Each step of the reaction has been furtherly studied with single point calculations performed on the geometries obtained from the IRC results.

3.3.1 Imidazopyrazinone core oxygenation

Furimazine oxidation does not involve any additional cofactor or reactant other than oxygen, thus meaning that the model system can be easily designed and includes only the substrate itself and an O₂ molecule. The reaction is supposed to start after the deprotonation of the substrate, which was hypothesized to be one of the main roles played by NanoLuc residues in the previous sections of this thesis work. The first step after the anion generation is the oxygen attack to the central ring of the molecule and two possible pathways have been proposed for other ImPy core species: one involving a direct attack of O₂, even if it is in a triplet GS, the other considering a radical based pathway with a SET taking place before the oxygen attack. Previous studies have proven that the latter is the most probable option, from both a computational²⁵³ and experimental point of view⁵⁶ and this is why this kind of mechanism was proposed for this study. The entire reaction mechanism was modeled in DMSO since a polar solvent is required to make the SET an exergonic process²⁵³ and DMSO is a common solvent used in experimental studies of ImPys bioluminescence²⁵⁰. Not only the oxidation process can be initiated through a SET or not, but the oxygen attack can also occur at either the C2 or C3 atom of furimazine, giving rise to distinct pathways for dioxetanone formation. Numbering of atoms for the following section of the thesis is reported in Figure 3.37.

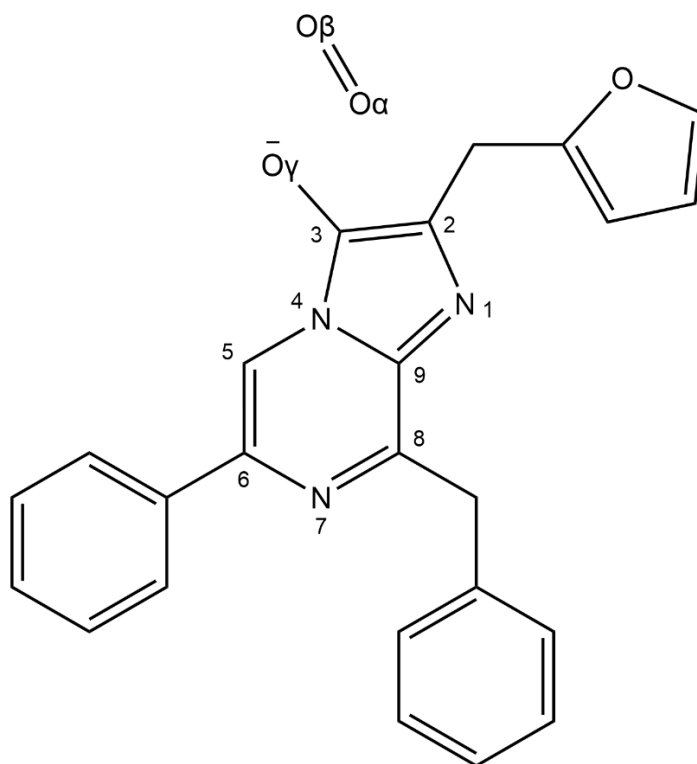


Figure 3. 37 - Atom numbering for anionic furimazine-O₂ exciplex

3.3.1.1 C2 oxygenation

The oxygen attack on C2 is the mainly investigated one due to the detection of the 2-peroxide intermediate in previous experimental investigations²⁵⁴. This pathway starts with the formation of the C2-O α bond, followed by the rotation of the O β atom towards the centre of the imidazopyrazinone and the formation of the four members ring with O β binding to C3. The last step is the cleavage of N4-C3 bond, leading to the formation of the dioxetanone (DE) intermediate. The free energy profile for the reaction with intermediates and transition states representations is reported in Figure 3.38.

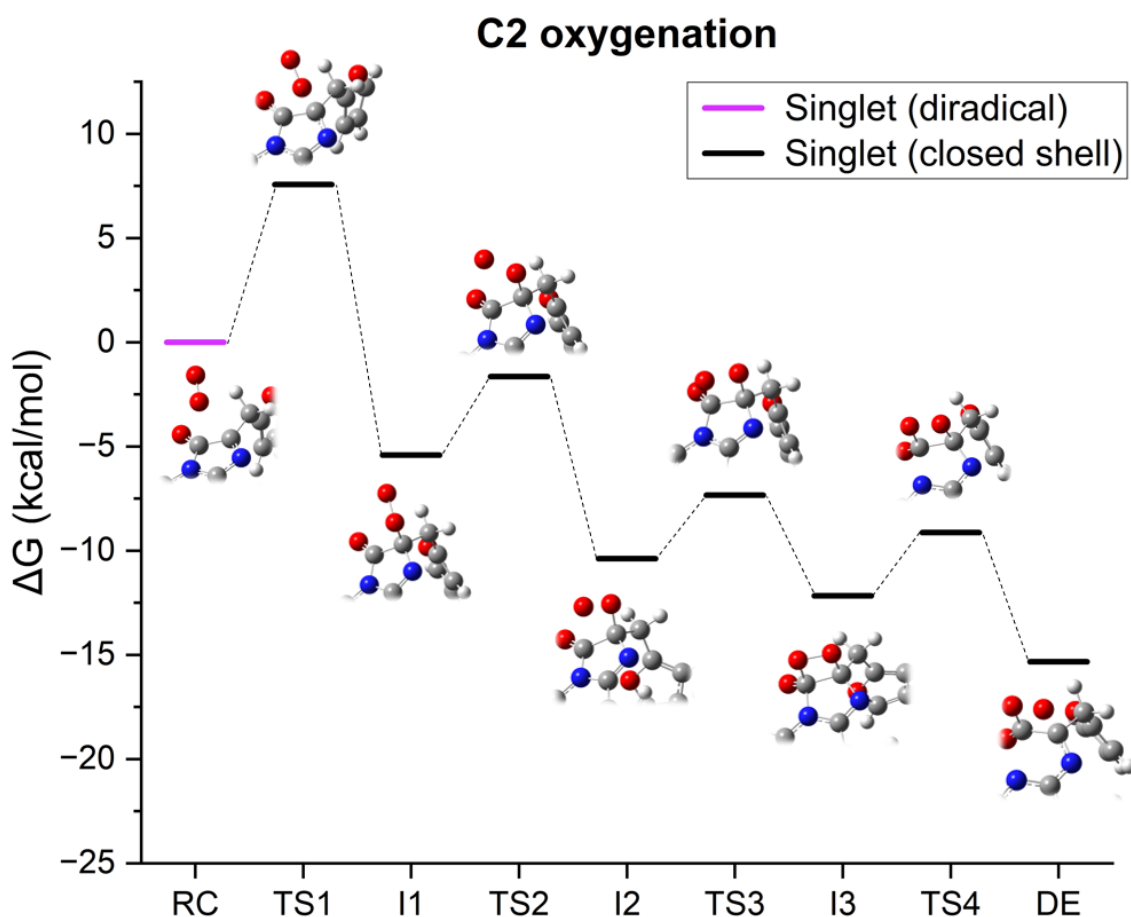


Figure 3. 38 - Free energy profile of the C2 oxygenation process. Energy of the optimized geometry for the diradical exciplex used as reference. Results at CAM-B3LYP/6-31+G(d,p) level of theory, implicit solvation in DMSO with CPCM model

After the formation of the singlet exciplex involving oxygen and the deprotonated furimazine, the oxygenation process begins with the shortening of the C2-O α distance, which leads to the formation of this bond during the first step of the process. This happens to be accompanied by the elongation of the C2-C3 bond and especially the O α -O β one, as expected. The shortening of the distance leads to the first transition state geometry, at which the energy difference between the singlet diradical state and the closed shell singlet one becomes zero. Moreover, this transition state coincides with the crossing of the singlet and triplet potential energy surfaces, with the triplet state becoming far less favourable, a feature which will be observed for the rest of the process, leading to the development of the reaction only on the singlet GS potential energy surface. After the formation of the anion peroxide intermediate, the electron density on the O₂ atoms increases, as expected and demonstrated by the Mulliken charges calculation. Trends of energies, distances and charges on O₂ atoms are reported in Figure 3.39.

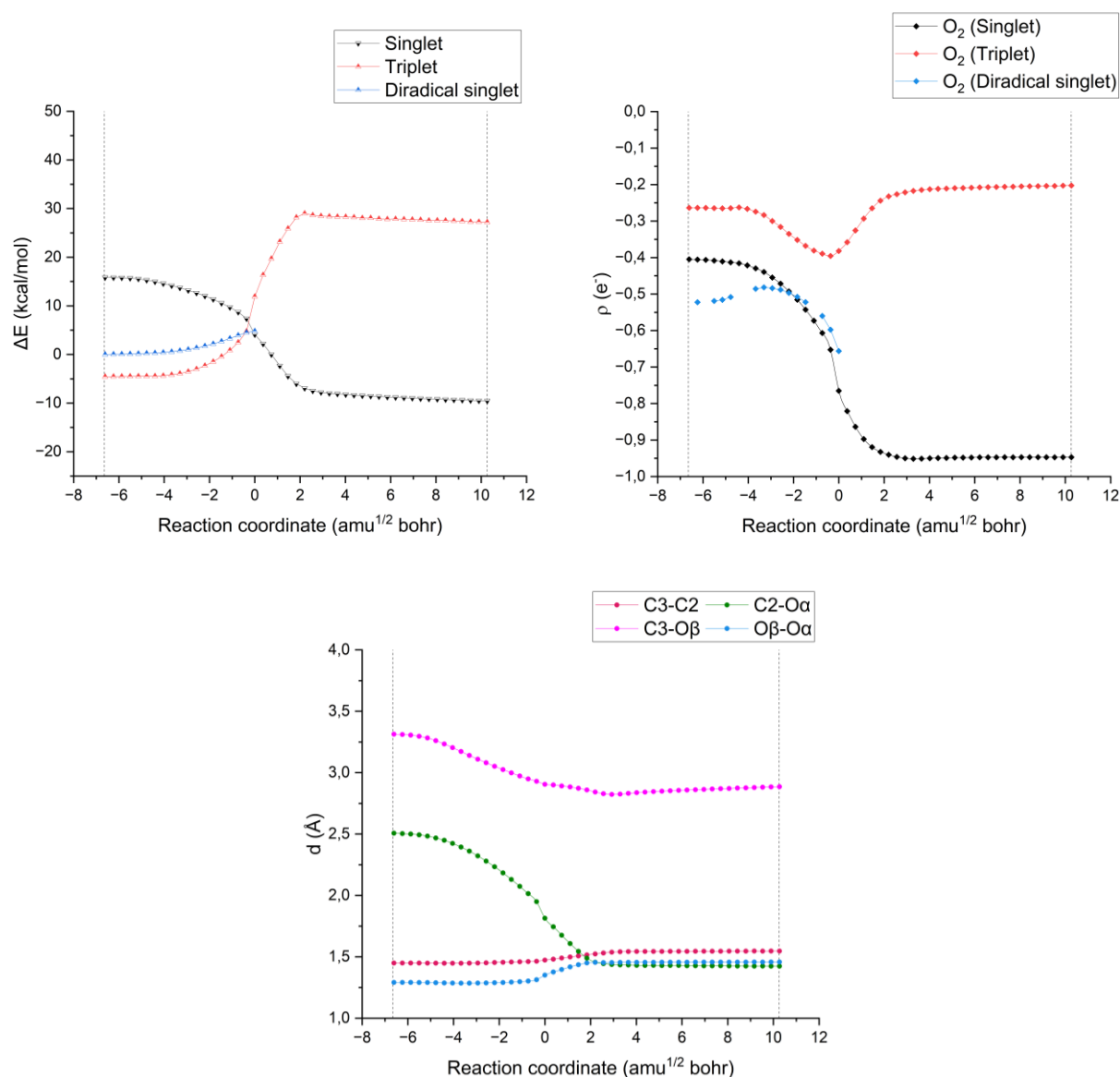


Figure 3. 39 - Trends for energy (a - in singlet, triplet and diradical singlet states), electron density on the O₂ atoms (b - measured in terms of Mulliken charges sum) and distances of interest (c) for the first step of C2 oxygenation. Each point corresponds to a different geometry given by the IRC calculation. Missing points for charges in diradical state due to convergence failure of the single point calculations. Dashed lines indicate the conclusion of the IRC calculation.

Once the peroxide has been formed, the closure of the four members ring, with the formation of C3-O β bond, happens in two steps, the first of them being the rotation of the terminal oxygen towards the

centre of the ring. The energy barrier for this rotation is very low, less than 3 kcal/mol, and the consequences of this step are mainly the reduction of the C3-O β distance as a result of the O β -O α -C2-C3 dihedral modification and a small reduction of the C2-O α bond length as shown in Figure 3.40.

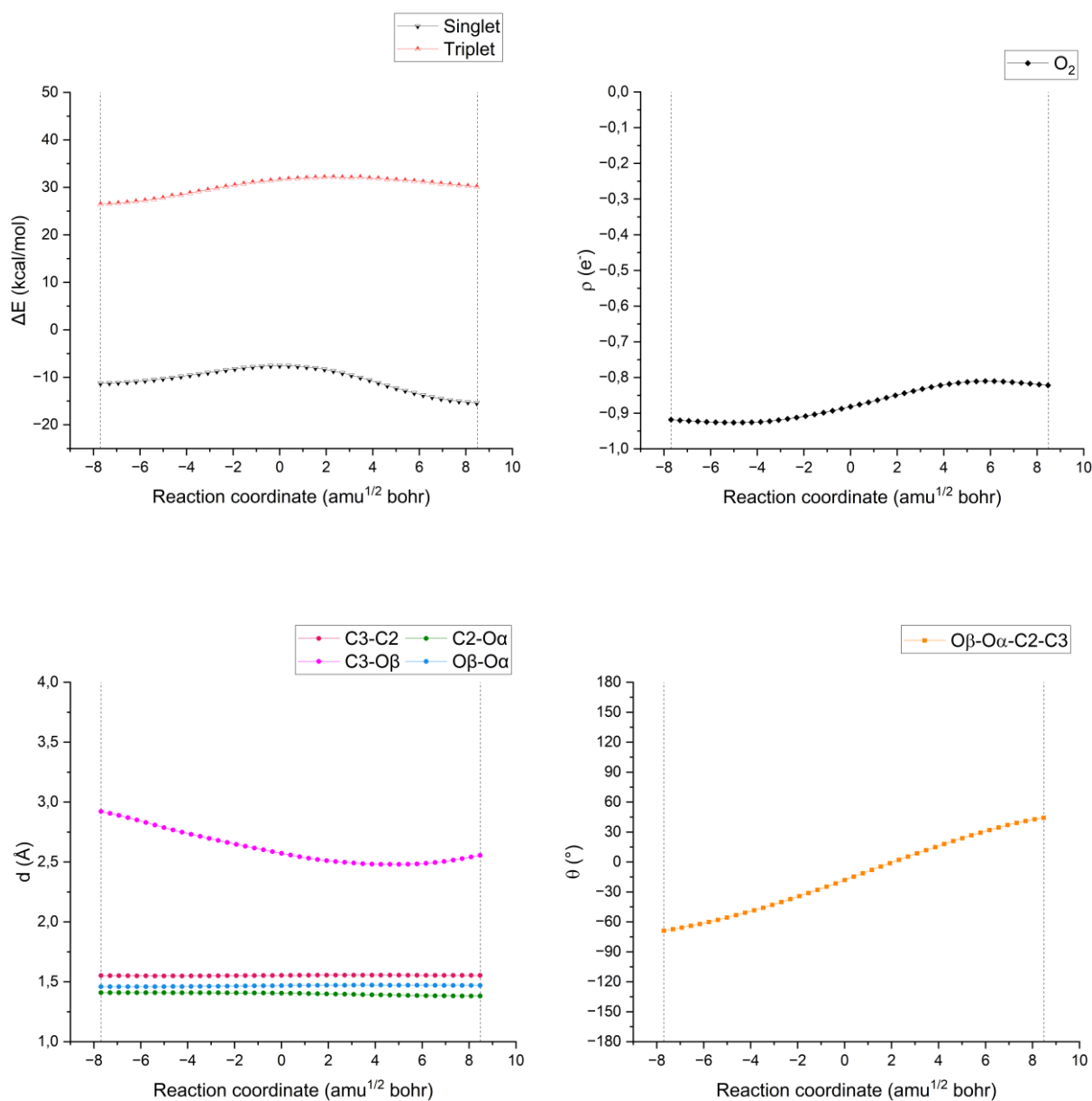


Figure 3. 40 - Trends for energy (a - in singlet and triplet states), electron density on the O₂ atoms (b - measured in terms of Mulliken charges sum), distances of interest (c) and dihedrals of interest (d) for the second step of C2 oxygenation. Each point corresponds to a different geometry given by the IRC calculation. Dashed lines indicate the conclusion of the IRC calculation.

The O β -C3 bond is then formed as it can be easily seen from the distances trends reported in Figure 3.41c, combined with the strong reduction in the dihedral angle O β -O α -C2-C3 (Figure 3.41d). This step leads to a redistribution of the electron density, as certified by the relevant change in O2 atoms charge (Figure 3.41b), lengthening in N4-C3 distance and decrease in the C6-C5-N4-O γ dihedral, which are the prelude to the following step, the opening of the five members ring resulting in the formation of the negatively charged dioxetanone intermediate (DE).

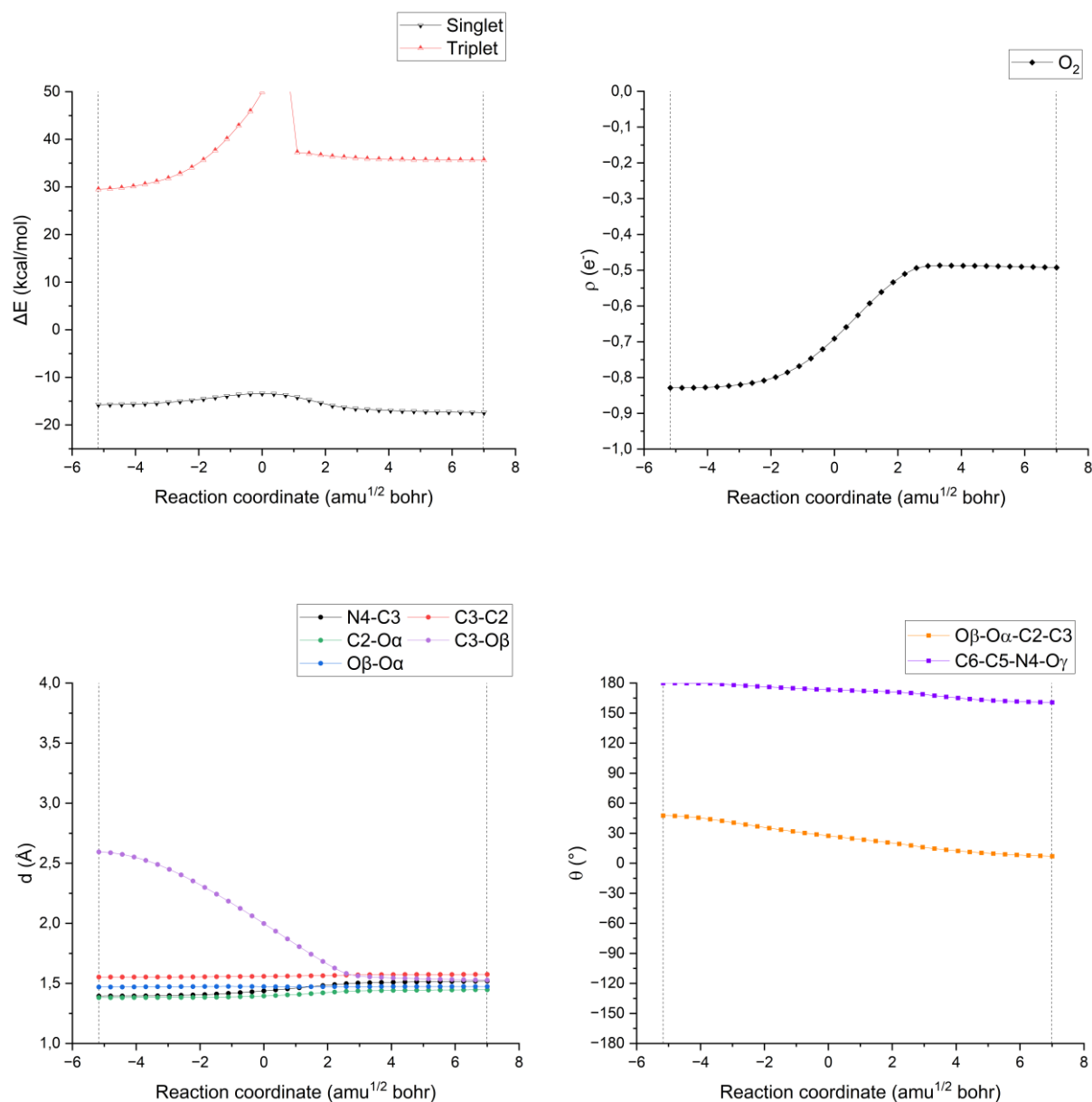


Figure 3. 41 - Trends for energy (a - in singlet and triplet states), electron density on the O2 atoms (b - measured in terms of Mulliken charges sum), distances of interest (c) and dihedrals of interest (d) for the third step of C2 oxygenation. Each point corresponds to a different geometry given by the IRC calculation. Dashed lines indicate the conclusion of the IRC calculation.

The final step of the C2 oxygenation pathway is the formation of the dioxetanone, the only other species which has been experimentally detected in similar systems²⁵⁴. The geometrical parameters whose change are more evident during this final step are the N4-C3 distance, obviously, but also the C3-O β one, which is reduced too (Figure 3.42c). To a smaller extent, also the O2 charge trend (Figure 3.42b) and the C6-N5-N4-O γ dihedral (Figure 3.42d) tell the same tale, since part of the electron density localized on O α and O β is redistributed over the molecule and the O γ atom definitively loses its planarity with the central core.

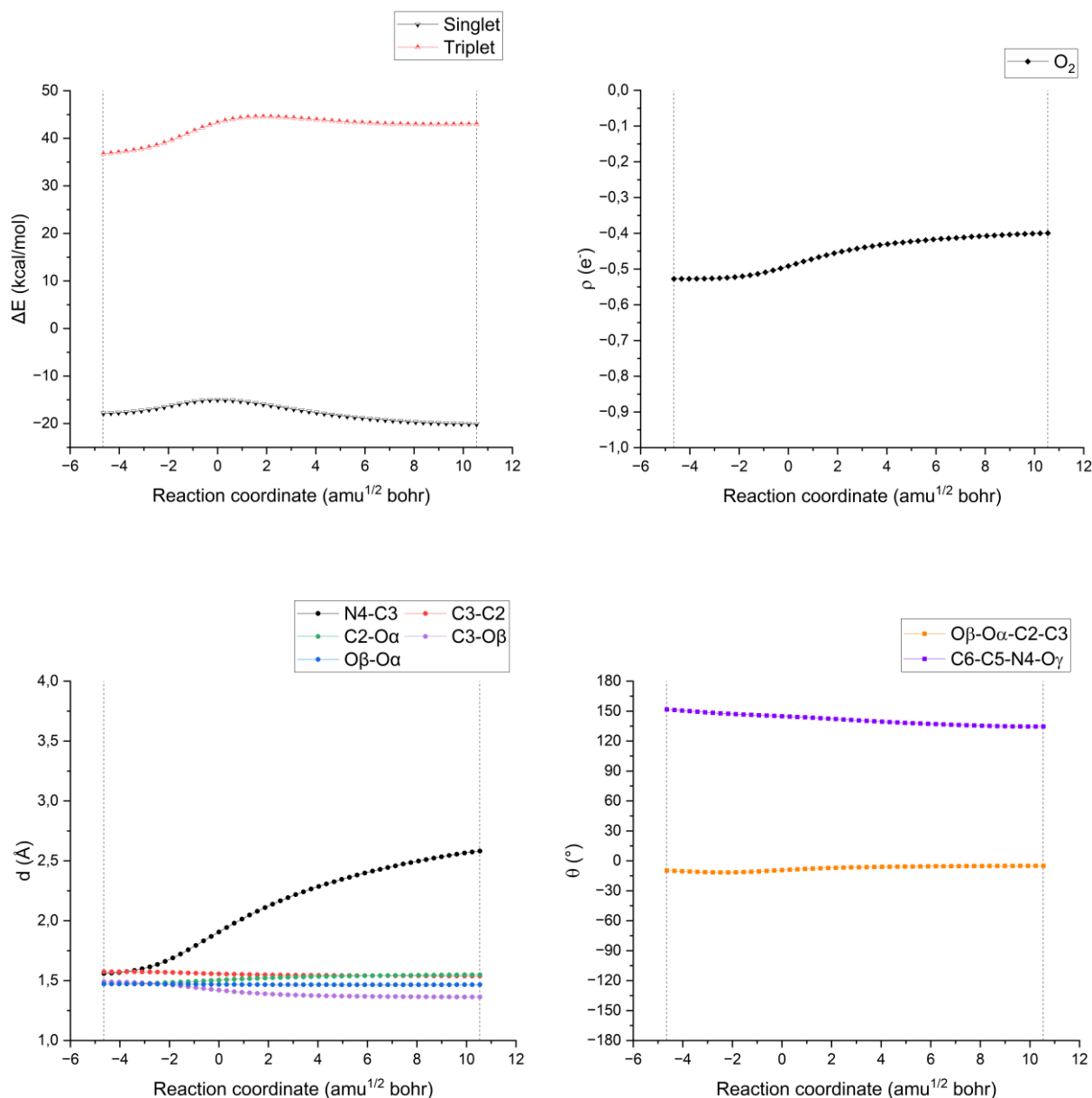


Figure 3. 42 - Trends for energy (a - in singlet and triplet states), electron density on the O₂ atoms (b - measured in terms of Mulliken charges sum), distances of interest (c) and dihedrals of interest (d) for the fourth step of C2 oxygenation. Each point corresponds to a different geometry given by the IRC calculation. Dashed lines indicate the conclusion of the IRC calculation.

Figure 3.43 shows the global trends of the different values of interest over the entire reaction path, highlighting some interesting features which seem to be characterizing the furimazine C2 oxygenation process.

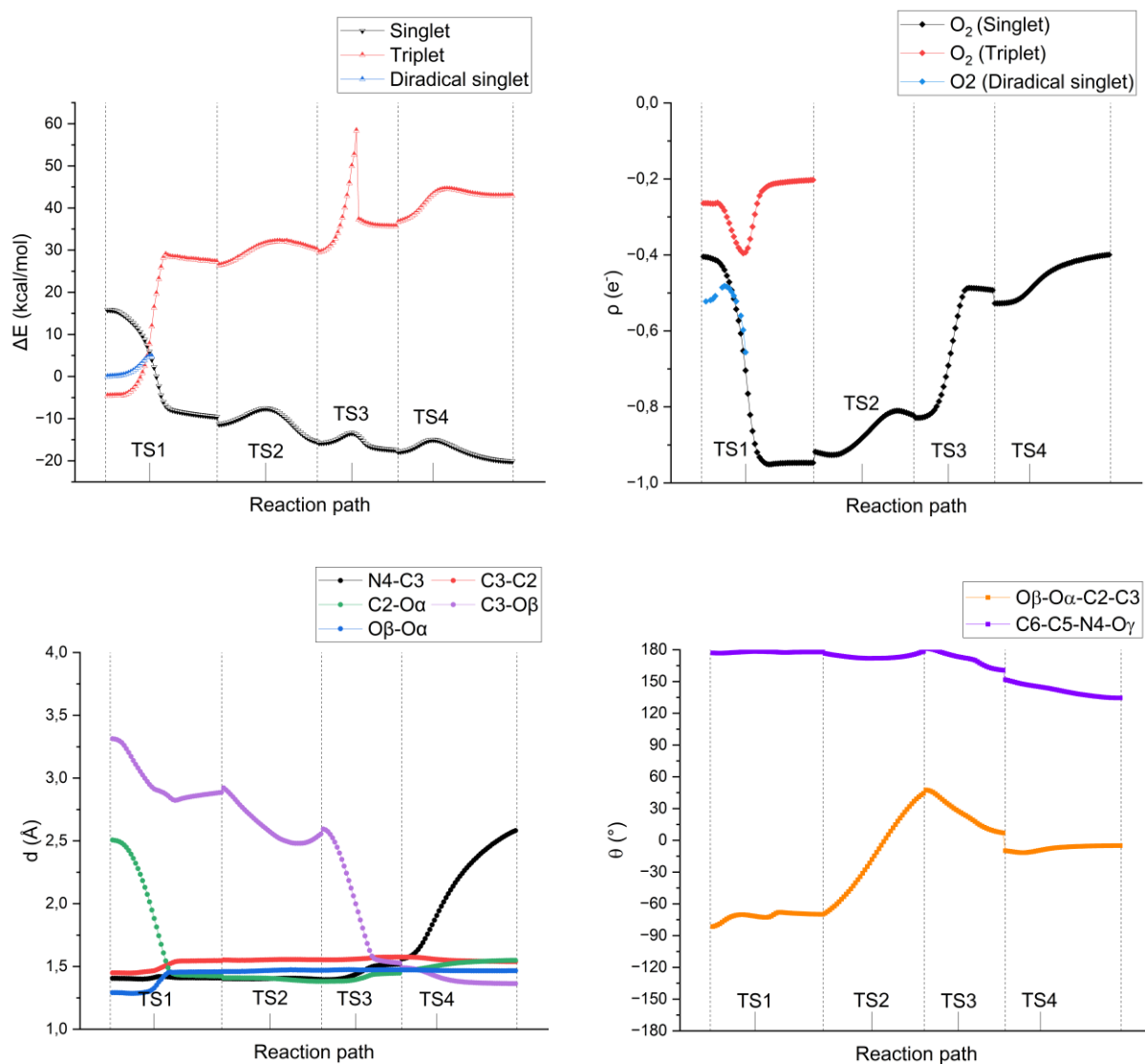


Figure 3. 43 - Trends for energy (a - in singlet, triplet and diradical singlet states), electron density on the O_2 atoms (b - measured in terms of Mulliken charges sum), distances of interest (c) and dihedrals of interest (d) for the C2 oxygenation. Each point corresponds to a different geometry given by the IRC calculations. Dashed lines indicate the conclusion of the IRC calculations.

The process is entirely developing on the closed shell singlet PES ever since the first transition state, which leads to the formation of the peroxide, which sees a charge transfer to the oxygen atoms followed by a gradual redistribution of the electron density in the following steps. The distances trends, together with the dihedrals ones, clearly show that the five-members ring opening, resulting in DE formation, takes place only after the closure of the four atoms one.

3.3.1.2 C3 oxygenation

Since the reaction has not been previously investigated in furimazine, the C3 oxygenation path (whose free energy profile is shown in Figure 3.44) could not be excluded as it could be a more straightforward option for the dioxetanone formation. In this case the first bond to be formed between the deprotonated core of furimazine and the molecular oxygen is not the C2-O α one, but the C3-O β counterpart, which obviously leads to a different reaction mechanism due to the impossibility for C3 to form another bond without breaking one. Thus, the C3-O β bond happens to form simultaneously to the elongation not only of the C2-C3 and O α -O β ones, as expected and already seen in the C2 mechanism, but also of the N4-C3 one, which is progressively broken to give a new intermediate. This intermediate is involved in a dihedral rotation with an energy barrier and gradient of the PES so small that the IRC calculation was not able to follow it, succeeded by the four members ring closure which forms the DE intermediate.

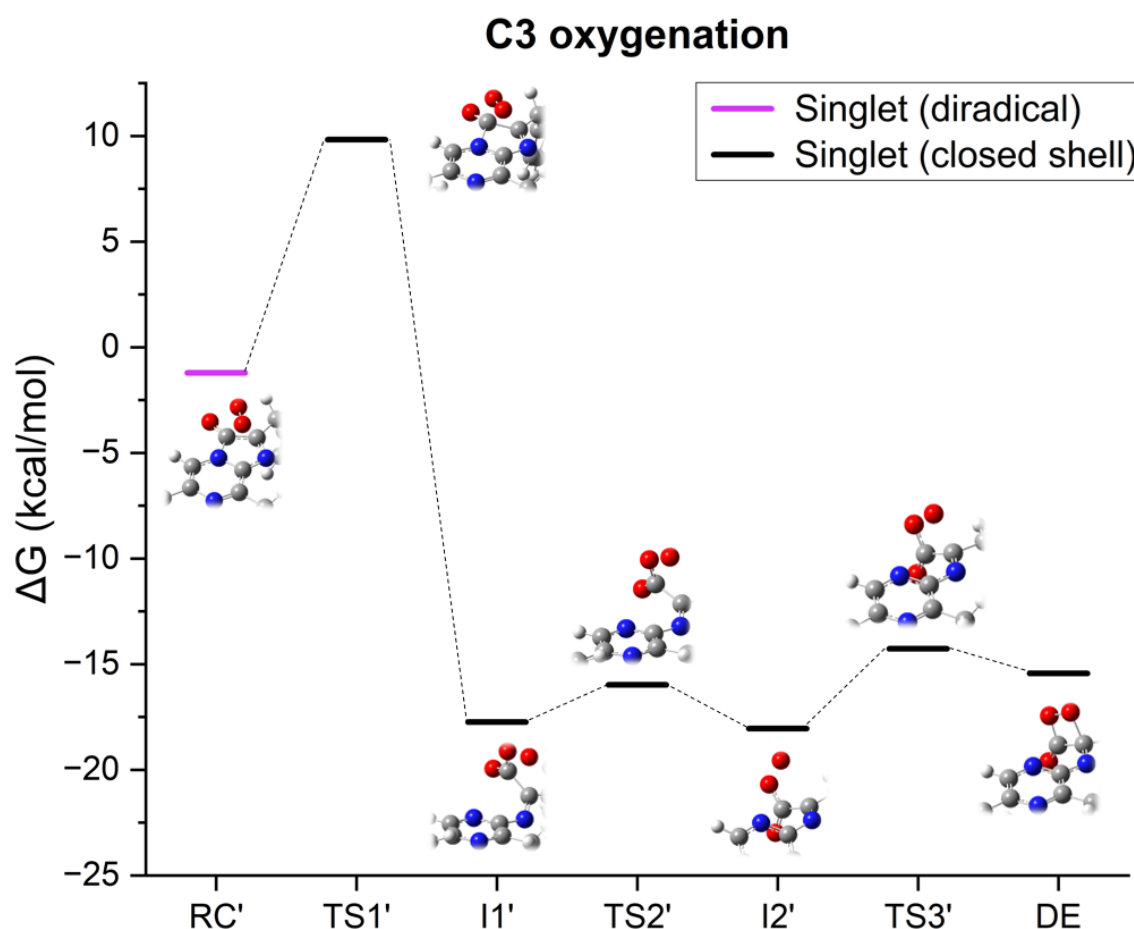


Figure 3. 44 - Free energy profile of the C3 oxygenation process. Free energy of the optimized geometry for the diradical exciplex in C2 mechanism used as reference. Results at CAM-B3LYP/6-31+G(d,p) level of theory, implicit solvation in DMSO with CPCM model

The first transition state of the C3 pathway is preceded by the crossing of the singlet, triplet and diradical singlet potential energy surfaces, with the closed shell singlet state becoming the most favourable close to the transition state. After the formation of the anion peroxide intermediate, which is formed even in this case, the electron density on the O₂ atoms increases, as expected and demonstrated by the Mulliken charges calculation. Trends of energies, distances and charges on O₂ atoms are reported in Figure 3.45. In this case the transition state is already located on the singlet potential energy curve, since the crossing between the diradical and closed shell singlet curves takes place before the transition state, as confirmed by a scan on the distance C3-O β which shows how its reduction is associated to a reduction in the energy gap between the two states (Figure 3.46). The presence of another oxygen

on the carbon bearing the peroxide anion leads to a reduction of the charge localized on O α and O β , as evidenced by the O2 charges trend (Figure 3.45b), which highlights the charge transfer and back charge transfer process taking place during the first step of the reaction. The key feature of the C3 process is the simultaneous formation of the C3-O β bond and cleavage of the N4-C3 one, clearly visible in Figure 3.45c, which leads to a remarkable structural reorganization of the molecule, confirmed by the dihedrals trends (Figure 3.45d).

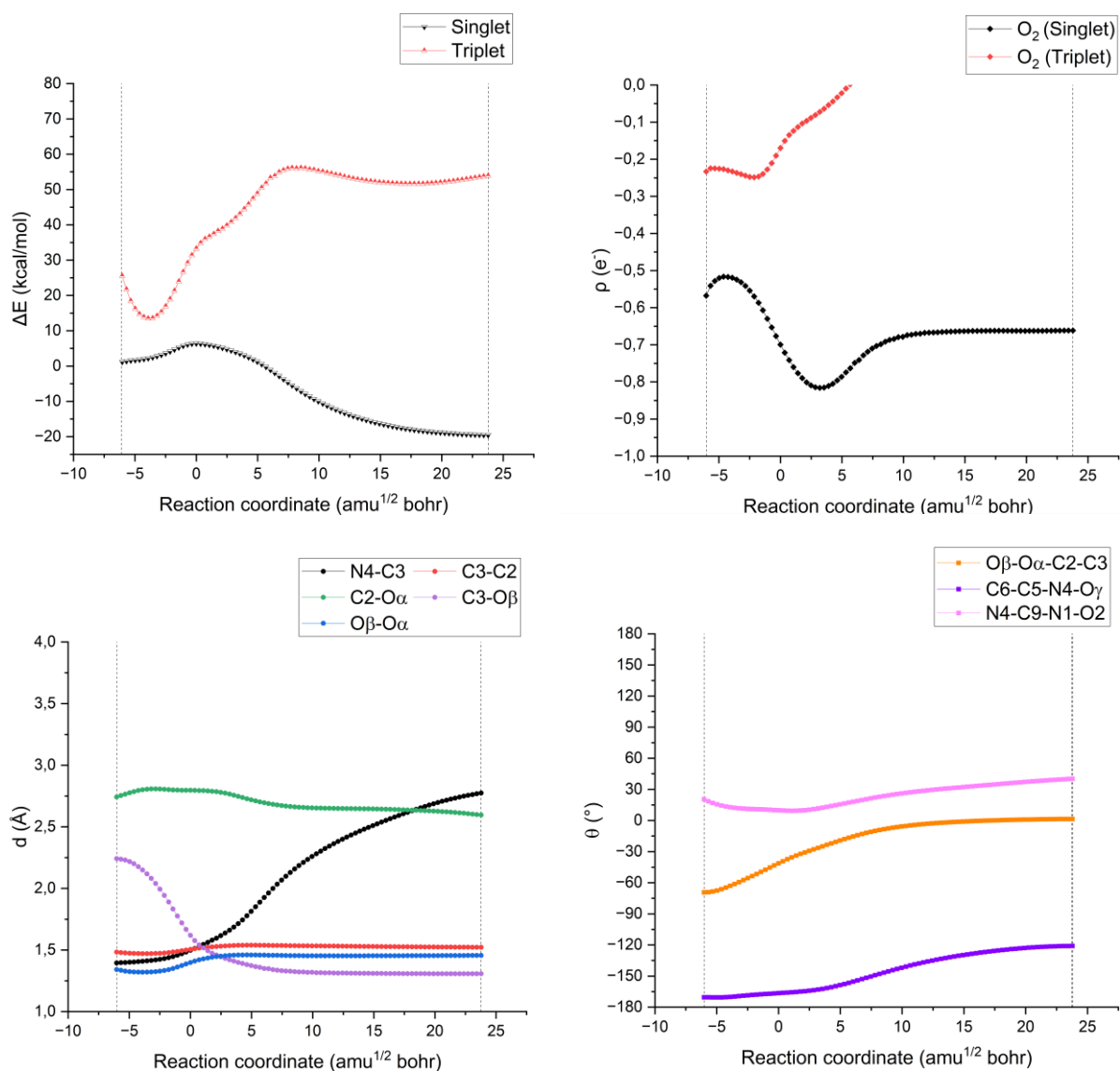


Figure 3. 45 - Trends for energy (a - in singlet and triplet states), electron density on the O2 atoms (b - measured in terms of Mulliken charges sum), distances of interest (c) and dihedrals of interest (d) for the first step of the C3 oxygenation. Each point corresponds to a different geometry given by the IRC calculation. Dashed lines indicate the conclusion of the IRC calculation.

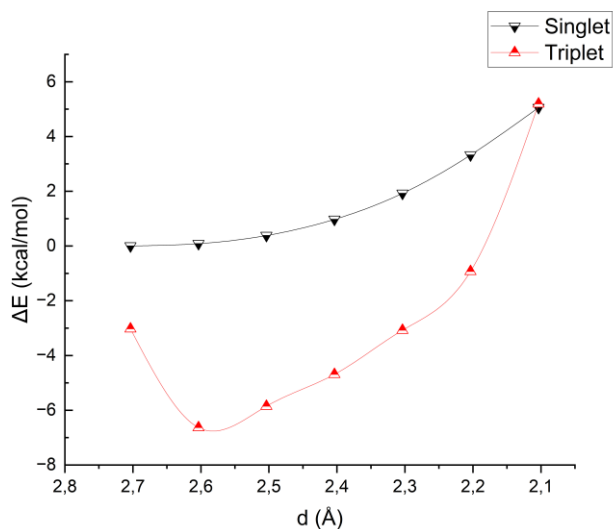


Figure 3. 46 - Trend for triplet and singlet energy as a function of C3-O β distance. The energy gap gets lower the closer the two atoms are.

Once the peroxide is formed and stabilized by a rearrangement of the molecule, the closure of the four members ring takes place, leading to the formation of DE, the same crucial intermediate which was obtained in the C2 process. As it was for C2, even in C3 oxygenation this step takes place with a negative charge withdrawal from the oxygen atoms (Figure 3.47b) and a redistribution of the electron density, consequence of the bond formation.

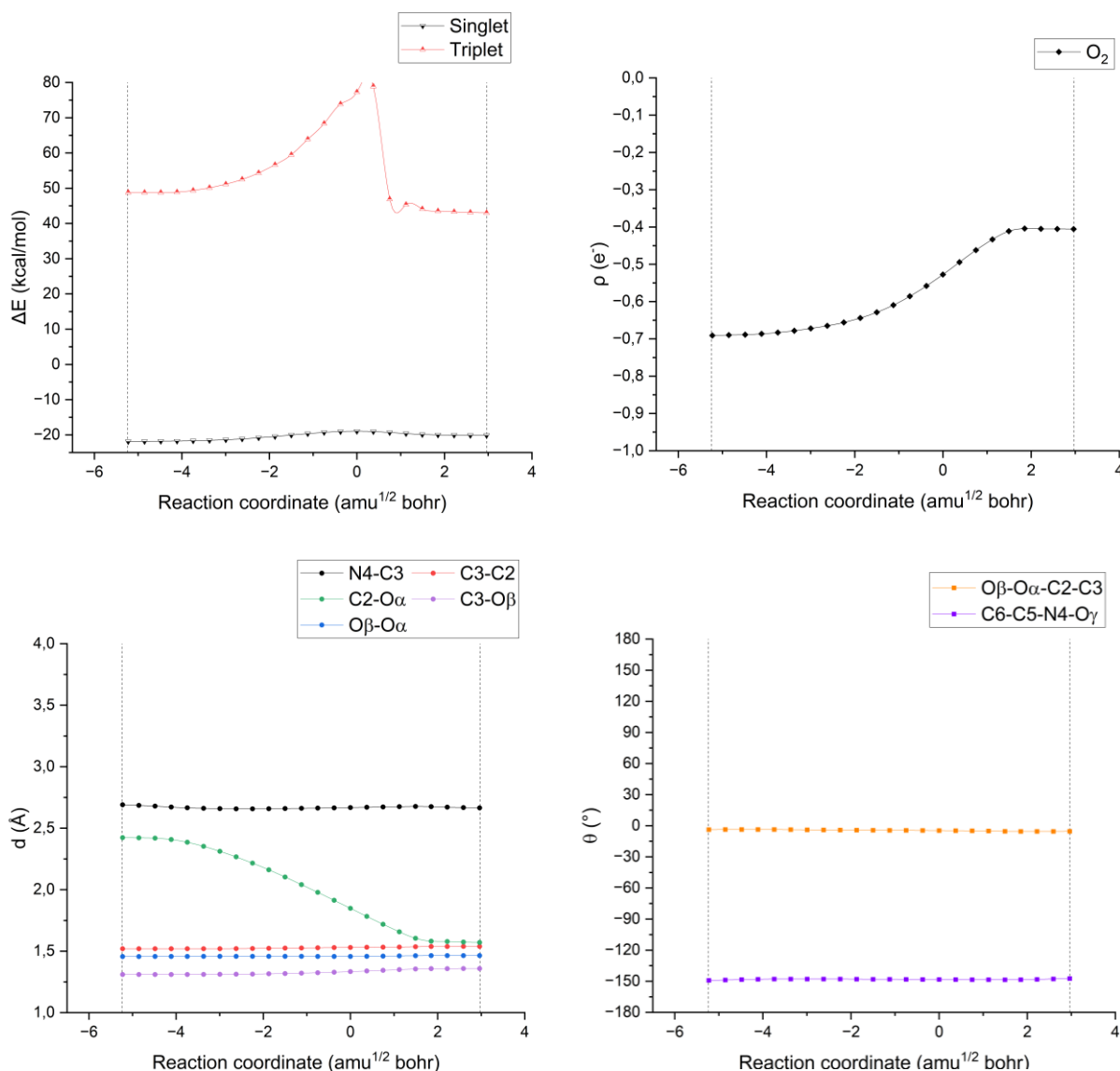


Figure 3. 47 - Trends for energy (a - in singlet and triplet states), electron density on the O₂ atoms (b - measured in terms of Mulliken charges sum), distances of interest (c) and dihedrals of interest (d) for the third step of the C3 oxygenation. Each point corresponds to a different geometry given by the IRC calculation. Dashed lines indicate the conclusion of the IRC calculation.

Ring closure does not involve any other moiety or particular geometrical rearrangement, as shown in Figures 3.47c and 3.47d. Indeed, a remarkable detail of this third step is the absence of other atom groups involved in it, which is probably the reason behind the low energy barrier which characterizes it.

Figure 3.48 shows the global trends of the different values of interest over the entire reaction path for furimazine C3 oxygenation process, which shows some crucial differences with the C2 pathway.

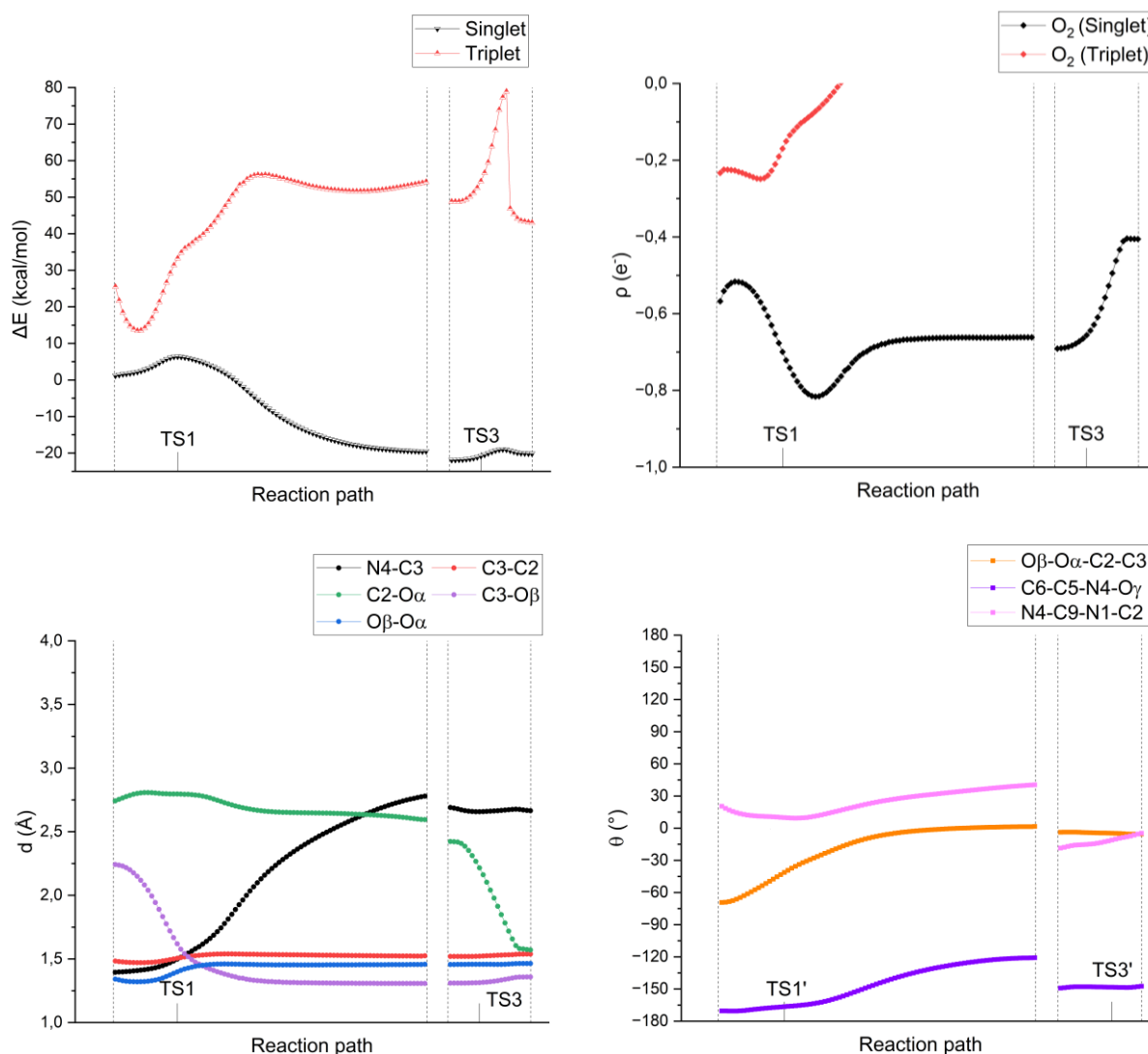


Figure 3. 48 - Trends for energy (a - in singlet and triplet states), electron density on the O_2 atoms (b - measured in terms of Mulliken charges sum), distances of interest (c) and dihedrals of interest (d) for the first and third step of the C3 oxygenation. Each point corresponds to a different geometry given by the IRC calculations. Dashed lines indicate the conclusion of the IRC calculations.

The main difference between the C2 and the C3 pathways is in the energy trends: in the case of C2 oxygenation, the process is a four-step one and almost each step is exergonic, while in the case of C3 the steps are three, only the first being relevantly exergonic. Despite this crucial difference, the final product of two pathways is the same, the anionic dioxetanone which is the key intermediate for the chemiluminescent process of furimazine oxidation.

3.3.2 Decomposition of dioxetanone intermediate

The most important step of the furimazine oxidation in terms of a chemiluminescent process is the decomposition of the dioxetanone intermediate, which is characterized by the crossing between the GS and the ES PESs. In this study, three different possible cases have been studied, differing for the protonation state of DE before the decomposition, in order to check the viability of the formation of FurA, FurC and FurD starting from their corresponding precursors, i.e. differently protonated DE intermediates. As the decomposition process is supposed to give the final product in excited state, the ES geometry for the $\text{Fur}^* + \text{CO}_2$ exciplex was optimized and used as reference structure for the transition state optimization of geometry with QST3 algorithm.

3.3.2.1 FurA formation

FurA formation process was studied starting from a neutral dioxetanone structure, obtained with the protonation of N1 and subsequent optimization of the ground state geometry for the resulting molecular species (DE'). A possible pathway for its decomposition was found, involving a single step process, with a pretty high energy barrier ($\Delta G^\ddagger = 36.1$ kcal/mol), a result similar to one of the proposed mechanisms for another neutral dioxetanone derivative decomposition²⁵⁵.

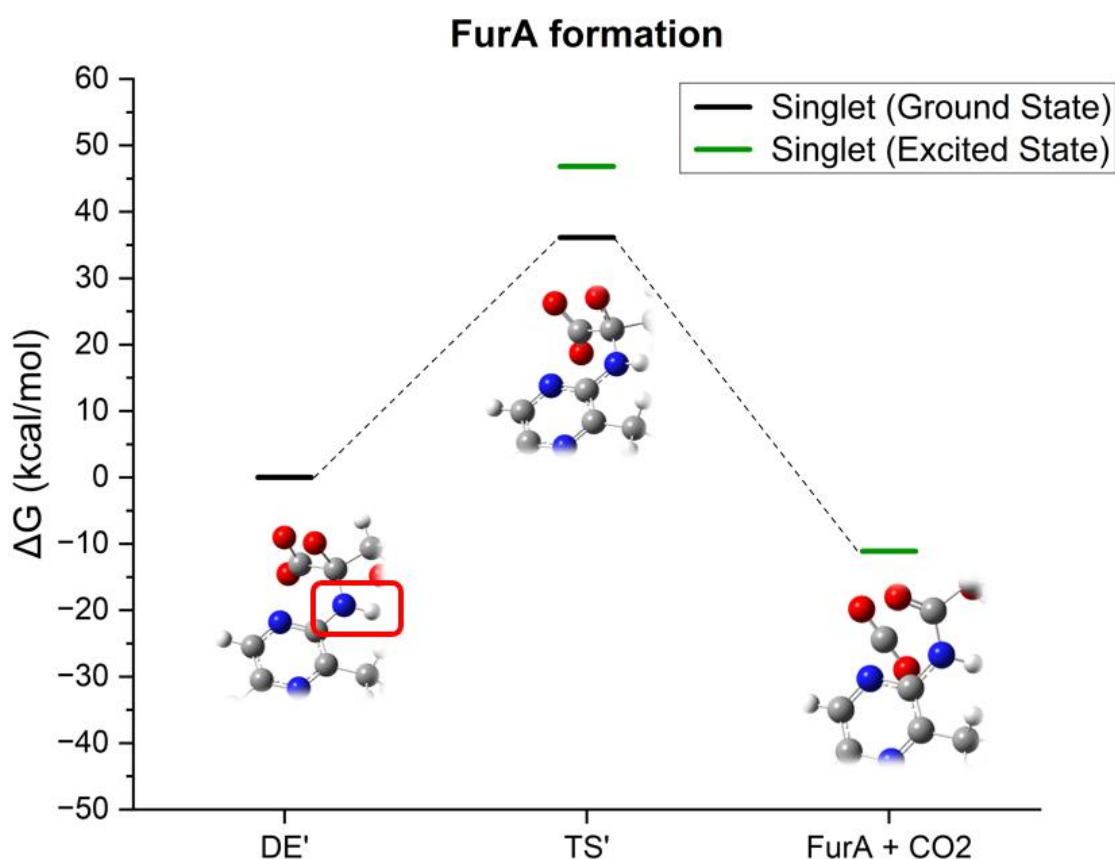


Figure 3. 49 - Free energy profile of the FurA formation process. Free energy of the corresponding optimized dioxetanonic precursor (DE') used as reference. Results at CAM-B3LYP/6-31+G(d,p) level of theory, implicit solvation in DMSO with CPCM model

The reaction is an asynchronous concerted process with CT character, with the elongation of the O β -O α bond followed by the cleavage of the C2-C3 one, as shown in Figure 3.50b. The bond breaking is simultaneous to the linearization of the CO₂ group, as confirmed by the trend reported in Figure 3.50c.

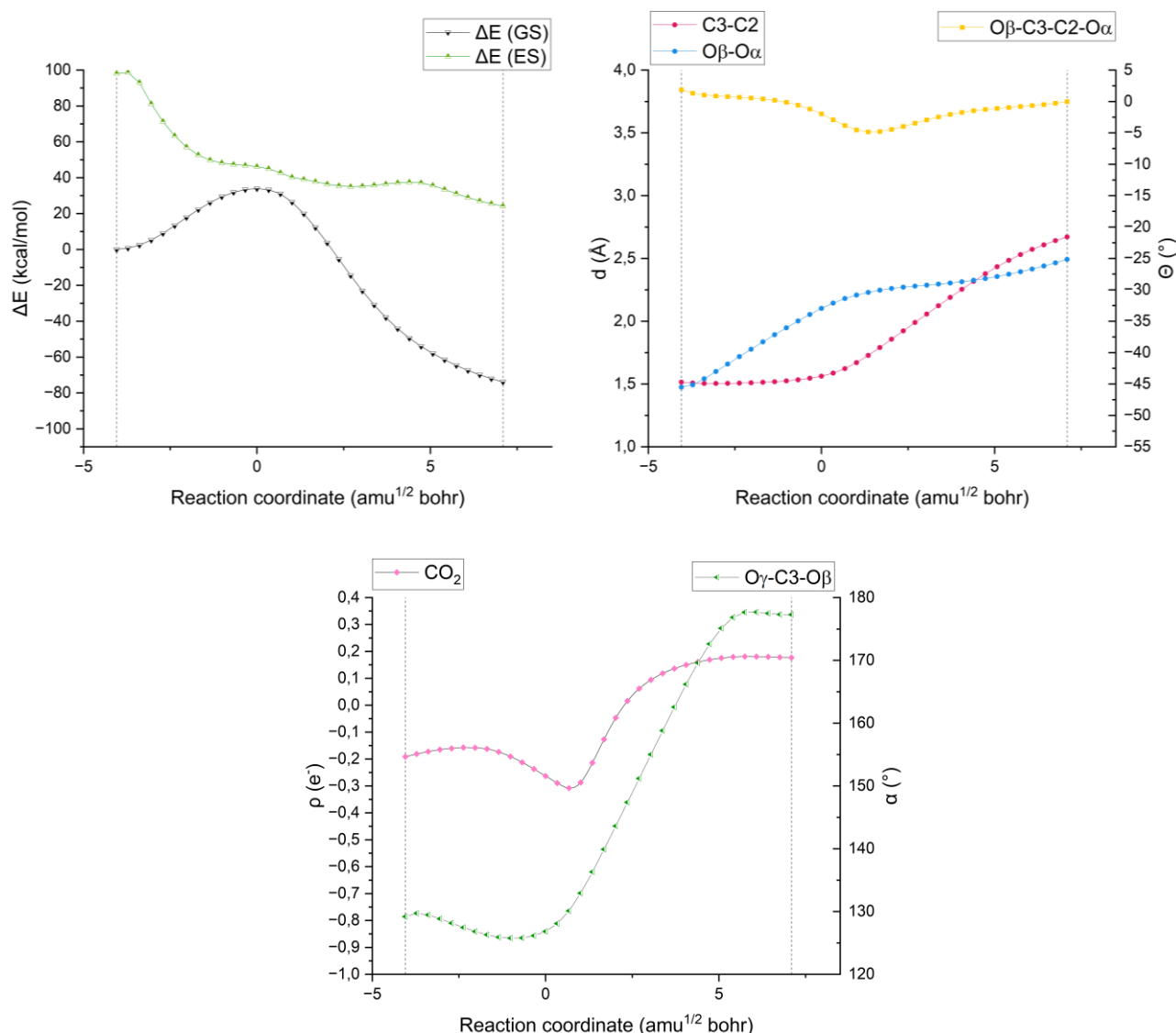


Figure 3. 50 - Trends for energy (a - in ground and excited states), distances of interest and dihedrals of interest (b) and CO₂ fragment electron density (in terms of Mulliken charges sum) and O-C-O angle (c) for FurA formation process. Each point corresponds to a different geometry given by the IRC calculations. Dashed lines indicate the conclusion of the IRC calculations.

The energy gap between the GS potential energy curve and the ES one is at its lowest immediately after the transition state (0.67 amu^{1/2} bohr), as shown in Figure 3.50a and is sufficiently reduced to not be able to exclude the possibility of a crossing between the two curves, also due to the possibility of a diradical character for the transition state, an hypothesis based on the charge transfer and back charge transfer character of the decomposition, highlighted by the trend in CO₂ charge shown in Figure 3.50c.

3.3.2.2 FurC formation

An analogous approach to FurA formation was employed for the FurC one. In particular, this time the neutral dioxetanone (DE'') was obtained with the protonation of N4, a less favourable option in terms of free energy difference between the two structures ($\Delta G = 12.1$ kcal/mol). On the other hand, the final product is at a lower energy than the FurA-CO₂ exciplex ($\Delta G = -6.3$ kcal/mol), as expected. Even in this case the identified mechanism is a single step one, with a lower energy barrier than the previously illustrated pathway ($\Delta G^\ddagger = 26.2$ kcal/mol).

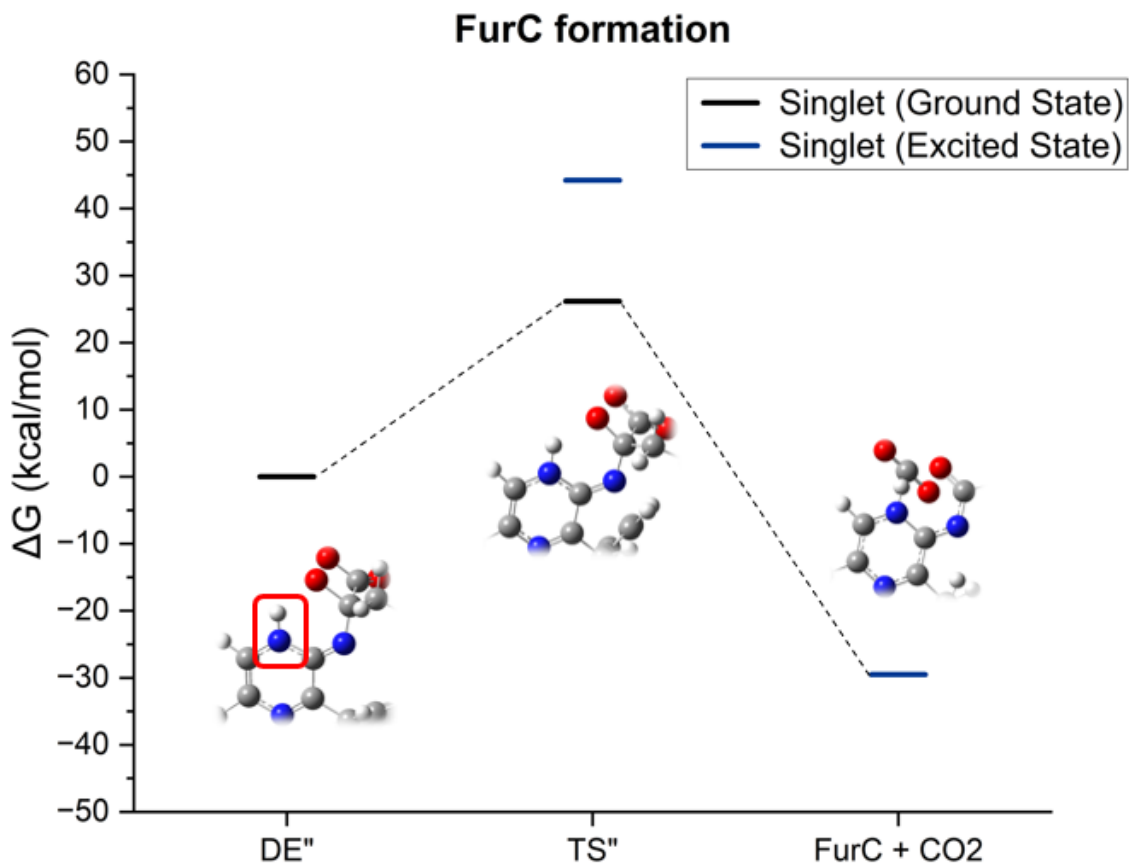


Figure 3. 51 - Free energy profile of the FurC formation process. Free energy of the corresponding optimized dioxetanonic precursor (DE'') used as reference. Results at CAM-B3LYP/6-31+G(d,p) level of theory, implicit solvation in DMSO with CPCM model

The process is more exergonic than the one for FurA formation, partially due to the greater stability of the final exciplex product and in part due to the higher energy of the neutral dioxetanone, but the two decompositions for neutral dioxetanones are both displaying an asynchronous concerted character. Even for FurC formation, the elongation of the O β -O α bond is followed by the cleavage of the C2-C3 one (as confirmed in Figure 3.52b) and the detachment of CO₂ happens after a charge transfer and back charge transfer process, reported in Figure 3.52c.

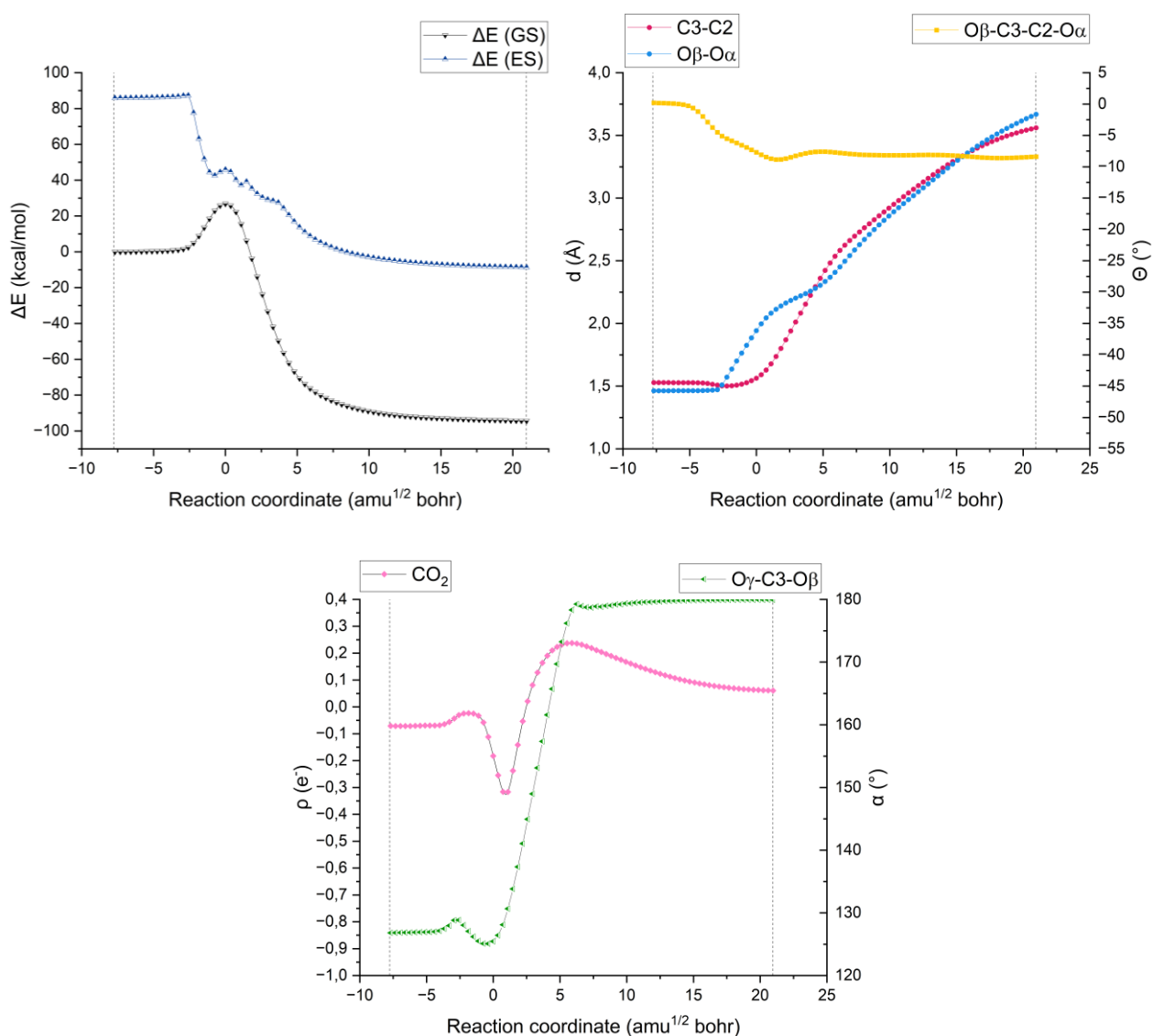


Figure 3. 52 - Trends for energy (a - in ground and excited states), distances of interest and dihedrals of interest (b) and CO₂ fragment electron density (in terms of Mulliken charges sum) and O-C-O angle (c) for FurC formation process. Each point corresponds to a different geometry given by the IRC calculations. Dashed lines indicate the conclusion of the IRC calculations.

In general, the trends are consistent with the ones observed for FurA formation process.

3.3.2.3 FurD formation

The dioxetanone decomposition can happen not only in case of a protonation, either on N1 or N4, but also for the anionic intermediate which was obtained as the result of the C2 and C3 oxygenation pathways. In this case, the process is no more a single step one: it shows two barriers and an intermediate.

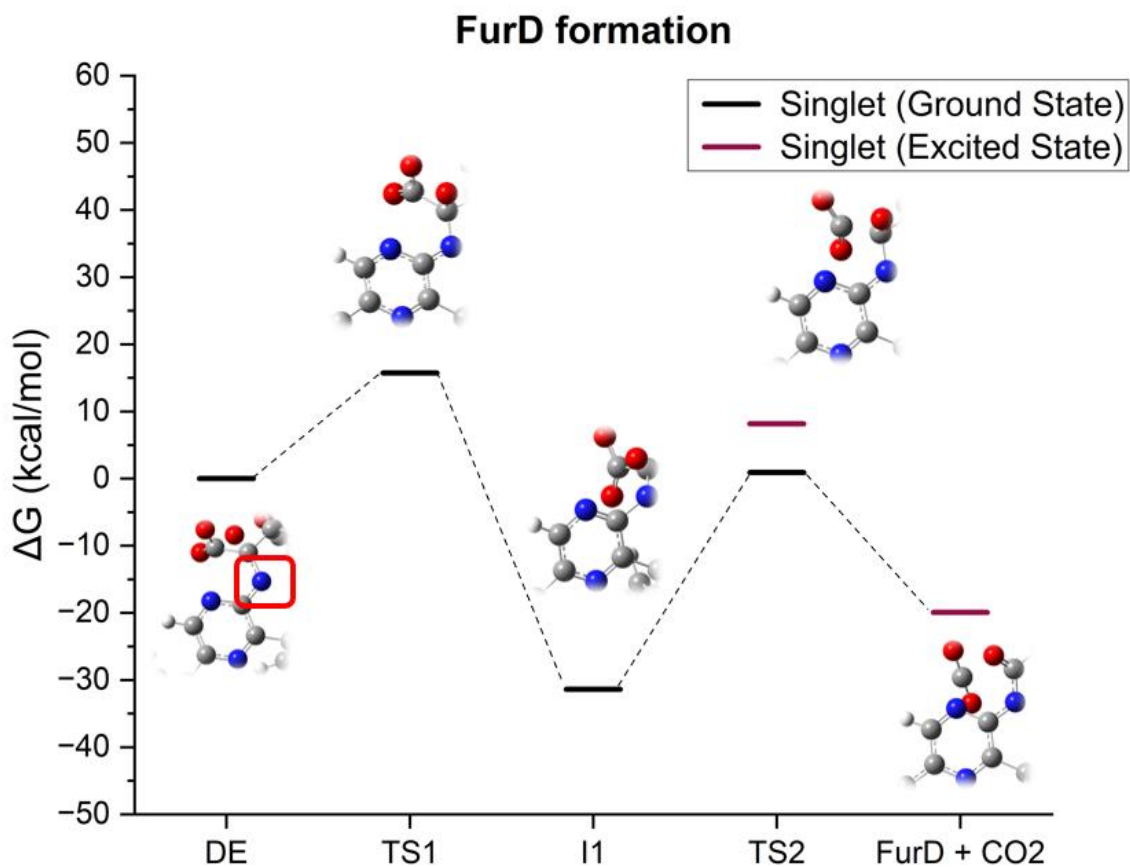


Figure 3. 53 - Free energy profile of the FurD formation process. Free energy of the corresponding optimized dioxetanonic precursor (DE) used as reference. Results at CAM-B3LYP/6-31+G(d,p) level of theory, implicit solvation in DMSO with CPCM model

The first step of the process is characterized by a strong charge transfer character, as reflected by the trend in CO₂ group Mulliken charges, reported in Figure 3.54c: the elongation of the O α -O β bond is accompanied by a strong distortion of the O β -C3-C2-O α dihedral (Figure 3.54b) and leads to the localization of a 0.75 e⁻ charge on the CO₂ atom group.

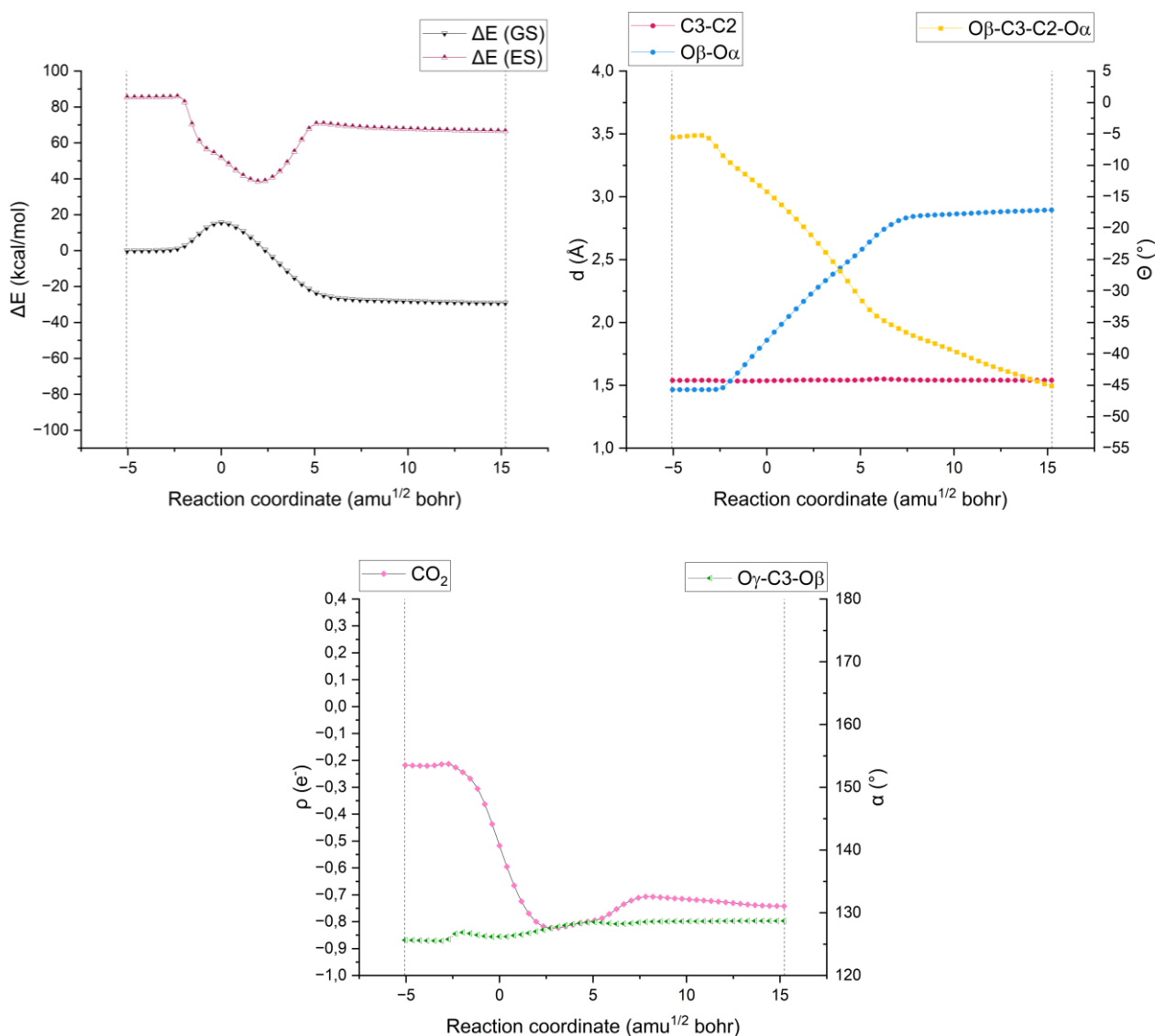


Figure 3. 54 - Trends for energy (a - in ground and excited states), distances of interest and dihedrals of interest (b) and CO₂ fragment electron density (in terms of Mulliken charges sum) and O-C-O angle (c) for the first step of FurD formation process. Each point corresponds to a different geometry given by the IRC calculations. Dashed lines indicate the conclusion of the IRC calculations.

While the first step is characterized by a great separation between the ground and the excited state potential energy curves, the second transition state is corresponding to a strong reduction in the gap between the two, which closes up to 6.4 kcal/mol immediately after the TS geometry along the reaction coordinate (0.3 amu^{1/2} bohr), as shown in Figure 3.55a.

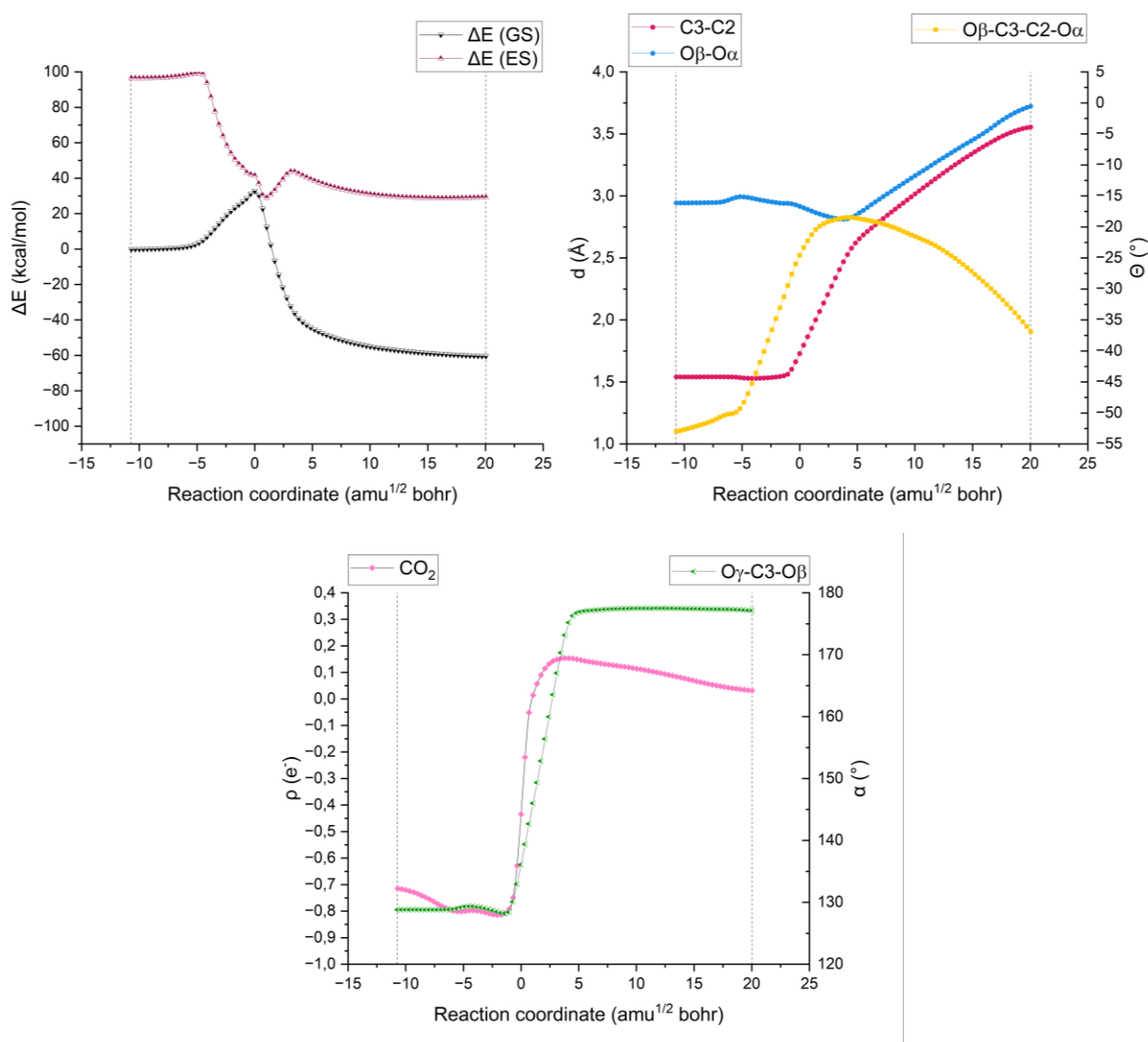


Figure 3. 55 - Trends for energy (a - in ground and excited states), distances of interest and dihedrals of interest (b) and CO₂ fragment electron density (in terms of Mulliken charges sum) and O-C-O angle (c) for the second step of FurD formation process. Each point corresponds to a different geometry given by the IRC calculations. Dashed lines indicate the conclusion of the IRC calculations.

The charge transfer which took place during the previous step is followed by a quick back charge transfer with the CO₂ group rapidly becoming neutral and detaching from the rest of the molecular structure. The decomposition takes place with a partial realignment of the two oxygen atoms, with the reduction of the O β -C3-C2-O α dihedral, followed by a new increase subsequent to the carbon dioxide elimination (Figure 3.55c).

The global trends for energy and relevant geometrical parameters for the anionic dioxetanone decomposition are shown in Figure 3.56.

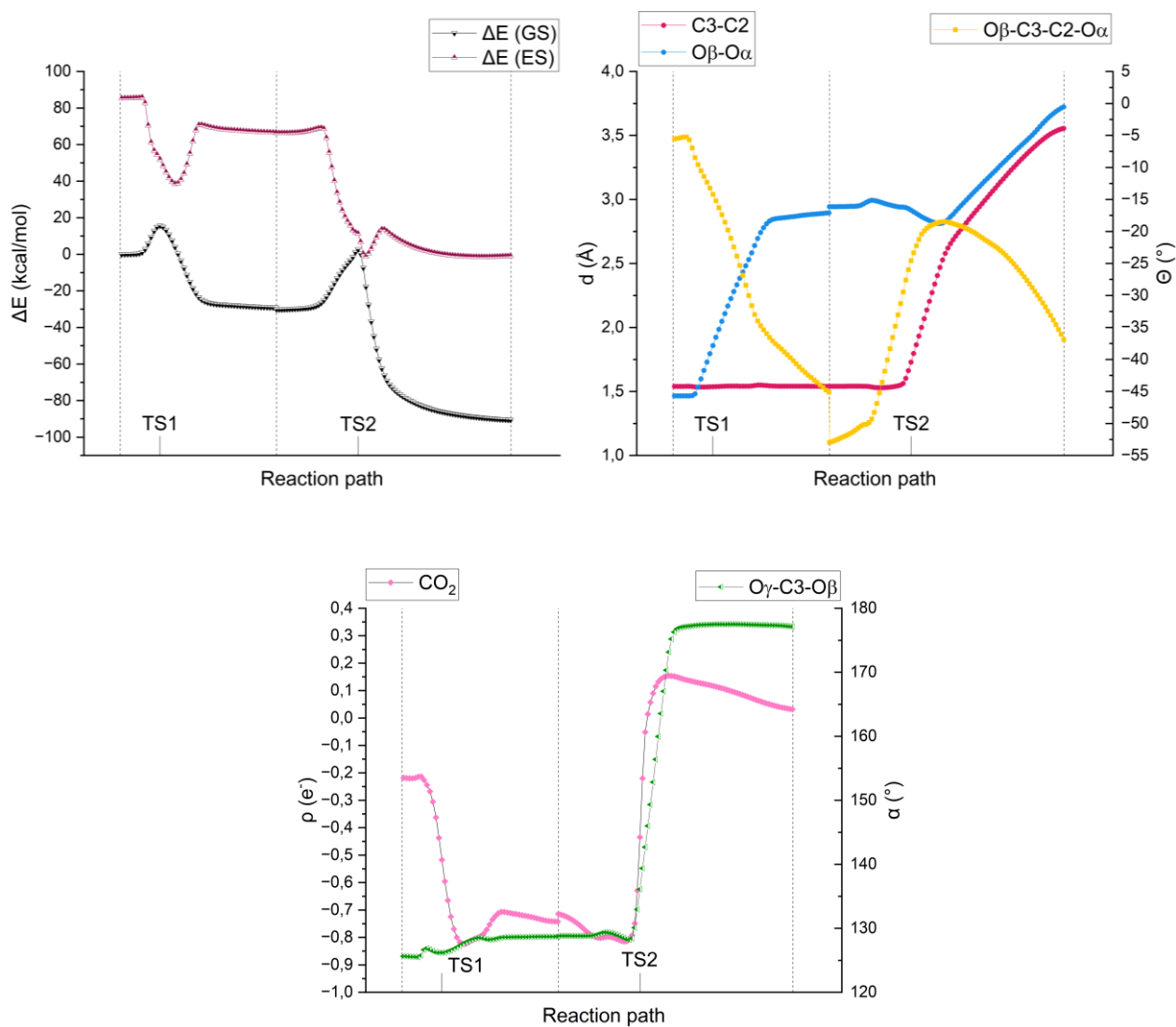


Figure 3. 56 - Trends for energy (a - in ground and excited states), distances of interest and dihedrals of interest (b) and CO_2 fragment electron density (in terms of Mulliken charges sum) and O-C-O angle (c) for FurD formation process. Each point corresponds to a different geometry given by the IRC calculations. Dashed lines indicate the conclusion of the IRC calculations.

3.3.3 Conclusions

The study of the reaction mechanism for furimazine oxidation was focused on two different parts, the evaluation of possible pathways for anionic furimazine oxygenation to give the dioxetanone intermediate and the study of its subsequent decomposition. The first process is started by a single electron transfer to give a diradical pair, whose diradical character is maintained while approaching the transition state associated to C-O bond formation. After the bond formation in the case of C2-oxygenation, and even earlier for the C3-process, it disappears leading to the development of the reaction on the GS singlet PES until the formation of the crucial dioxetanone intermediate. This is a four-step process in the case of the C2-attack, with very low energy barriers, the highest of them being 7.5 kcal/mol, and a three step one for C3-oxygenation, characterized by an higher, but still feasible, energy barrier for the rate determining step (15 kcal/mol). The process is in both cases involving the formation of a peroxide, but the experimental detection of a C3 peroxide could possibly be hard due to the very low energy barriers of the steps separating it from the dioxetanone intermediate. Both pathways have to be considered while studying the reaction mechanism in a protein environment since they are both feasible and the protein residues could play a role in altering the energy barriers values favouring C2 or C3-oxygenation.

The second part of the mechanism is the decomposition of the dioxetanone, which was considered in three different protonation states in order to evaluate the feasibility of FurA, FurC and FurD formation processes. In particular, the crucial feature which was carefully analysed, together with the energy barrier for the CO₂ elimination, is the energy gap between the GS and the ES. According to the obtained results all three mechanisms cannot be excluded since the barriers are high, but not unreasonable, especially considering the possible lowering effect of the protein environment, and the gap is low enough to justify the possibility of having a crossing between GS and ES because the differences could be a consequence of the level of theory employed for the calculations, especially with the crossing points being potentially presenting a diradical character, as shown by the charge transfer and back charge transfer displayed by the region near the TSs of the decompositions pathways.

Future developments preceding the evaluation of the mechanism in the protein environment could include the further evaluation of these regions characteristics focusing on the possibility of a diradical description.

4. Conclusions and perspectives

This thesis investigated some key aspects of the bioluminescence mechanism in the NanoLuc-furimazine system, to provide a detailed picture of the molecular processes that drive this phenomenon. The first part of the study was focused on the identification of the most probable protonation state of both furimazine, the substrate of the enzymatic reaction, and furimamide, the emissive product. Quantum mechanical results suggested that the enol form of furimazine could be involved in the reaction mechanism due to an almost absent energy difference with respect to the corresponding keto form in polar solvents, such as aqueous solutions, once explicit molecules of the solvent are added. This hypothesis was supported by docking studies and molecular dynamics simulations, which showed that the enol form establishes a key hydrogen bond with Asp139 in the NanoLuc active site accounting for the correct ligand placement in the protein pocket and for the beginning of the catalytic process. Indeed, the presence of this H-bond supports a straightforward deprotonation step, required for substrate activation.

As far as it concerns the product protonation state, FurA, the neutral form of furimamide, was ruled out due to the significant blue-shift characterizing its predicted emission with respect to the experimental data. Moreover, when evaluating the relative stability of the globally neutral forms in the ES, FurC, the zwitterionic form of furimamide, emerged as the most stable one. On top of that, FurC simulated emission spectrum was the best fitting one with respect to experimental data, making it the most probable emitter of the NanoLuc-furimazine pair bioluminescent process. While FurD, the anionic form, shared similar emission properties, its basic nature, already investigated in previous studies too, makes it prone to rapid protonation in the physiological pH range, making it a less probable final product. Collectively, these findings strongly support the role of the enol form of furimazine as the reactant and point at FurC as the likely luminescent product.

The second part of the work further investigated the keto-enol tautomerism of furimazine and provided a detailed characterization of furimamide, especially with respect to the tendency to deprotonation. UV/Vis and NMR data delineated a complex picture of keto-enol equilibrium, which could be justified only considering the enol form as the dominant one, with the keto form present in smaller amounts, particularly in polar solvents. Electronic energy calculations with microsolvation cluster models supported this equilibrium, with the enol tautomer being energetically favourable in the presence of explicit solvent interactions and an adequate basis set for their description.

Furimamide protonation state was then further assessed through experimental deprotonation studies. In particular, the neutral form, FurA, showed extreme resistance to deprotonation under basic conditions, reinforcing the idea that the anionic FurD is strongly unlikely to be the final product of the bioluminescent reaction. These results aligned with the computational predictions that FurC, rather than FurD, is the most probable emissive species in the bioluminescent reaction of NanoLuc.

Third and last, the mechanism of furimazine oxidation was investigated, placing special emphasis on the formation and decomposition of the crucial dioxetanone intermediate. The process was examined basing on the hypothesis that a single electron transfer from the deprotonated core of furimazine to molecular oxygen activates the system, leading at subsequent oxygenation at either the C2 or C3 positions. Both pathways were found to be energetically feasible, with C2-oxygenation being slightly preferable than C3 one, thus making it not possible to exclude either of the two options, especially considering the possible role of the protein environment in lowering the energy barriers.

The dioxetanone intermediate decomposition was investigated considering three different protonation states of it, corresponding to FurA, FurC, and FurD precursors. Although high energy barriers were associated with the decomposition process, they were not prohibitive, especially in the context of a

protein environment that could lower these barriers, as it was supposed for the oxygenation process. Additionally, the small energy gap between the ground state and excited state suggested the possibility of a crossing point during the reaction, with a probable diradical character as charge transfer dynamics are evident in all cases.

In conclusion, this work:

- i) provides new insights into the intricate mechanisms governing the NanoLuc-furimazine bioluminescent system by supporting the preference for the enol tautomer of furimazine as substrate
- ii) identifies zwitterionic furimamide (FurC) as the likely emissive product
- iii) explores multiple oxidation pathways, collectively advancing our understanding of this important luciferase-luciferin pair

The mechanistic exploration of furimazine oxidation has highlighted the importance of considering protein environments in future studies, as these could play a pivotal role in modulating reaction barriers and stabilizing intermediates. Future work will aim to build on these results, integrating diradical character analysis for the decomposition pathways to have a more comprehensive evaluation and protein residues effects to provide a more accurate model of NanoLuc bioluminescence mechanism. Ultimately, this deeper understanding could inspire the optimization and engineering of luciferase-luciferin systems for enhanced bioluminescent applications.

References

- [1] A. Bonardi, M. Turelli, G. Moro, U. Cosentino, C. Greco and C. Adamo, Behind the glow: unveiling the nature of NanoLuc reactants and products, *Phys. Chem. Chem. Phys.*, DOI: 10.1039/D4CP02551B
- [2] J. W. Verhoeven, Glossary of terms used in photochemistry (IUPAC Recommendations 1996), *Pure Appl. Chem.*, 1996, **68**, 2223–2286.
- [3] P. J. Herring and E. A. Widder, in *Encyclopedia of Ocean Sciences (Third Edition)*, eds. J. K. Cochran, H. J. Bokuniewicz and P. L. Yager, Academic Press, Oxford, 2001, pp. 508–515.
- [4] E. A. Widder, in *Adaptive Mechanisms in the Ecology of Vision*, eds. S. N. Archer, M. B. A. Djamgoz, E. R. Loew, J. C. Partridge and S. Vallergera, Springer Netherlands, Dordrecht, 1999, pp. 555–581.
- [5] E. A. Widder, Bioluminescence in the Ocean: Origins of Biological, Chemical, and Ecological Diversity, *Science*, 2010, **328**, 704–708.
- [6] S. Martini and S. H. D. Haddock, Quantification of bioluminescence from the surface to the deep sea demonstrates its predominance as an ecological trait, *Sci. Rep.*, 2017, **7**, 45750.
- [7] M. Pointer, L. Carvalho, J. Cowing, J. Bowmaker and D. Hunt, The visual pigments of a deep-sea teleost, the pearl eye Scopelarchus analis, *J. Exp. Biol.*, 2007, **210**, 2829–35.
- [8] A. Apprill, Marine Animal Microbiomes: Toward Understanding Host–Microbiome Interactions in a Changing Ocean, *Front. Mar. Sci.*, 2017, **4**.
- [9] J. Paitio, D. Yano, E. Muneyama, S. Takei, H. Asada, M. Iwasaka and Y. Oba, Reflector of the body photophore in lanternfish is mechanistically tuned to project the biochemical emission in photocytes for counterillumination, *Biochem. Biophys. Res. Commun.*, 2020, **521**, 821–826.
- [10] J. M. Claes and J. Mallefet, Early development of bioluminescence suggests camouflage by counter-illumination in the velvet belly lantern shark *Etmopterus spinax* (Squaloidea: Etmopteridae), *J. Fish Biol.*, 2008, **73**, 1337–1350.
- [11] S. Schramm and D. Weiß, Bioluminescence - The Vibrant Glow of Nature and its Chemical Mechanisms, *ChemBioChem*, 2024, **25**, e202400106.
- [12] L. J. Baker, L. L. Freed, C. G. Easson, J. V. Lopez, D. Fenolio, T. T. Sutton, S. V. Nyholm and T. A. Hendry, Diverse deep-sea anglerfishes share a genetically reduced luminous symbiont that is acquired from the environment, *eLife*, **8**, e47606.
- [13] W. B. Ludt and T. R. Clardy, First detection of biofluorescence in a deep-sea anglerfish, *J. Fish Biol.*, 2022, **100**, 843–846.
- [14] O. Shimomura, T. Masugi, F. H. Johnson and Y. Haneda, Properties and reaction mechanism of the bioluminescence system of the deep-sea shrimp *Oplophorus gracilorostris*, *Biochemistry*, 1978, **17**, 994–998.
- [15] E. M. Thompson, S. Nagata and F. I. Tsuji, *Vargula hilgendorffii* luciferase: a secreted reporter enzyme for monitoring gene expression in mammalian cells, *Gene*, 1990, **96**, 257–262.
- [16] M. Valiadi and D. Iglesias-Rodriguez, Understanding Bioluminescence in Dinoflagellates - How Far Have We Come?, *Microorganisms*, 2013, **1**, 3–25.
- [17] M. D. Burkenroad, A possible function of bioluminescence, *Journ. Mar. Res.*, 1943, **5**, 612.

- [18] M. R. Bowlby, E. A. Widder and J. F. Case, Patterns of Stimulated Bioluminescence in Two Pyrosomes (Tunicata: Pyrosomatidae), *Biol. Bull.*, 1990, **179**, 340–350.
- [19] S. M. Lewis and C. K. Cratsley, Flash Signal Evolution, Mate Choice, and Predation in Fireflies, *Annu. Rev. Entomol.*, 2008, **53**, 293–321.
- [20] H. H. Seliger, The origin of bioluminescence, *Photochem. Photobiol.*, 1975, **21**, 355–361.
- [21] O. Shimomura, Presence of coelenterazine in non-bioluminescent marine organisms, *Comp. Biochem. Physiol. Part B Comp. Biochem.*, 1987, **86**, 361–363.
- [22] J.-F. Rees, B. D. Wergifosse, O. Noiset, M. Dubuisson, B. Janssens and E. M. Thompson, The Origins of Marine Bioluminescence: Turning Oxygen Defence Mechanisms Into Deep-Sea Communication Tools, *J. Exp. Biol.*, 1998, **201**, 1211–1221.
- [23] O. Shimomura, The discovery of aequorin and green fluorescent protein, *J. Microsc.*, 2005, **2017**, 1–15.
- [24] O. Shimomura and I. Yampolsky, *Bioluminescence - Chemical Principles and Methods*, World Scientific, Third Edition., 2019.
- [25] O. Shimomura and F. H. Johnson, Regeneration of the photoprotein aequorin, *Nature*, 1975, **256**, 236–238.
- [26] M. T. Alonso, P. Torres-Vidal, B. Calvo, C. Rodriguez, A. Delrio-Lorenzo, J. Rojo-Ruiz, J. Garcia-Sancho and S. Patel, Use of aequorin-based indicators for monitoring Ca²⁺ in acidic organelles, *Biochim. Biophys. Acta BBA - Mol. Cell Res.*, 2023, **1870**, 119481.
- [27] D. Sanglard, Aequorin as a Useful Calcium-Sensing Reporter in *Candida albicans*, *J. Fungi*, 2021, **7**, 319.
- [28] T. Miyashiro and E. G. Ruby, Shedding light on bioluminescence regulation in *Vibrio fischeri*, *Mol. Microbiol.*, 2012, **84**, 795–806.
- [29] K. L. Visick and M. J. McFall-Ngai, An Exclusive Contract: Specificity in the *Vibrio fischeri*-*Euprymna scolopes* Partnership, *J. Bacteriol.*, 2000, **182**, 1779–1787.
- [30] M. McFall-Ngai, E. A. C. Heath-Heckman, A. A. Gillette, S. M. Peyer and E. A. Harvie, The secret languages of coevolved symbioses: Insights from the *Euprymna scolopes*-*Vibrio fischeri* symbiosis, *Semin. Immunol.*, 2012, **24**, 3–8.
- [31] E. G. Ruby and K.-H. Lee, The *Vibrio fischeri*-*Euprymna scolopes* Light Organ Association: Current Ecological Paradigms, *Appl. Environ. Microbiol.*, 1998, **64**, 805–812.
- [32] M. P. Davis, N. I. Holcroft, E. O. Wiley, J. S. Sparks and W. Leo Smith, Species-specific bioluminescence facilitates speciation in the deep sea, *Mar. Biol.*, 2014, **161**, 1139–1148.
- [33] S. D. Miller, S. H. D. Haddock, C. D. Elvidge and T. F. Lee, Detection of a bioluminescent milky sea from space, *Proc. Natl. Acad. Sci.*, 2005, **102**, 14181–14184.
- [34] S. L. Hinder, G. C. Hays, M. Edwards, E. C. Roberts, A. W. Walne and M. B. Gravenor, Changes in marine dinoflagellate and diatom abundance under climate change, *Nat. Clim. Change*, 2012, **2**, 271–275.
- [35] C. D. van der Lingen, L. Hutchings, T. Lamont and G. C. Pitcher, Climate change, dinoflagellate blooms and sardine in the southern Benguela Current Large Marine Ecosystem, *Environ. Dev.*, 2016, **17**, 230–243.
- [36] J. Lee, *Bioluminescence, the nature of the light*, University of Georgia, 2020.
- [37] W. Lu, A. K. H. Priyashantha, M. C. A. Galappaththi, S. Tibpromma, D.-Q. Dai, N. M. Patabendige, B. M. Premarathne, D. M. Kulasuriya, A. N. Ediriweera, T. S. Nimalrathna, N.

Suwannarach, S. Lumyong, A. Tang, S.-C. Shao and S. C. Karunarathna, Fungal Bioluminescence: Past, Present, and Future, *Diversity*, 2024, **16**, 539.

- [38] M. Mishra and D. Srivastava, Bioluminescent fungi: reviewing nature's riddle!, *J. Mycopathol. Res.*, 2021, **59**, 199–206.
- [39] A. Verdes and D. F. Gruber, Glowing Worms: Biological, Chemical, and Functional Diversity of Bioluminescent Annelids, *Integr. Comp. Biol.*, 2017, **57**, 18–32.
- [40] L. Duchatelet, J. M. Claes, J. Delroisse, P. Flammang and J. Mallefet, Glow on Sharks: State of the Art on Bioluminescence Research, *Oceans*, 2021, **2**, 822–842.
- [41] G. Morciano, A. C. Sarti, S. Marchi, S. Missiroli, S. Falzoni, L. Raffaghello, V. Pistoia, C. Giorgi, F. Di Virgilio and P. Pinton, Use of luciferase probes to measure ATP in living cells and animals, *Nat. Protoc.*, 2017, **12**, 1542–1562.
- [42] E. Brodl, A. Winkler and P. Macheroux, Molecular Mechanisms of Bacterial Bioluminescence, *Comput. Struct. Biotechnol. J.*, 2018, **16**, 551–564.
- [43] H. Fraga, Firefly luminescence: A historical perspective and recent developments, *Photochem. Photobiol. Sci.*, 2008, **7**, 146–158.
- [44] A. Schenkmyerova, M. Toul, D. Pluskal, R. Baatallah, G. Gagnot, G. P. Pinto, V. T. Santana, M. Stuchla, P. Neugebauer, P. Chaiyen, J. Damborsky, D. Bednar, Y. L. Janin, Z. Prokop and M. Marek, Catalytic mechanism for Renilla-type luciferases, *Nat. Catal.*, 2023, **6**, 23–38.
- [45] B.-W. Ding and Y.-J. Liu, Bioluminescence of Firefly Squid via Mechanism of Single Electron-Transfer Oxygenation and Charge-Transfer-Induced Luminescence, *J. Am. Chem. Soc.*, 2017, **139**, 1106–1119.
- [46] D. J. Merritt, E. M. Rodgers, A. F. Amir and A. K. Clarke, Same Temporal Niche, Opposite Rhythmicity: Two Closely Related Bioluminescent Insects With Opposite Bioluminescence Propensity Rhythms, *Chronobiol. Int.*, 2012, **29**, 1336–1344.
- [47] S. Inouye, K. Watanabe, H. Nakamura and O. Shimomura, Secretional luciferase of the luminous shrimp *Oplophorus gracilirostris*: cDNA cloning of a novel imidazopyrazinone luciferase, *FEBS Lett.*, 2000, **481**, 19–25.
- [48] G. Zambito, C. Chawda and L. Mezzanotte, Emerging tools for bioluminescence imaging, *Curr. Opin. Chem. Biol.*, 2021, **63**, 86–94.
- [49] C. Carrasco-López, N. M. Lui, S. Schramm and P. Naumov, The elusive relationship between structure and colour emission in beetle luciferases, *Nature Rev. Chem.*, 2021, **5**, 4–20.
- [50] Y. Jiang, X. Shi, C. Tang and F. Wang, Beyond luciferase-luciferin system: Modification, improved imaging and biomedical application, *Coord. Chem. Rev.*, 2023, **481**, 215045.
- [51] J. He, Multiple Origins of Bioluminescence in Beetles and Evolution of Luciferase Function.
- [52] S. H. D. Haddock, M. A. Moline and J. F. Case, Bioluminescence in the Sea, *Ann. Rev. Mar. Sci.*, 2010, **2**, 443–493.
- [53] A. J. Syed and J. C. Anderson, Applications of bioluminescence in biotechnology and beyond, *Chem. Soc. Rev.*, 2021, **50**, 5668–5705.
- [54] Y. Ando, K. Niwa, N. Yamada, T. Enomoto, T. Irie, H. Kubota, Y. Ohmiya and H. Akiyama, Firefly bioluminescence quantum yield and colour change by pH-sensitive green emission, *Nat. Photonics*, 2008, **2**, 44–47.
- [55] L. Pinto da Silva, A. J. M. Santos and J. C. G. Esteves da Silva, Efficient Firefly Chemi/Bioluminescence: Evidence for Chemiexcitation Resulting from the Decomposition of a Neutral Firefly Dioxetanone Molecule, *J. Phys. Chem. A*, 2013, **117**, 94–100.

- [56] B. R. Branchini, C. E. Behney, T. L. Southworth, D. M. Fontaine, A. M. Gulick, D. J. Vinyard and G. W. Brudvig, Experimental Support for a Single Electron-Transfer Oxidation Mechanism in Firefly Bioluminescence, *J. Am. Chem. Soc.*, 2015, **137**, 7592–7595.
- [57] G. Morciano, H. Imamura, S. Patergnani, G. Pedriali, C. Giorgi and P. Pinton, in *Methods in Cell Biology*, Elsevier, 2020, vol. 155, pp. 199–219.
- [58] B. R. Branchini, T. L. Southworth, D. M. Fontaine, D. Kohrt, M. Talukder, E. Michelini, L. Cevenini, A. Roda and M. J. Grossel, An enhanced chimeric firefly luciferase-inspired enzyme for ATP detection and bioluminescence reporter and imaging applications, *Anal. Biochem.*, 2015, **484**, 148–153.
- [59] A. Lundin, in *Bioluminescence: Fundamentals and Applications in Biotechnology - Volume 2*, eds. G. Thouand and R. Marks, Springer Berlin Heidelberg, Berlin, Heidelberg, 2014, vol. 145, pp. 31–62.
- [60] H. W. Kim and Y. M. Rhee, On the pH Dependent Behavior of the Firefly Bioluminescence: Protein Dynamics and Water Content in the Active Pocket, *J. Phys. Chem. B*, 2013, **117**, 7260–7269.
- [61] D. K. Welsh and T. Noguchi, Cellular Bioluminescence Imaging, *Cold Spring Harb. Protoc.*, 2012, **2012**, pdb.top070607.
- [62] W. W. Lorenz, R. O. McCann, M. Longiaru and M. J. Cormier, Isolation and expression of a cDNA encoding Renilla reniformis luciferase., *Proc. Natl. Acad. Sci.*, 1991, **88**, 4438–4442.
- [63] X. Yang, X. Qin, H. Ji, L. Du and M. Li, Constructing firefly luciferin bioluminescence probes for in vivo imaging, *Org. Biomol. Chem.*, 2022, **20**, 1360–1372.
- [64] Z. Yuan, Q. Jiang and G. Liang, Inspired by nature: Bioluminescent systems for bioimaging applications, *Talanta*, 2025, **281**, 126821.
- [65] J. J. O’Sullivan, V. Medici and M. C. Heffern, A caged imidazopyrazinone for selective bioluminescence detection of labile extracellular copper(II), *Chem. Sci.*, 2022, **13**, 4352–4363.
- [66] S.-B. Kim and T. Furuta, Bioluminescence from the bright and dark sides, *Front. Chem. Biol.*, 2024, **3**, 1459397.
- [67] R. Weissleder and V. Ntziachristos, Shedding light onto live molecular targets, *Nat. Med.*, 2003, **9**, 123–128.
- [68] L. Mezzanotte, M. van ‘t Root, H. Karatas, E. A. Goun and C. W. G. M. Löwik, In Vivo Molecular Bioluminescence Imaging: New Tools and Applications, *Trends Biotechnol.*, 2017, **35**, 640–652.
- [69] S. Golovynskyi, I. Golovynska, L. I. Stepanova, O. I. Datsenko, L. Liu, J. Qu and T. Y. Ohulchansky, Optical windows for head tissues in near-infrared and short-wave infrared regions: Approaching transcranial light applications, *J. Biophotonics*, 2018, **11**, e201800141.
- [70] R. Shinde, J. Perkins and C. H. Contag, Luciferin Derivatives for Enhanced in Vitro and in Vivo Bioluminescence Assays, *Biochemistry*, 2006, **45**, 11103–11112.
- [71] M. S. Evans, J. P. Chaurette, S. T. Adams, G. R. Reddy, M. A. Paley, N. Aronin, J. A. Prescher and S. C. Miller, A synthetic luciferin improves bioluminescence imaging in live mice, *Nat. Methods*, 2014, **11**, 393–395.
- [72] F. Berger, R. Paulmurugan, S. Bhaumik and S. S. Gambhir, Uptake kinetics and biodistribution of ¹⁴C-d-luciferin - a radiolabeled substrate for the firefly luciferase catalyzed bioluminescence reaction: impact on bioluminescence based reporter gene imaging, *Eur. J. Nucl. Med. Mol. Imaging*, 2008, **35**, 2275–2285.

- [73] S. Iwano, M. Sugiyama, H. Hama, A. Watakabe, N. Hasegawa, T. Kuchimaru, K. Z. Tanaka, M. Takahashi, Y. Ishida, J. Hata, S. Shimozono, K. Namiki, T. Fukano, M. Kiyama, H. Okano, S. Kizaka-Kondoh, T. J. McHugh, T. Yamamori, H. Hioki, S. Maki and A. Miyawaki, Single-cell bioluminescence imaging of deep tissue in freely moving animals, *Science*, 2018, **359**, 935–939.
- [74] H. Zhao, T. C. Doyle, O. Coquoz, F. Kalish, B. W. Rice and C. H. Contag, Emission spectra of bioluminescent reporters and interaction with mammalian tissue determine the sensitivity of detection in vivo, *J. Biomed. Opt.*, 2005, **10**, 041210.
- [75] E. H. White, H. Wörther, H. H. Seliger and W. D. McElroy, Amino Analogs of Firefly Luciferin and Biological Activity Thereof1, *J. Am. Chem. Soc.*, 1966, **88**, 2015–2019.
- [76] Z. M. Kaskova, A. S. Tsarkova and I. V. Yampolsky, 1001 lights: luciferins, luciferases, their mechanisms of action and applications in chemical analysis, biology and medicine, *Chem. Soc. Rev.*, 2016, **45**, 6048–6077.
- [77] M. P. Hall, C. C. Woodroffe, M. G. Wood, I. Que, M. van't Root, Y. Ridwan, C. Shi, T. A. Kirkland, L. P. Encell, K. V. Wood, C. Löwik and L. Mezzanotte, Click beetle luciferase mutant and near infrared naphthyl-luciferins for improved bioluminescence imaging, *Nat. Commun.*, 2018, **9**, 132.
- [78] R. Geiger, E. Schneider, K. Wallenfels and W. Miska, A New Ultrasensitive Bioluminogenic Enzyme Substrate for β -Galactosidase, 1992, **373**, 1187–1192.
- [79] T. Monsees, W. Miska and R. Geiger, Synthesis and Characterization of a Bioluminogenic Substrate for α -Chymotrypsin, *Anal. Biochem.*, 1994, **221**, 329–334.
- [80] J. Li, L. Chen, W. Wu, W. Zhang, Z. Ma, Y. Cheng, L. Du and M. Li, Discovery of Bioluminogenic Probes for Aminopeptidase N Imaging, *Anal. Chem.*, 2014, **86**, 2747–2751.
- [81] P. Feng, H. Zhang, Q. Deng, W. Liu, L. Yang, G. Li, G. Chen, L. Du, B. Ke, M. Li, Real-Time Bioluminescence Imaging of Nitroreductase in Mouse Model, *Anal. Chem.*, 2016, **88**, 5610–5614.
- [82] Y. Gao, Y. Lin, T. Liu, H. Chen, X. Yang, C. Tian, L. Du and M. Li, Bioluminescent Probe for Tumor Hypoxia Detection via CYP450 Reductase in Living Animals, *Anal. Chem.*, 2017, **89**, 12488–12493.
- [83] W. R. Wilson and M. P. Hay, Targeting hypoxia in cancer therapy, *Nat. Rev. Cancer*, 2011, **11**, 393–410.
- [84] W. Wenxiao, J. Li, L. Chen, Z. Ma, W. Zhang, L. Du, M. Li, Bioluminescent Probe for Hydrogen Peroxide Imaging in Vitro and in Vivo, *Anal. Chem.*, 2014, **86**.
- [85] S. G. Rhee, H₂O₂, a Necessary Evil for Cell Signaling, *Science*, 2006, **312**, 1882–1883.
- [86] Y. Wang, V. Hodgkinson, S. Zhu, G. A. Weisman and M. J. Petris, Advances in the Understanding of Mammalian Copper Transporters, *Adv. Nutr.*, 2011, **2**, 129–137.
- [87] M. A. Sellmyer, L. Bronsart, H. Imoto, C. H. Contag, T. J. Wandless, J. A. Prescher, Visualizing cellular interactions with a generalized proximity reporter, *PNAS*, **110**, 8567–8572.
- [88] M. Verhaegen and T. K. Christopoulos, Recombinant Gaussia Luciferase. Overexpression, Purification, and Analytical Application of a Bioluminescent Reporter for DNA Hybridization, *Anal. Chem.*, 2002, **74**, 4378–4385.
- [89] S. V. Markova and E. S. Vysotski, Coelenterazine-dependent luciferases, *Biochem. Mosc.*, 2015, **80**, 714–732.
- [90] A. M. Loening, Consensus guided mutagenesis of Renilla luciferase yields enhanced stability and light output, *Protein Eng. Des. Sel.*, 2006, **19**, 391–400.

- [91] B. A. Tannous and J. Teng, Secreted blood reporters: Insights and applications, *Biotechnol. Adv.*, 2011, **29**, 997–1003.
- [92] M. Hattori and T. Ozawa, Split Luciferase Complementation for Analysis of Intracellular Signaling, *Anal. Sci.*, 2014, **30**, 539–544.
- [93] K. E. Luker, L. A. Mihalko, B. T. Schmidt, S. A. Lewin, P. Ray, D. Shcherbo, D. M. Chudakov and G. D. Luker, In vivo imaging of ligand receptor binding with Gaussia luciferase complementation, *Nat. Med.*, 2012, **18**, 172–177.
- [94] K. A. Jones, D. J. Li, E. Hui, M. A. Sellmyer and J. A. Prescher, Visualizing Cell Proximity with Genetically Encoded Bioluminescent Reporters, *ACS Chem. Biol.*, 2015, **10**, 933–938.
- [95] E. S. Vysotski and J. Lee, Ca²⁺-Regulated Photoproteins: Structural Insight into the Bioluminescence Mechanism, *Acc. Chem. Res.*, 2004, **37**, 405–415.
- [96] A. Samanta and I. L. Medintz, Bioluminescence-Based Energy Transfer Using Semiconductor Quantum Dots as Acceptors, *Sensors*, 2020, **20**, 2909.
- [97] Y. Xu, D. W. Piston and C. H. Johnson, A bioluminescence resonance energy transfer (BRET) system: Application to interacting circadian clock proteins, *Proc. Natl. Acad. Sci.*, 1999, **96**, 151–156.
- [98] A. Taylor, J. Sharkey, A. Plagge, B. Wilm and P. Murray, Multicolour In Vivo Bioluminescence Imaging Using a NanoLuc-Based BRET Reporter in Combination with Firefly Luciferase, *Contrast Media Mol. Imaging*, 2018, **2018**, e2514796.
- [99] R. Sleno, D. Pétrin, D. Devost, E. Goupil, A. Zhang and T. E. Hébert, Designing BRET-based conformational biosensors for G protein-coupled receptors, *Methods*, 2016, **92**, 11–18.
- [100] T. Azad, H. J. Janse van Rensburg, J. Morgan, R. Rezaei, M. J. F. Crupi, R. Chen, M. Ghahremani, M. Jamalkhah, N. Forbes, C. Ilkow, J. C. Bell, Luciferase-Based Biosensors in the Era of the COVID-19 Pandemic, *ACS Nanosci. Au.*, 2021, **1**, 15–37.
- [101] A. Salahpour, S. Espinoza, B. Masri, V. Lam, L. Barak and R. R. Gainetdinov, BRET biosensors to study GPCR biology, pharmacology, and signal transduction, *Front. Endocrinol.*, 2012, **3**.
- [102] Y.-J. Liu, Understanding the complete bioluminescence cycle from a multiscale computational perspective: A review, *J. Photochem. Photobiol. C Photochem. Rev.*, 2022, **52**, 100537.
- [103] M. Yu and Y. Liu, A QM/MM Study on the Initiation Reaction of Firefly Bioluminescence - Enzymatic Oxidation of Luciferin, *Molecules*, 2021, **26**, 4222.
- [104] G. Orlova, J. D. Goddard and L. Yu. Brovko, Theoretical Study of the Amazing Firefly Bioluminescence: The Formation and Structures of the Light Emitters, *J. Am. Chem. Soc.*, 2003, **125**, 6962–6971.
- [105] I. Navizet, D. Roca-Sanjuán, L. Yue, Y.-J. Liu, N. Ferré and R. Lindh, Are the Bio- and Chemiluminescence States of the Firefly Oxyluciferin the Same as the Fluorescence State?, *Photochem. Photobiol.*, 2013, **89**, 319–325.
- [106] F. Duarte, A. Barrozo, J. Åqvist, N. H. Williams and S. C. L. Kamerlin, The Competing Mechanisms of Phosphate Monoester Dianion Hydrolysis, *J. Am. Chem. Soc.*, 2016, **138**, 10664–10673.
- [107] A. Ghose, M. Rebarz, O. V. Maltsev, L. Hintermann, C. Ruckebusch, E. Fron, J. Hofkens, Y. Mély, P. Naumov, M. Sliwa and P. Didier, Emission Properties of Oxyluciferin and Its Derivatives in Water: Revealing the Nature of the Emissive Species in Firefly Bioluminescence, *J. Phys. Chem. B*, 2015, **119**, 2638–2649.
- [108] L. Pinto Da Silva and J. C. G. Esteves Da Silva, Computational Investigation of the Effect of pH on the Color of Firefly Bioluminescence by DFT, *ChemPhysChem*, 2011, **12**, 951–960.

- [109] Z.-S. Li, L.-Y. Zou, C.-G. Min and A.-M. Ren, The effect of micro-environment on luminescence of aequorin: The role of amino acids and explicit water molecules on spectroscopic properties of coelenteramide, *J. Photochem. Photobiol. B*, 2013, **127**, 94–99.
- [110] U. T. Bornscheuer and M. Pohl, Improved biocatalysts by directed evolution and rational protein design, *Curr. Opin. Chem. Biol.*, 2001, **5**, 137–143.
- [111] H. Yu and H. Huang, Engineering proteins for thermostability through rigidifying flexible sites, *Biotechnol. Adv.*, 2014, **32**, 308–315.
- [112] A. A. Dombkowski, K. Z. Sultana and D. B. Craig, Protein disulfide engineering, *FEBS Lett.*, 2014, **588**, 206–212.
- [113] H. Cui, L. Eltoukhy, L. Zhang, U. Markel, K.-E. Jaeger, M. D. Davari and U. Schwaneberg, Less Unfavorable Salt Bridges on the Enzyme Surface Result in More Organic Cosolvent Resistance, *Angew. Chem. Int. Ed.*, 2021, **60**, 11448–11456.
- [114] W. Yao, L. Mo, L. Fang and L. Qin, Molecular dynamics simulations on interactions of five antibiotics with luciferase of *Vibrio Qinghaiensis* sp.-Q67, *Ecotoxicol. Environ. Saf.*, 2023, **256**, 114910.
- [115] B. R. Branchini, T. L. Southworth, M. H. Murtiashaw, H. Boije and S. E. Fleet, A Mutagenesis Study of the Putative Luciferin Binding Site Residues of Firefly Luciferase, *Biochemistry*, 2003, **42**, 10429–10436.
- [116] M. Rahban, N. Salehi, A. A. Saboury, S. Hosseinkhani, M. H. Karimi-Jafari, R. Firouzi, N. Rezaei-Ghaleh and A. A. Moosavi-Movahedi, Histidine substitution in the most flexible fragments of firefly luciferase modifies its thermal stability, *Arch. Biochem. Biophys.*, 2017, **629**, 8–18.
- [117] M. T. Muhammed and E. Aki-Yalcin, Homology modeling in drug discovery: Overview, current applications, and future perspectives, *Chem. Biol. Drug Des.*, 2019, **93**, 12–20.
- [118] S. Inouye and S. Sasaki, Overexpression, purification and characterization of the catalytic component of Oplophorus luciferase in the deep-sea shrimp, *Oplophorus gracilirostris*, *Protein Expr. Purif.*, 2007, **56**, 261–268.
- [119] M. A. Kurowski and J. M. Bujnicki, GeneSilico protein structure prediction meta-server, *Nucleic Acids Res.*, 2003, **31**, 3305–3307.
- [120] D. R. Flower, A. C. T. North and C. E. Sansom, The lipocalin protein family: structural and sequence overview, *Biochim. Biophys. Acta BBA - Protein Struct. Mol. Enzymol.*, 2000, **1482**, 9–24.
- [121] M. P. Hall, J. Unch, B. F. Binkowski, M. P. Valley, B. L. Butler, M. G. Wood, P. Otto, K. Zimmerman, G. Vidugiris, T. Machleidt, M. B. Robers, H. A. Benink, C. T. Eggers, M. R. Slater, P. L. Meisenheimer, D. H. Klaubert, F. Fan, L. P. Encell and K. V. Wood, Engineered Luciferase Reporter from a Deep Sea Shrimp Utilizing a Novel Imidazopyrazinone Substrate, *ACS Chem. Biol.*, 2012, **7**, 1848–1857.
- [122] D. R. Mills, R. L. Peterson and S. Spiegelman, An extracellular Darwinian experiment with a self-duplicating nucleic acid molecule., *Proc. Natl. Acad. Sci.*, 1967, **58**, 217–224.
- [123] L. Sellés Vidal, M. Isalan, J. T. Heap and R. Ledesma-Amaro, A primer to directed evolution: current methodologies and future directions, *RSC Chem. Biol.*, **4**, 271–291.
- [124] L. Yu, Z. Liu, W. Xu, K. Jin, J. Liu, X. Zhu, Y. Zhang and Y. Wu, Towards overcoming obstacles of type II photodynamic therapy: Endogenous production of light, photosensitizer, and oxygen, *Acta Pharm. Sin. B*, 2024, **14**, 1111–1131.
- [125] S. Inouye and O. Shimomura, The Use of Renilla Luciferase, Oplophorus Luciferase, and Apoequorin as Bioluminescent Reporter Protein in the Presence of Coelenterazine Analogues as Substrate, *Biochem. Biophys. Res. Commun.*, 1997, **233**, 349–353.

- [126] A. C. Stacer, S. Nyati, P. Moudgil, R. Iyengar, K. E. Luker, A. Rehemtulla and G. D. Luker, NanoLuc Reporter for Dual Luciferase Imaging in Living Animals, *Mol. Imaging*, 2013, **12**, 1-13.
- [127] V. Tran, L. A. Moser, D. S. Poole and A. Mehle, Highly sensitive real-time in vivo imaging of an influenza reporter virus reveals dynamics of replication and spread, *J. Virol.*, 2013, **87**, 13321–13329.
- [128] A. S. Dixon, M. K. Schwinn, M. P. Hall, K. Zimmerman, P. Otto, T. H. Lubben, B. L. Butler, B. F. Binkowski, T. Machleidt, T. A. Kirkland, M. G. Wood, C. T. Eggers, L. P. Encell and K. V. Wood, NanoLuc Complementation Reporter Optimized for Accurate Measurement of Protein Interactions in Cells, *ACS Chem. Biol.*, 2015, **11**, 400–408.
- [129] H. T. Nguyen, S. Hurh, L. P. Nguyen, T. U. Nguyen, H.-K. Park, J. Y. Seong, C. S. Lee, B.-J. Ham and J.-I. Hwang, Functional Analysis of CXCR3 Splicing Variants and Their Ligands Using NanoBiT-Based Molecular Interaction Assays, *Mol. Cells*, 2023, **46**, 281–297.
- [130] J. Norisada, K. Fujimura, F. Amaya, H. Kohno, Y. Hirata and K. Oh-hashii, Application of NanoBiT for Monitoring Dimerization of the Null Hong Kong Variant of α -1-Antitrypsin, NHK, in Living Cells, *Mol. Biotechnol.*, 2018, **60**, 539–549.
- [131] R. Rozbeh and K. Forchhammer, In Vivo Detection of Metabolic Fluctuations in Real Time Using the NanoBiT Technology Based on PII Signalling Protein Interactions, *Int. J. Mol. Sci.*, 2024, **25**, 3409.
- [132] G. Song, Q. Jiang, T. Xu, Y.-L. Liu, Z.-G. Xu and Z.-Y. Guo, A convenient luminescence assay of ferroportin internalization to study its interaction with hepcidin, *FEBS J.*, 2013, **280**, 1773–1781.
- [133] Y. Yasuzaki, Y. Yamada, T. Ishikawa and H. Harashima, Validation of Mitochondrial Gene Delivery in Liver and Skeletal Muscle via Hydrodynamic Injection Using an Artificial Mitochondrial Reporter DNA Vector, *Mol. Pharm.*, 2015, **12**, 4311–4320.
- [134] C. G. England, E. B. Ehlerding and W. Cai, NanoLuc: A Small Luciferase Is Brightening Up the Field of Bioluminescence, *Bioconjug. Chem.*, 2016, **27**, 1175–1187.
- [135] L. Biewenga, B. J. H. M. Rosier and M. Merckx, Engineering with NanoLuc: a playground for the development of bioluminescent protein switches and sensors, *Biochem. Soc. Trans.*, 2020, **48**, 2643–2655.
- [136] A. den Hamer, P. Dierickx, R. Arts, J. S. P. M. de Vries, L. Brunsveld and M. Merckx, Bright Bioluminescent BRET Sensor Proteins for Measuring Intracellular Caspase Activity, *ACS Sens.*, 2017, **2**, 729–734.
- [137] J. H. Correia, J. A. Rodrigues, S. Pimenta, T. Dong and Z. Yang, Photodynamic Therapy Review: Principles, Photosensitizers, Applications, and Future Directions, *Pharmaceutics*, 2021, **13**, 1332.
- [138] C.-Y. Hsu, C.-W. Chen, H.-P. Yu, Y.-F. Lin and P.-S. Lai, Bioluminescence resonance energy transfer using luciferase-immobilized quantum dots for self-illuminated photodynamic therapy, *Biomaterials*, 2013, **34**, 1204–1212.
- [139] B. Gidwani, V. Sahu, S. S. Shukla, R. Pandey, V. Joshi, V. K. Jain and A. Vyas, Quantum dots: Prospectives, toxicity, advances and applications, *J. Drug Deliv. Sci. Technol.*, 2021, **61**, 102308.
- [140] X. Shu, V. Lev-Ram, T. J. Deerinck, Y. Qi, E. B. Ramko, M. W. Davidson, Y. Jin, M. H. Ellisman and R. Y. Tsien, A Genetically Encoded Tag for Correlated Light and Electron Microscopy of Intact Cells, Tissues, and Organisms, *PLOS Biol.*, 2011, **9**, e1001041.
- [141] E. A. Souslova, K. E. Mironova and S. M. Deyev, Applications of genetically encoded photosensitizer miniSOG: from correlative light electron microscopy to immunophotosensitizing, *J. Biophotonics*, 2017, **10**, 338–352.

- [142] G. M. Proshkina, E. I. Shramova, O. N. Shilova, A. V. Ryabova and S. M. Deyev, Phototoxicity of flavoprotein miniSOG induced by bioluminescence resonance energy transfer in genetically encoded system NanoLuc-miniSOG is comparable with its LED-excited phototoxicity, *J. Photochem. Photobiol. B*, 2018, **188**, 107–115.
- [143] A. Shakhmin, M. P. Hall, T. Machleidt, J. R. Walker, K. V. Wood and T. A. Kirkland, Coelenterazine analogues emit red-shifted bioluminescence with NanoLuc, *Org. Biomol. Chem.*, 2017, **15**, 8559–8567.
- [144] C. Yan, L. Du and M. Li, Novel NanoLuc-type substrates with various C-6 substitutions, *Bioorg. Med. Chem. Lett.*, 2020, **30**, 127085.
- [145] J. Chu, Y. Oh, A. Sens, N. Ataie, H. Dana, J. J. Macklin, T. Laviv, E. S. Welf, K. M. Dean, F. Zhang, B. B. Kim, C. T. Tang, M. Hu, M. A. Baird, M. W. Davidson, M. A. Kay, R. Fiolka, R. Yasuda, D. S. Kim, H.-L. Ng and M. Z. Lin, A bright cyan-excitable orange fluorescent protein facilitates dual-emission microscopy and enhances bioluminescence imaging in vivo, *Nat. Biotechnol.*, 2016, **34**, 760–767.
- [146] H.-W. Yeh, O. Karmach, A. Ji, D. Carter, M. M. Martins-Green and H. Ai, Red-shifted luciferase–luciferin pairs for enhanced bioluminescence imaging, *Nat. Methods*, 2017, **14**, 971–978.
- [147] K. Suzuki, T. Kimura, H. Shinoda, G. Bai, M. J. Daniels, Y. Arai, M. Nakano and T. Nagai, Five colour variants of bright luminescent protein for real-time multicolour bioimaging, *Nat. Commun.*, 2016, **7**, 13718.
- [148] F. Jensen, *Introduction to computational chemistry*, Wiley, Chichester, West Sussex Hoboken, NJ Oxford, Third edition., 2017.
- [149] A. D. Mackerell, Empirical force fields for biological macromolecules: overview and issues, *J. Comput. Chem.*, 2004, **25**, 1584–1604.
- [150] R. C. Tryon, Cluster analysis: correlation profile and orthometric (factor) analysis for the isolation of unities in mind and personality., Edwards brothers inc., Ann Arbor, Mich., 1939.
- [151] T. S. Madhulatha, An overview on clustering methods, *IOSR J. Eng.*, 2012, **2**, 719–725.
- [152] T. Hansson, C. Oostenbrink and W. van Gunsteren, Molecular dynamics simulations, *Curr. Opin. Struct. Biol.*, 2002, **12**, 190–196.
- [153] J.-P. Ryckaert, G. Ciccotti and H. J. C. Berendsen, Numerical integration of the cartesian equations of motion of a system with constraints: molecular dynamics of *n*-alkanes, *J. Comput. Phys.*, 1977, **23**, 327–341.
- [154] L. Verlet, Computer “Experiments” on Classical Fluids. I. Thermodynamical Properties of Lennard-Jones Molecules, *Phys. Rev.*, 1967, **159**, 98–103.
- [155] W. Kabsch and C. Sander, Dictionary of Protein Secondary Structure: Pattern Recognition of Hydrogen-Bonded and Geometrical Features, *Biopolymers*, 1983, **22**, 2577–2637.
- [156] S. K. Popat, Review and Comparative Study of Clustering Techniques, *IJCSIT*, 2014, **5**, 805–812.
- [157] X. Daura, K. Gademann, B. Jaun, D. Seebach, W. F. van Gunsteren and A. E. Mark, Peptide Folding: When Simulation Meets Experiment, *Angew. Chem. Int. Ed.*, 1999, **38**, 236–240.
- [158] C. J. Cramer, *Essentials of computational chemistry: theories and models*, Wiley, Chichester, West Sussex, England ; Hoboken, NJ, 2nd ed., 2004.
- [159] E. Engel and R. M. Dreizler, *Density Functional Theory: An Advanced Course*, Springer, Berlin, Heidelberg, 2011.
- [160] S. Genheden and U. Ryde, The MM/PBSA and MM/GBSA methods to estimate ligand-binding affinities, *Expert Opin. Drug Discov.*, 2015, **10**, 449–461.

- [161] M. R. Shirts, D. L. Mobley and J. D. Chodera, in *Annual Reports in Computational Chemistry*, eds. D. C. Spellmeyer and R. Wheeler, Elsevier, 2007, vol. 3, pp. 41–59.
- [162] J. Srinivasan, T. E. Cheatham, P. Cieplak, P. A. Kollman and D. A. Case, Continuum Solvent Studies of the Stability of DNA, RNA, and Phosphoramidate–DNA Helices, *J. Am. Chem. Soc.*, 1998, **120**, 9401–9409.
- [163] S.-Y. Huang and X. Zou, Advances and Challenges in Protein-Ligand Docking, *Int. J. Mol. Sci.*, 2010, **11**, 3016–3034.
- [164] I. Halperin, B. Ma, H. Wolfson and R. Nussinov, Principles of docking: An overview of search algorithms and a guide to scoring functions, *Proteins Struct. Funct. Bioinforma.*, 2002, **47**, 409–443.
- [165] C.-J. Tsai, B. Ma, Y. Y. Sham, S. Kumar and R. Nussinov, Structured disorder and conformational selection, *Proteins Struct. Funct. Bioinforma.*, 2001, **44**, 418–427.
- [166] M. J. Todd, N. Semo and E. Freire, The structural stability of the HIV-1 protease¹, *J. Mol. Biol.*, 1998, **283**, 475–488.
- [167] M. J. Todd and E. Freire, The effect of inhibitor binding on the structural stability and cooperativity of the HIV-1 protease, *Proteins Struct. Funct. Bioinforma.*, 1999, **36**, 147–156.
- [168] C. Baysal, A. R. Atilgan, Coordination topology and stability for the native and binding conformers of chymotrypsin inhibitor 2, *Proteins*, 2001, **45**, 62–70.
- [169] J. H. Holland, *Adaptation in Natural and Artificial Systems: An Introductory Analysis with Applications to Biology, Control, and Artificial Intelligence*, MIT Press, 1992.
- [170] E. J. Gardiner, P. Willett and P. J. Artymiuk, Protein docking using a genetic algorithm, *Proteins Struct. Funct. Bioinforma.*, 2001, **44**, 44–56.
- [171] I. W. Davis and D. Baker, RosettaLigand Docking with Full Ligand and Receptor Flexibility, *J. Mol. Biol.*, 2009, **385**, 381–392.
- [172] J. Apostolakis, A. Plückthun and A. Caflisch, Docking small ligands in flexible binding sites, *J. Comput. Chem.*, 1998, **19**, 21–37.
- [173] H. Claußen, C. Buning, M. Rarey and T. Lengauer, FlexE: efficient molecular docking considering protein structure variations¹, *J. Mol. Biol.*, 2001, **308**, 377–395.
- [174] B. Q. Wei, L. H. Weaver, A. M. Ferrari, B. W. Matthews and B. K. Shoichet, Testing a Flexible-receptor Docking Algorithm in a Model Binding Site, *J. Mol. Biol.*, 2004, **337**, 1161–1182.
- [175] N. Huang, C. Kalyanaraman, K. Bernacki and M. Jacobson, Molecular mechanics methods for predicting protein-ligand binding, *Phys. Chem. Chem. Phys. PCCP*, 2006, **8**, 5166–77.
- [176] E. Meng, B. K. Shoichet, I.D. Kuntz, Automated docking with grid-based energy evaluation, *J. Comput. Chem.*, 1992, **13**, 505–524
- [177] H.-J. Böhm, The development of a simple empirical scoring function to estimate the binding constant for a protein-ligand complex of known three-dimensional structure, *J. Comput. Aided Mol. Des.*, 1994, **8**, 243–256.
- [178] R. Wang, L. Lai and S. Wang, Further development and validation of empirical scoring functions for structure-based binding affinity prediction, *J. Comput. Aided Mol. Des.*, 2002, **16**, 11–26.
- [179] R. A. Friesner, J. L. Banks, R. B. Murphy, T. A. Halgren, J. J. Klicic, D. T. Mainz, M. P. Repasky, E. H. Knoll, M. Shelley, J. K. Perry, D. E. Shaw and P. S. Shenkin, Glide: A New Approach for Rapid, Accurate Docking and Scoring. 1. Method and Assessment of Docking Accuracy, *J. Med. Chem.*, 2004, **47**, 1739–1749.

- [180] T. A. Halgren, R. B. Murphy, R. A. Friesner, H. S. Beard, L. L. Frye, W. T. Pollard and J. L. Banks, Glide: A New Approach for Rapid, Accurate Docking and Scoring. 2. Enrichment Factors in Database Screening, *J. Med. Chem.*, 2004, **47**, 1750–1759.
- [181] M. D. Eldridge, C. W. Murray, T. R. Auton, G. V. Paolini and R. P. Mee, Empirical scoring functions: I. The development of a fast empirical scoring function to estimate the binding affinity of ligands in receptor complexes, *J. Comput. Aided Mol. Des.*, 1997, **11**, 425–445.
- [182] S. Raub, A. Steffen, A. Kämper and C. M. Marian, AIScore - Chemically Diverse Empirical Scoring Function Employing Quantum Chemical Binding Energies of Hydrogen-Bonded Complexes, *J. Chem. Inf. Model.*, 2008, **48**, 1492–1510.
- [183] S. Tanaka and H. A. Scheraga, Medium- and Long-Range Interaction Parameters between Amino Acids for Predicting Three-Dimensional Structures of Proteins, *Macromolecules*, 1976, **9**, 945–950.
- [184] S.-Y. Huang and X. Zou, in *Annual Reports in Computational Chemistry*, ed. R. A. Wheeler, Elsevier, 2010, **6**, 280–296.
- [185] P. D. Thomas and K. A. Dill, Statistical Potentials Extracted From Protein Structures: How Accurate Are They?, *J. Mol. Biol.*, 1996, **257**, 457–469.
- [186] W. Sherman, T. Day, M. P. Jacobson, R. A. Friesner and R. Farid, Novel Procedure for Modeling Ligand/Receptor Induced Fit Effects, *J. Med. Chem.*, 2006, **49**, 534–553.
- [187] Maestro, Schrödinger Release 2021-1, 2021, LLC, NY.
- [188] M. P. Jacobson, D. L. Pincus, C. S. Rapp, T. J. F. Day, B. Honig, D. E. Shaw and R. A. Friesner, A hierarchical approach to all-atom protein loop prediction, *Proteins*, 2004, **55**, 351–367.
- [189] M. P. Jacobson, R. A. Friesner, Z. Xiang and B. Honig, On the Role of the Crystal Environment in Determining Protein Side-chain Conformations, *J. Mol. Biol.*, 2002, **320**, 597–608.
- [190] W. C. Still, A. Tempczyk, R. C. Hawley and T. Hendrickson, Semianalytical treatment of solvation for molecular mechanics and dynamics, *J. Am. Chem. Soc.*, 1990, **112**, 6127–6129.
- [191] M. Born and R. Oppenheimer, Zur Quantentheorie der Molekeln, *Ann. Phys.*, 1927, **389**, 457–484.
- [192] V. I. Minkin, Glossary of terms used in theoretical organic chemistry, *Pure Appl. Chem.*, 1999, **71**, 1919–1981.
- [193] J. C. Slater, A Generalized Self-Consistent Field Method, *Phys. Rev.*, 1953, **91**, 528–530.
- [194] P. Hohenberg and W. Kohn, Inhomogeneous Electron Gas, *Phys. Rev.*, 1964, **136**, B864–B871.
- [195] W. Kohn and L. J. Sham, Self-Consistent Equations Including Exchange and Correlation Effects, *Phys. Rev.*, 1965, **140**, A1133–A1138.
- [196] I. N. Levine, *Quantum chemistry*, Pearson, Boston, Seventh edition., 2014.
- [197] A. D. Becke, Perspective: Fifty years of density-functional theory in chemical physics, *J. Chem. Phys.*, 2014, **140**, 18A301.
- [198] J. P. Perdew, in *AIP Conference Proceedings*, AIP, Antwerp (Belgium), 2001, vol. 577, pp. 1–20.
- [199] M. Jäger, L. Freitag and L. González, Using computational chemistry to design Ru photosensitizers with directional charge transfer, *Coord. Chem. Rev.*, 2015, **304–305**, 146–165.
- [200] J. P. Perdew and W. Yue, Accurate and simple density functional for the electronic exchange energy: Generalized gradient approximation, *Phys. Rev. B*, 1986, **33**, 8800–8802.
- [201] D. C. Langreth and M. J. Mehl, Beyond the local-density approximation in calculations of ground-state electronic properties, *Phys. Rev. B*, 1983, **28**, 1809–1834.

- [202] A. D. Becke, Density-functional thermochemistry. III. The role of exact exchange, *J. Chem. Phys.*, 1993, **98**, 5648–5652.
- [203] C. Adamo and V. Barone, Toward reliable density functional methods without adjustable parameters: The PBE0 model, *J. Chem. Phys.*, 1999, **110**, 6158–6170.
- [204] C. Lee, W. Yang and R. G. Parr, Development of the Colle-Salvetti correlation-energy formula into a functional of the electron density, *Phys. Rev. B*, 1988, **37**, 785–789.
- [205] A. D. Becke, Density-functional exchange-energy approximation with correct asymptotic behavior, *Phys. Rev. A*, 1988, **38**, 3098–3100.
- [206] I. C. Gerber and J. G. Ángyán, Hybrid functional with separated range, *Chem. Phys. Lett.*, 2005, **415**, 100–105.
- [207] T. Yanai, D. P. Tew and N. C. Handy, A new hybrid exchange–correlation functional using the Coulomb-attenuating method (CAM-B3LYP), *Chem. Phys. Lett.*, 2004, **393**, 51–57.
- [208] E. Brémond and C. Adamo, Seeking for parameter-free double-hybrid functionals: The PBE0-DH model, *J. Chem. Phys.*, 2011, **135**, 024106.
- [209] M. Alipour, On the performance of time-dependent double-hybrid density functionals for description of absorption and emission spectra of heteroaromatic compounds, *Theor. Chem. Acc.*, 2016, **135**, 67.
- [210] S. Grimme and F. Neese, Double-hybrid density functional theory for excited electronic states of molecules, *J. Chem. Phys.*, 2007, **127**, 154116.
- [211] E. Runge and E. K. U. Gross, Density-Functional Theory for Time-Dependent Systems, *Phys. Rev. Lett.*, 1984, **52**, 997–1000.
- [212] R. van Leeuwen, Mapping from Densities to Potentials in Time-Dependent Density-Functional Theory, *Phys. Rev. Lett.*, 1999, **82**, 3863–3866.
- [213] M. E. Casida in *Recent Advances in Density Functional Methods*, World scientific, 1995, **1**, 155–192.
- [214] H. Lin and D. G. Truhlar, QM/MM: what have we learned, where are we, and where do we go from here?, *Theor. Chem. Acc.*, 2007, **117**, 185–199.
- [215] R. P. Magalhães, H. S. Fernandes and S. F. Sousa, Modelling Enzymatic Mechanisms with QM/MM Approaches: Current Status and Future Challenges, *Isr. J. Chem.*, 2020, **60**, 655–666.
- [216] M. Zemmouche, C. García-Iriepa and I. Navizet, Light emission colour modulation study of oxyluciferin synthetic analogues via QM and QM/MM approaches, *Phys. Chem. Chem. Phys.*, 2020, **22**, 82–91.
- [217] S.-F. Chen, N. Ferré and Y.-J. Liu, QM/MM Study on the Light Emitters of Aequorin Chemiluminescence, Bioluminescence, and Fluorescence: A General Understanding of the Bioluminescence of Several Marine Organisms, *Chem. Eur. J.*, 2013, **19**, 8466–8472.
- [218] B. Honig and M. Karplus, Implications of torsional potential of retinal isomers for visual excitation, *Nature*, 1971, **229**, 558–560.
- [219] A. Warshel and M. Levitt, Theoretical studies of enzymic reactions: dielectric, electrostatic and steric stabilization of the carbonium ion in the reaction of lysozyme, *J. Mol. Biol.*, 1976, **103**, 227–249.
- [220] M. Karplus, Development of Multiscale Models for Complex Chemical Systems: From H+H2 to Biomolecules (Nobel Lecture), *Angew. Chem. Int. Ed.*, 2014, **53**, 9992–10005.

- [221] T. S. Hofer and A. O. Tirlir, Combining 2d-Periodic Quantum Chemistry with Molecular Force Fields: A Novel QM/MM Procedure for the Treatment of Solid-State Surfaces and Interfaces, *J. Chem. Theory Comput.*, 2015, **11**, 5873–5887.
- [222] F. R. S. Purtscher and T. S. Hofer, Probing the range of applicability of structure- and energy-adjusted QM/MM link bonds III: QM/MM MD simulations of solid-state systems at the example of layered carbon structures, *J. Comput. Chem.*, 2024, **45**, 2186–2197.
- [223] J. M. Boereboom, P. Fleurat-Lessard and R. E. Bulo, Explicit Solvation Matters: Performance of QM/MM Solvation Models in Nucleophilic Addition, *J. Chem. Theory Comput.*, 2018, **14**, 1841–1852.
- [224] C. E. Tzeliou, M. A. Mermigki and D. Tzeli, Review on the QM/MM Methodologies and Their Application to Metalloproteins, *Molecules*, 2022, **27**, 2660.
- [225] S. A. Barr, G. S. Kedziora, A. M. Ecker, J. C. Moller, R. J. Berry and T. D. Breitzman, Bond breaking in epoxy systems: A combined QM/MM approach, *J. Chem. Phys.*, 2016, **144**, 244904.
- [226] U. Ryde, in *Methods in Enzymology*, ed. G. A. Voth, Academic Press, 2016, **577**, 119–158.
- [227] R. Zhang, B. Lev, J. E. Cuervo, S. Y. Noskov and D. R. Salahub, in *Advances in Quantum Chemistry*, eds. J. R. Sabin and E. Brändas, Academic Press, 2010, **59**, 353–400.
- [228] C. N. Rowley and B. Roux, The Solvation Structure of Na⁺ and K⁺ in Liquid Water Determined from High Level *ab Initio* Molecular Dynamics Simulations, *J. Chem. Theory Comput.*, 2012, **8**, 3526–3535.
- [229] L. W. Chung, W. M. C. Sameera, R. Ramozzi, A. J. Page, M. Hatanaka, G. P. Petrova, T. V. Harris, X. Li, Z. Ke, F. Liu, H.-B. Li, L. Ding and K. Morokuma, The ONIOM Method and Its Applications, *Chem. Rev.*, 2015, **115**, 5678–5796.
- [230] F. Maseras and K. Morokuma, IMOMM: A new integrated *ab initio* + molecular mechanics geometry optimization scheme of equilibrium structures and transition states, *J. Comput. Chem.*, 1995, **16**, 1170–1179.
- [231] S. Humbel, S. Sieber and K. Morokuma, The IMOMO method: Integration of different levels of molecular orbital approximations for geometry optimization of large systems: Test for n-butane conformation and SN2 reaction: RCl+Cl⁻, *J. Chem. Phys.*, 1996, **105**, 1959–1967.
- [232] G. Madhavi Sastry, M. Adzhigirey, T. Day, R. Annabhimoju and W. Sherman, Protein and ligand preparation: parameters, protocols, and influence on virtual screening enrichments, *J. Comput. Aided Mol. Des.*, 2013, **27**, 221–234.
- [233] M. H. M. Olsson, C. R. Søndergaard, M. Rostkowski and J. H. Jensen, PROPKA3: Consistent Treatment of Internal and Surface Residues in Empirical pKa Predictions, *J. Chem. Theory Comput.*, 2011, **7**, 525–537.
- [234] J. Tomasi, B. Mennucci and R. Cammi, Quantum Mechanical Continuum Solvation Models, *Chem. Rev.*, 2005, **105**, 2999–3094.
- [235] V. Barone and M. Cossi, Quantum Calculation of Molecular Energies and Energy Gradients in Solution by a Conductor Solvent Model, *J. Phys. Chem. A*, 1998, **102**, 1995–2001.
- [236] P. C. Hariharan and J. A. Pople, The influence of polarization functions on molecular orbital hydrogenation energies, *Theor. Chim. Acta*, 1973, **28**, 213–222.
- [237] R. Ditchfield, W. J. Hehre and J. A. Pople, Self-Consistent Molecular-Orbital Methods. IX. An Extended Gaussian-Type Basis for Molecular-Orbital Studies of Organic Molecules, *J. Chem. Phys.*, 2003, **54**, 724–728.

- [238] W. J. Hehre, R. Ditchfield and J. A. Pople, Self-Consistent Molecular Orbital Methods. XII. Further Extensions of Gaussian-Type Basis Sets for Use in Molecular Orbital Studies of Organic Molecules, *J. Chem. Phys.*, 2003, **56**, 2257–2261.
- [239] M. M. Francl, W. J. Pietro, W. J. Hehre, J. S. Binkley, M. S. Gordon, D. J. DeFrees and J. A. Pople, Self-consistent molecular orbital methods. XXIII. A polarization-type basis set for second-row elements, *J. Chem. Phys.*, 1982, **77**, 3654–3665.
- [240] W. Sherman, H. S. Beard and R. Farid, Use of an Induced Fit Receptor Structure in Virtual Screening, *Chem. Biol. Drug Des.*, 2006, **67**, 83–84.
- [241] D. K. Lokwani, A. P. Sarkate, K. S. Karnik, A. P. G. Nikalje and J. A. Seijas, Structure-Based Site of Metabolism (SOM) Prediction of Ligand for CYP3A4 Enzyme: Comparison of Glide XP and Induced Fit Docking (IFD), *Molecules*, 2020, **25**, 1622–1634.
- [242] T. M. Griffiths, A. J. Oakley and H. Yu, Atomistic Insights into Photoprotein Formation: Computational Prediction of the Properties of Coelenterazine and Oxygen Binding in Obelin, *J. Comput. Chem.*, 2020, **41**, 587–603.
- [243] T. A. Halgren, Identifying and Characterizing Binding Sites and Assessing Druggability, *J. Chem. Inf. Model.*, 2009, **49**, 377–389.
- [244] T. Halgren, New Method for Fast and Accurate Binding-site Identification and Analysis, *Chem. Biol. Drug Des.*, 2007, **69**, 146–148.
- [245] F. McCapra and Y. C. Chang, The Chemiluminescence of a Cypridina Luciferin Analogue, *Chem. Commun. (London)*, 1967, 1011–1012.
- [246] M. Sahihi, J. S. García and I. Navizet, Bioluminescent Nanoluciferase–Furimamide Complex: A Theoretical Study on Different Protonation States, *J. Phys. Chem. B*, 2020, **124**, 2539–2548.
- [247] E. Frieden and C. Walter, Prevalence and Significance of the Product Inhibition of Enzymes, *Nature*, 1963, **198**, 834–837.
- [248] Y. Liu, F. Zhang, L. Jiang, J. J. P. Perry, Z. Zhao and J. Liao, Product inhibition kinetics determinations - Substrate interaction affinity and enzymatic kinetics using one quantitative FRET assay, *Int. J. Biol. Macromol.*, 2021, **193**, 1481–1487.
- [249] Y. Xiong, Y. Zhang, Z. Li, M. S. Reza, X. Li, X. Tian and H. Ai, Engineered Amber-Emitting Nano Luciferase and Its Use for Immunobioluminescence Imaging In Vivo, *J. Am. Chem. Soc.*, 2022, **144**, 14101–14111.
- [250] T. Jiang, L. Du and M. Li, Lighting up bioluminescence with coelenterazine: strategies and applications, *Photochem. Photobiol. Sci.*, 2016, **15**, 466–480.
- [251] Y. C. Martin, Experimental and pKa prediction aspects of tautomerism of drug-like molecules, *Drug Discov. Today Technol.*, 2018, **27**, 59–64.
- [252] S. Nakai, M. Yasui, M. Nakazato, F. Iwasaki, S. Maki, H. Niwa, M. Ohashi and T. Hirano, Fundamental Studies on the Structures and Spectroscopic Properties of Imidazo[1,2-*a*]pyrazin-3(7*H*)-one Derivatives, *Bull. Chem. Soc. Jpn.*, 2003, **76**, 2361–2387.
- [253] H. Isobe, S. Yamanaka, S. Kuramitsu and K. Yamaguchi, Regulation Mechanism of Spin–Orbit Coupling in Charge-Transfer-Induced Luminescence of Imidazopyrazinone Derivatives, *J. Am. Chem. Soc.*, 2008, **130**, 132–149.
- [254] K. Usami and M. Isobe, Low-temperature photooxygenation of coelenterate luciferin analog synthesis and proof of 1,2-dioxetanone as luminescence intermediate, *Tetrahedron*, 1996, **52**, 12061–12090.

- [255] B.-W. Ding, P. Naumov and Y.-J. Liu, Mechanistic Insight into Marine Bioluminescence: Photochemistry of the Chemiexcited *Cypridina* (Sea Firefly) Lumophore, *J. Chem. Theory Comput.*, 2015, **11**, 591–599.
- [256] E. A. Jares-Erijman and T. M. Jovin, FRET imaging, *Nat. Biotechnol.*, 2003, **21**, 1387–1395.
- [257] C. E. Rowland, C. W. Brown, I. L. Medintz and J. B. Delehanty, Intracellular FRET-based probes: a review, *Methods Appl. Fluoresc.*, 2015, **3**, 042006.
- [258] B. W. Van Der Meer, in *FRET - Förster Resonance Energy Transfer*, eds. I. Medintz and N. Hildebrandt, Wiley, 1st edn., 2013, pp. 23–62.
- [259] Y. Wu and T. Jiang, Developments in FRET- and BRET-Based Biosensors, *Micromachines*, 2022, **13**, 1789.
- [260] P. Rajdev and S. Ghosh, Fluorescence Resonance Energy Transfer (FRET): A Powerful Tool for Probing Amphiphilic Polymer Aggregates and Supramolecular Polymers, *J. Phys. Chem. B*, 2019, **123**, 327–342.
- [261] M. Imani, N. Mohajeri, M. Rastegar and N. Zarghami, Recent advances in FRET-Based biosensors for biomedical applications, *Anal. Biochem.*, 2021, **630**, 114323.
- [262] T. Förster, Zwischenmolekulare Energiewanderung und Fluoreszenz, *Ann. Phys.*, 1948, **437**, 55–75.
- [263] S. A. Latt, H. T. Cheung and E. R. Blout, Energy Transfer. A System with Relatively Fixed Donor-Acceptor Separation, *J. Am. Chem. Soc.*, 1965, **87**, 995–1003.
- [264] D. Shrestha, A. Jenei, P. Nagy, G. Vereb and J. Szöllősi, Understanding FRET as a Research Tool for Cellular Studies, *Int. J. Mol. Sci.*, 2015, **16**, 6718–6756.
- [265] L. Stryer, Fluorescence Energy Transfer as a Spectroscopic Ruler, *Annu. Rev. Biochem.*, 1978, **47**, 819–846.
- [266] O. Shimomura, Cause of spectral variation in the luminescence of semisynthetic aequorins, *Biochem. J.*, 1995, **306**, 537–543.
- [267] M. J. Frisch, G. W. Trucks, H. B. Schlegel, G. E. Scuseria, M. A. Robb, J. R. Cheeseman, G. Scalmani, V. Barone, G. A. Petersson, H. Nakatsuji, X. Li, M. Caricato, A. V. Marenich, J. Bloino, B. G. Janesko, R. Gomperts, B. Mennucci, H. P. Hratchian, J. V. Ortiz, A. F. Izmaylov, J. L. Sonnenberg, D. Williams-Young, F. Ding, F. Lipparini, F. Egidi, J. Goings, B. Peng, A. Petrone, T. Henderson, D. Ranasinghe, V. G. Zakrzewski, J. Gao, N. Rega, G. Zheng, W. Liang, M. Hada, M. Ehara, K. Toyota, R. Fukuda, J. Hasegawa, M. Ishida, T. Nakajima, Y. Honda, O. Kitao, H. Nakai, T. Vreven, K. Throssell, J. A. Montgomery, Jr., J. E. Peralta, F. Ogliaro, M. J. Bearpark, J. J. Heyd, E. N. Brothers, K. N. Kudin, V. N. Staroverov, T. A. Keith, R. Kobayashi, J. Normand, K. Raghavachari, A. P. Rendell, J. C. Burant, S. S. Iyengar, J. Tomasi, M. Cossi, J. M. Millam, M. Klene, C. Adamo, R. Cammi, J. W. Ochterski, R. L. Martin, K. Morokuma, O. Farkas, J. B. Foresman, and D. J. Fox, Gaussian, Inc., Wallingford CT, Gaussian 16, Revision C.01, 2016.
- [268] R. Farid, T. Day, R. A. Friesner and R. A. Pearlstein, New insights about HERG blockade obtained from protein modeling, potential energy mapping, and docking studies, *Bioorg. Med. Chem.*, 2006, **14**, 3160–3173.
- [269] E. Lindahl, M. J. Abraham, B. Hess and D. van der Spoel, GROMACS 2020 Manual , Version 2020, Zenodo.
- [270] J. A. Lemkul, From Proteins to Perturbed Hamiltonians: A Suite of Tutorials for the GROMACS-2018 Molecular Simulation Package [Article v1.0], *Living J. Comput. Mol. Sci.*, 2019, **1**, 5068–5068.

- [271] K. Lindorff-Larsen, S. Piana, K. Palmo, P. Maragakis, J. L. Klepeis, R. O. Dror and D. E. Shaw, Improved side-chain torsion potentials for the Amber ff99SB protein force field, *Proteins*, 2010, **78**, 1950–1958.
- [272] A. W. Sousa da Silva and W. F. Vranken, ACPYPE - AnteChamber PYthon Parser interface, *BMC Res. Notes*, 2012, **5**, 367.
- [273] J. Wang, R. M. Wolf, J. W. Caldwell, P. A. Kollman and D. A. Case, Development and testing of a general amber force field, *J. Comput. Chem.*, 2004, **25**, 1157–1174.
- [274] M. Parrinello and A. Rahman, Polymorphic transitions in single crystals: A new molecular dynamics method, *J. Appl. Phys.*, 1981, **52**, 7182–7190.
- [275] H. J. C. Berendsen, J. P. M. Postma, W. F. van Gunsteren, A. DiNola and J. R. Haak, Molecular dynamics with coupling to an external bath, *J. Chem. Phys.*, 1984, **81**, 3684–3690.
- [276] T. Darden, D. York and L. Pedersen, Particle mesh Ewald: An $N \cdot \log(N)$ method for Ewald sums in large systems, *J. Chem. Phys.*, 1993, **98**, 10089–10092.
- [277] B. Hess, H. Bekker, H. J. C. Berendsen and J. G. E. M. Fraaije, LINCS: A linear constraint solver for molecular simulations, *J. Comput. Chem.*, 1997, **18**, 1463–1472.
- [278] C. Peng and H. Bernhard Schlegel, Combining Synchronous Transit and Quasi-Newton Methods to Find Transition States, *Isr. J. Chem.*, 1993, **33**, 449–454.
- [279] L. Deng, T. Ziegler and L. Fan, A combined density functional and intrinsic reaction coordinate study on the ground state energy surface of H₂CO, *J. Chem. Phys.*, 1993, **99**, 3823–3835.

A1. Appendix 1 - Basics of FRET photophysics

Förster-resonance-energy transfer is a distance-dependent photophysical process consisting in a radiationless energy transfer from a donor (D) to an acceptor (A) via long-range dipole-dipole coupling mechanism^{2,256,257}. The two molecular species involved must present different absorption and emission spectra, with the acceptor displaying a redshifted absorption with respect to the donor. Furthermore, the emission spectrum of D must be well overlapping with the absorption spectrum of A. Despite these specifications, the process is a non-radiative one and takes place only in a specific range of distance between the two molecular species, which goes from 1 to 10 nm, since at higher values photon emission becomes dominant and at lower ones the collision would prevail, leading to collisional quenching²⁵⁸. Depending on the spectral properties of D and A, the two must be different or not. Whenever the molecular species are identical, something which can happen only if the Stokes' shift of the molecules is small enough to permit it, the process is called homoFRET, otherwise it is termed heteroFRET. Being a strongly distance-dependent process, FRET is widely used for the investigation of conformational changes in macromolecules and for monitoring aggregation processes either in vivo or in vitro²⁵⁹⁻²⁶¹. The crucial equations in describing FRET process, whose Jablonski diagram is depicted in Figure A1.1, are the ones proposed in 1948 by Theodore Förster²⁶², whose theoretical results were demonstrated only in the late sixties²⁶³.

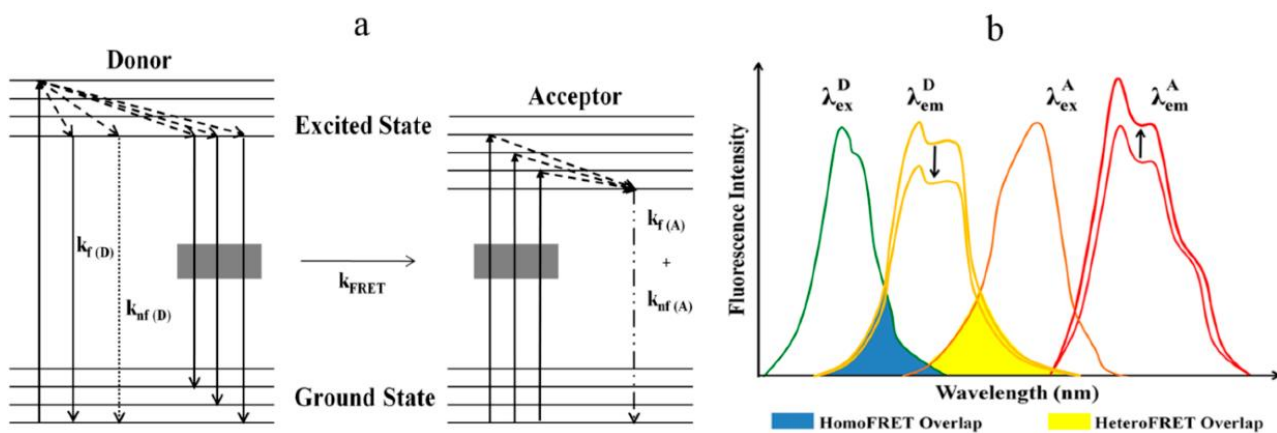


Figure A1. 1 - Jablonski diagram of FRET process (a) and representation of spectral overlap (b). First, a donor molecule absorbs energy, reaching an excited state, then, the return to the ground state can be either a radiative fluorescent process ($k_f(D)$ being its kinetic constant), a non-radiative internal conversion ($k_{nr}(D)$) or the non-radiative FRET (k_{FRET}) where energy is transferred to an acceptor, which reaches the corresponding excited state and then decays to the ground state with the emission of light (or with non-radiative pathways). The FRET mechanism is possible only when the spectral properties of D and A allow it, with a partial overlap of the D emission spectrum and A absorption one. Image taken from reference²⁶⁴.

The dependence of the efficiency of the energy transfer on the distance between D and A is described by this relation:

$$E = \frac{R_0^6}{R^6 + R_0^6} \quad (A1.1)$$

Where R is the distance between the two involved molecular species and R_0 is the Forster distance, a characteristic value indicating the distance at which the probability of having the transfer is 50%, which is usually in the range of 4-8 nm. This distance can be recast in function of the dipole orientation and spectroscopic properties of D and A, in particular it is possible to write R_0 as:

$$R_0 = 0.02108 [k^2 \Phi_D n^{-4}]^{\frac{1}{6}} \quad (A1.2)$$

Where k^2 represents an orientation factor related to the dipoles of D emission and A absorption and depends on the angle between D and A (θ_R) and the ones between each dipole and the segment joining the dipole origins (θ_D and θ_A) according to the equation:

$$k^2 = (\cos(\theta_R) - 3 \cos(\theta_D) \cos(\theta_A))^2 \quad (\text{A1.3})$$

The other significant term in equation A1.2 is J, which is the spectral overlap term, defined as:

$$J = \int_0^{\infty} F_D(\lambda) \varepsilon_A(\lambda) \lambda^4 d\lambda \quad (\text{A1.4})$$

Where λ is the wavelength, $\varepsilon_A(\lambda)$ is the molar absorption coefficient of the acceptor and $F_D(\lambda)$ is the normalized fluorescence emission of the donor, both depending on λ . Looking at these equations, some key features of donor and acceptor to have a highly efficient FRET process become evident. First of all, the overlap between the donor fluorescence spectrum and the acceptor absorption spectrum, whose crucial role emerges from equation A1.4. Another significant feature accounting for the efficiency of the energy transfer is the quantum yield of D: the higher, the better. Last but not least, the orientation of the interacting dipoles should be favourable, which means that they should ideally be parallel, to maximize the orientation factor k^2 , ranging from 0 to 4. For most biological applications based on organic molecules as D and A, k^2 value is assumed to be 2/3, which is an isotropic dynamic averaging, based on the hypothesis that both molecules are free to rotate during the D excited state lifetime.

Since the efficiency of FRET is strongly affected by distance between D and A (see equation A1.1), this phenomenon can be used as a “spectroscopic ruler”, as stated by Stryer²⁶⁵. Anyway, this possibility is strongly related to the distance range object of investigation, since for R below $0.5 R_0$ the efficiency quickly raises to 1 and above $2 R_0$ it rapidly tends to 0²⁶⁴. This means that for distances outside the specified range, it is only possible to state whether R is smaller than $0.5 R_0$ or bigger than $2 R_0$. Interestingly, biological macromolecules are usually in the dimension range which characterizes FRET, making it a possible solution to overcome the limitations of optical microscopes, allowing the prediction of molecular species interactions or proximity in a range which is far shorter than the typical ~250nm lateral diffraction of optical instruments.

A2. Appendix 2 - Computational details

A2.1 Modeling of NanoLuc-furimazine and NanoLuc-furimamide complexes

DFT benchmark:

The aim was to select, within the Density Functional Theory framework, which functional was able to provide the most reliable results among the following: the range separated CAM-B3LYP²⁰⁷, that provides a good description of charge transfer (CT) excitations, and the global hybrid ones B3LYP^{202,204} and PBE0²⁰³. Since no experimental data is available for furimazine nor furimamide, calculation of the absorption spectra was performed on coelenterazine, whose corresponding experimental data is available²⁶⁶. Geometry optimization in methanol and absorption spectra prediction were carried out with each of the three aforementioned functionals and results are reported in Figure A2.1.

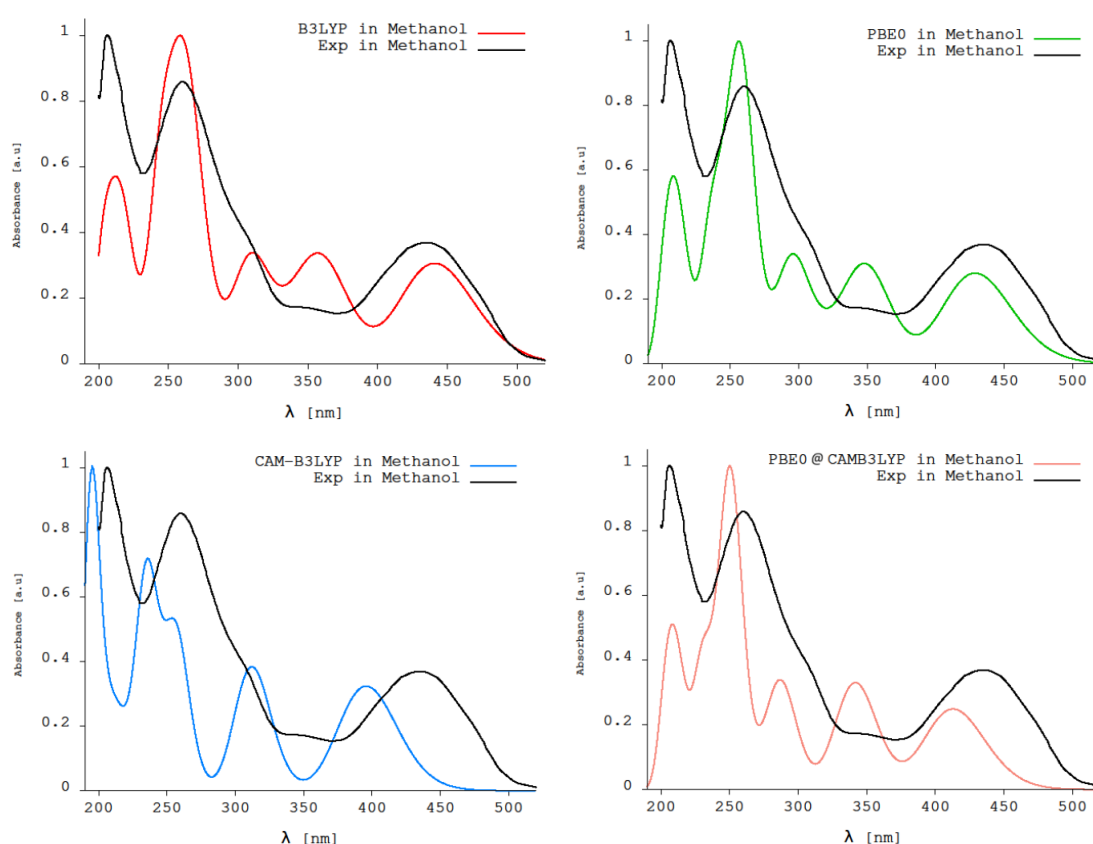


Figure A2. 1 - Calculated absorption spectra of coelenterazine in methanol using B3LYP/6-31+G(d), PBE0/6-31+G(d), and CAM-B3LYP/6-31+G(d) levels of theory after geometry optimization, and PBE0/6-31+G(d) on the CAM-B3LYP/6-31+G(d) optimized geometry. Each calculated spectrum is compared with the experimental one.

Spectra analysis shows that CAM-B3LYP best performs in reproducing the spectral shape, despite giving transition energies blue-shifted by ~0.3 eV, while PBE0 better predicts the energy of the transitions.

On the basis of these results, all geometry optimizations of the investigated systems were performed by using the CAM-B3LYP functional, and single point calculations with the PBE0 functional were used for the final emission spectra to partially correct the blue-shift which characterizes CAM-B3LYP results.

QM calculations:

QM geometry optimizations have been performed with the Gaussian16 package²⁶⁷ in the framework of DFT for GS and TD-DFT for ES. Basing on what was found in the literature and on the results of the performed benchmark, the range-separated hybrid functional CAM-B3LYP²⁰⁷ and the 6-31+G(d)^{236–239} basis set, were used, if not differently specified, for all optimizations. Orbitals involved in the electronic transitions were visualized with GaussView 6.1.1. Solvent effects were taken into account by means of the polarizable continuum model implemented according to the CPCM scheme^{234,235}. For all excited states' optimizations and spectra prediction in solution, linear response formalism was used.

Docking calculations:

The docking between NanoLuc and the keto and enol forms of furimazine was performed with the IFD^{186,240,268} module included in Maestro, which can account for the flexibility of both the ligand and the protein residues' sidechains for their rotatable bonds. The IFD procedure was carried out with the following parameters: 2.5 kcal/mol as energy window for the sampling of ring conformations, penalization for non-planar amide conformations, van der Waals scaling of 0.5 for the enzyme and the substrate atoms radius and optimization of sidechains closer than 0.5 nm from the substrate. The scoring parameter adopted for the calculated poses is the Glide Docking Score^{179,179,180}. Starting from these results, the docking procedure has been repeated by reducing the dimension of the box in which the centre of mass of the ligand was located after the first run, to get the final result.

MD Simulations:

All MD simulations were carried out with the GROMACS 2020.6^{269,270} package. The force field used is the AMBER ff99SB-ILDN²⁷¹, shown to behave well with stable proteins, as is the case for NanoLuc¹²¹; force field parameters for the substrates, all furimazine and furimamide forms, were all taken from the ACpype webserver²⁷², based on the General Amber Force Field (GAFF)²⁷³. In all the simulations, either with NanoLuc alone or NanoLuc plus ligand, the system was placed at the centre of a dodecahedral box, solvated with explicit water molecules (TIP3P) with a minimum distance of 1 nm between the solute and the box, neutrality was reached through the addition of appropriate amounts of sodium ions (Na⁺) and periodic boundary conditions in all directions were applied.

An energy minimization was performed and was followed by equilibration, first in the NVT and then in the NPT ensemble, and then by the effective 100 ns MD run, which was carried out at 1 bar and 300 K, controlled respectively by a Parrinello-Rahman barostat²⁷⁴ and a modified Berendsen thermostat²⁷⁵. The cutoffs for van der Waals (vdW) and Coulomb interactions were set at 1 nm. Particle Mesh Ewald (PME)²⁷⁶ was used for long-range electrostatics. The integration of the equations of motion was done with the leap-frog algorithm with a timestep of 2 fs and the constraints on H-bonds were applied exploiting the LINCS algorithm²⁷⁷. Coordinates were recorded to the trajectory file every 0.1 ps. Clustering analysis of the MD trajectories was performed with the GMX Cluster module of GROMACS²⁶⁹ to extract the most representative structures. The clustering algorithm employed is the GROMOS one¹⁵⁷, with a cutoff set at 0.145 nm for all the simulations, except for the case of NanoLuc alone where it was set at 0.2 nm. *Gmx energy* tool was employed to evaluate, at each frame of the simulations, the interaction energies between ligands and NanoLuc or ligands and NanoLuc and water molecules surrounding the ligands. In addition, the *gmx_MMPBSA* package⁴⁵ was also used to provide an estimate of binding free energies according to the MM/GBSA method^{46,47}, which should provide a more correct ranking of ligands⁴⁸. The 100 ns simulation was then supplemented by two other replicas of 90 ns each, yielding results consistent with the first ones in terms of energies and type of interactions (notably H-bonds), for all the considered species, so, for conciseness only the results of the first MD simulation have been reported.

QM/MM calculations:

QM/MM calculations on the protein-ligand pair were performed by using the ONIOM method²²⁹ as implemented in the Gaussian16 package. The model system, represented by the ligand, was described at the TD-DFT level with the CAM-B3LYP functional and the 6-31+G(d) basis set, while the MM portion, the protein and the water molecules, was modelled via the AMBER force field. All these calculations were performed with the inclusion of the MM charges from the whole MM region in the TD-DFT calculations on the model system. For the NanoLuc-furimamide (FurA, FurC, and FurD) systems obtained from the docking/MD protocol, the ES geometries of the ligands were fully optimized, keeping the MM part frozen. The QM/MM optimizations were followed by single-point calculations with the same settings, except for the employed functional, with PBE0 chosen to correct the CAM-B3LYP blue-shift. Electronic spectra were simulated by broadening the vertical excitation energies with a gaussian function of HWHM = 0.19 eV, extracted from the experimental spectra.

A2.2 Furimazine and furimamide characterization

QM calculations:

QM geometry optimizations have been performed with the Gaussian16 package²⁶⁷ in the framework of DFT for GS and of its time-dependent formulation for ES. The range-separated hybrid functional CAM-B3LYP²⁰⁷ and the 6-31+G(d)^{236–239} basis set were employed for all optimizations. UV/Vis final spectra predictions have been carried out by performing a single point PBE0 calculation with the same basis set on the CAM-B3LYP optimized geometry as done in the previous work, while NMR spectra were calculated at B3LYP/6-311+G(d,2p) level of theory. Solvent effects were taken into account by means of the CPCM model^{234,235}. For all excited states' optimizations and spectra prediction in solution, linear response formalism was used.

Experimental details:

The UV/Vis absorption spectrum of furimazine in DMSO was collected with a Cary 100 UV-Vis spectrometer, solution concentration being 2.62×10^{-3} M. The solution was placed in standard quartz glass cuvettes of 1 mm optical path length. For DCM results, the cuvette was a 10 mm optical path one, and absorption spectrum was collected with a Jasco V-570 spectrometer, solution concentration being 6.55×10^{-5} M. The NMR results were collected with a Bruker NMR Avance 400 NEO.

Furimamide absorption and emission spectra in acetonitrile were collected with a Cary 100 UV/Vis spectrometer, solution concentration being 1.3535×10^{-5} M. The solution was placed in standard quartz glass cuvettes of 1 mm optical path length. Evaluation of acid-base properties of furimamide were carried out with the same settings, by adding a 200-fold more concentrated solution of EtONa in EtOH to check the possible deprotonation of the chemical species.

A2.3 Modeling of furimazine oxidation mechanism in DMSO

QM geometry optimizations have been performed with the Gaussian16 package²⁶⁷ in the framework of DFT for GS and of TD-DFT for ES. The range-separated hybrid functional CAM-B3LYP²⁰⁷ and the 6-31+G(d,p)^{236–239} basis set, were employed. Solvent effects were taken into account by means of the conductor-like polarizable continuum model (CPCM)^{234,235}. The unrestricted open-shell approximation was used to treat biradical species, in which the broken symmetry technique was introduced to treat singlet-state biradical and S^2 values were used as index to confirm the biradical character of the found solution. The calculations of harmonic frequency, zero-point vibrational energy (ZVE), and thermodynamics data were carried out at the same theoretical level as the corresponding optimization. The transition state research was performed with the QST3 (Quadratic Synchronous Transit 3)²⁷⁸

method, and the results were checked with Intrinsic Reaction Coordinate (IRC)²⁷⁹ calculations, to verify that the found transition states were effectively connecting the hypothesised minima. Electronic structures, orbitals involved in the electronic transitions and emission spectra were visualized with GaussView 6.1.1.

Acknowledgements

This is the third time in my life that I've sat down to write acknowledgments, and I'm starting to run out of ways to make them original. That said, so much has changed over the last three years that I can't possibly skip this part. Finally, I get to drop the formal tone of the previous chapters and switch to something more conversational - a bit more me. Let's take it one step at a time, starting with a thank you to those who have read this far, and not just from here on out. That's especially the professors I've been lucky enough to have by my side, not only in the "authors" section of my first reference but throughout this entire three-year journey. Claudio, Ugo, Carlo - you had the misfortune of taking me on as a PhD student, and let's be honest, that's no easy task. Then, things got even more complicated when I went through some of the toughest days of my life during this period. When I was asked years ago if I'd ever consider doing a PhD, I said no. I wanted to try something different and, to be honest, reading about the horror stories of toxic work environments for PhD students made me think I'd never survive in one of those places. But then came an opportunity to work with Claudio and Ugo, whom I already knew and respected, and Carlo, with whom I had only exchanged a few words, but knew was held in high regard by professors I had worked with during my internship. And I thought, "You won't get a chance like this again. Go for it." Sure, decisions like this are usually based on scientific reasoning, and while I knew I'd reach the same conclusion on those grounds, my real concern was working with people who would allow me to contribute in the best way possible. Let's face it, my personality doesn't always make that easy. The ultimate proof that I'd made the right choice came during the worst time. I'll never forget Carlo's response when I called to tell him I wouldn't be in the office for a few days. There was no hesitation, just understanding and respect. "Take all the time you need." That same kindness and support was reflected by Claudio and Ugo during those days. I wish I'd been a better scientist over these three years to truly show my gratitude in more practical ways, but for now, you'll have to settle for these words. The work environment you created during my thesis is something I'll carry with me forever, and for that, I must also thank the people who made that environment what it was.

Anna, who knows how important she's been, unofficially running the "psychological help desk" all these years, and Fabio, now a cornerstone of the research group on this side of the Alps.

Across the Alps, first and foremost, there's Ilaria, who narrowly avoided a choreographed performance to Taylor Swift's music, purely for reasons of time and an unread e-mail I just found a few days ago, and Michele, who I was fortunate enough to share an office with throughout my stay in Paris, navigating through Outlook and HyperChem misadventures, TD, teaching duties, music, and word games. By the way, today's Cemantle word is "infection". And, of course, Feven, who will probably hate me for reminding her of that game, Irene, who made sure the word "coccolono" found its way into the Treccani dictionary and Tienhong, who I was supposed to mentor but somehow became better with Gaussian than I ever could in the span of *All Too Well*. There's also Yu, whose Ho-Oh proudly sits in my living room and will likely stay with me for a long time, and Francesca, who put up with me for months and even hosted me in Treviso. And let's not forget Pietro, the only person I know capable of using "revanchism" when talking about Franchino il Criminale's videos.

A huge thank you to Alistar, Davide, Daniele, Manolo, Zihan, Martina, Florence, and Jennifer. The 18 months in Paris were unforgettable because of all of you. And to people like Françoise, who helped me improve my very broken French, making it good enough to build a small, happy world in Paris that included volleyball matches. It was amazing how just a few words and the universal language of sports connected me with my teammates: Lucas, Hussein, Alexis, Justine, and especially Sasha, Samir, and Adèle. Thank you for making me feel like one of you from day one. Paris wouldn't have been the same without you all, and without Gaia, someone who risked hypothermia more than once just to be with me. I'm not sure she fully realizes just how much what she did for me meant.

While I was away, I was always looking for pieces of home - that home I missed so much, especially at the beginning of my Paris adventure and during the darkest times. It's made up of the incredible people around me here. Each one, in their own way, helped me get through the worst moments and made sure I enjoyed the once-in-a-lifetime experience I was living. *Mamma*, with her advice, seeing herself in me. *Papà*, always wanting to be by my side when I left and when I came back. Marco, with his late-night math questions to remind me what it's like to be the older brother. Anna, with her constant support, who never let her hard days (which I'm sure existed too) fall on me. *Nonna Lina*, who was always there to warmly welcome me home, every time. And, of course, nonna Lella, nonno Giacomo, and zia Irene, who never stopped cheering me on, even from afar.

My family was crucial, but my friends were too. I won't forget the love you showed me while I was away and how you welcomed me back. I still have the banner and the photos we took a few days later. But most of all, I hold onto the memories from these three years - moments worth treasuring. To my friends of a lifetime who started new jobs and climbed up to Serie D (Gando), got married and moved in with their partners (Luca) and started building up their adult life (Gabri), thank you for still being my constants. And to all the newer (and some not-so-new) certainties like Fede, Lisa, Isi, Sara, Nuzzo, Marta, Renée, Annina, Ale, Jacopo, Manu, thank you. I wish I will have the opportunity to cheer for you for a long time. Then, a special thanks goes to Rebe, who, despite everything, managed to find time to visit me, and Molte, who was my daily companion for more than a year, night after night...

These are, perhaps, the last acknowledgments I'll ever write, and I need to make sure there's room for everyone - literally everyone - because who I am today is thanks, even in small part, to all the people I've mentioned, and even to some who don't know me at all. Thanks, Zlatan, Anna, Jurgen, Mario, Donovan, Josep, Julio, Kylian, LeBron, Davide (look who I placed you next to), Francesca, Elvira, Emanuela, Maria Antonella, Kobe, Cristiano, Eden, Marco, Russell, Jenia, Aleksandar, Jimmy, Harry, Roger, Rafa, Novak, Michael, Kevin, DeMar, Kevin, Stephen, Kyrie, Ivan, Osmany, Damian, Andrés, Earvin, Manuel, Gregg, Phil, Denis, Marcel, Lionel, Michael, Johann, Maurizio, Stefano, José, Arrigo, Fabio, Francesco, Marshall, Fabrizio, Jacopo, Francesco, and so many, many more.

This time, despite all the names mentioned, this entire work - and this entire three-year journey - is dedicated to just one person. This is for you, *Nonno*. I still wear you proudly on my wrist almost every day, but more importantly, you're always in my mind, my heart, and in every single thing I do.

Bastava la metà, ma già lo sai.