Identification of natural product stereochemistry via calculation of ECD spectra

Riccardo Lolli

Master in Chemistry

UNIVERSITÀ DI PARMA

University of Cape Town
University of Parma
2018

Identification of natural product stereochemistry via calculation of ECD spectra

Riccardo Lolli

A dissertation submitted in fulfilment of the requirements of the degree

Master in Chemistry

UNIVERSITÀ
DI PARMA

University of Cape Town
University of Parma

Supervisor:
Dr. Gerhard A. Venter

Co-Supervisors:
Dr. Karl A. Wilkinson (posthumously)
Prof.ssa Francesca Terenziani

2018

Declaration

- *I know that plagiarism is wrong. Plagiarism is to use another's work and pretend that it is one's own.*
- *I have used the required convention for citation and referencing. Each contribution to and quotation in this assignment from the work(s) of other people has been attributed, and has been cited and referenced.*
- *This assignment is my own work.*
- *I have not allowed, and will not allow, anyone to copy my work with the intention of passing it off as his or her own work.*
- *I acknowledge that copying someone else's assignment or essay, or part of it, is wrong, and declare that this is my own work.*

02/07/2018

…… Signed by candidate

Date Riccardo Lolli

Dedication

Dedicated to my family and to Karl A. Wilkinson.

Acknowledgements

First of all, I would like to express my gratitude to Karl A.Wilkinson for having always been stimulating and encouraging regarding both the work and personal life in Cape Town. It has been an honour to work with him.

A special thanks to Dr. Gerhard A. Venter and Prof. Francesca Terenziani, for grabbing my hand during the last period of the work and for their big support in writing the thesis.

I would like to deeply thank the scientific computing research group for how they welcomed me as part of the group. In particular, Tharindu, Monde and Christopher, for the help they gave me and all the others in making daily life lighter.

I would like to thank Mrs. Deirdre Brooks for her essential help in managing the bureaucratic procedure to get to Cape Town.

I special thanks to all my friends in Cape Town for the very special moments we shared together and to all my friends in Italy who supported me these last years.

Lastly, I would like to express my biggest gratitude to my family for having always being present to encourage me in all aspects of my life.

Table of contents

Chapter 3

Chapter 4

Abstract

Most commercially available antibiotics are obtained from natural products, secondary metabolites of bacteria or other living organisms. Due to the importance of this class of compounds in medicinal chemistry and growing drug resistance, efforts to discover, characterize and isolate new or improved antibiotics are continually increasing. The assignment of the absolute configuration (AC) adopted by these compounds is a crucial aspect of the characterization step and knowledge of the stereochemistry is an important factor in deciphering the interaction of these compounds with the organism and thus, the mechanism of action. In order to assign the AC, several techniques, such as X-ray diffraction and NMR experiments as well as standard electronic spectroscopy experiments (UV-Vis, ECD, etc.) or less widespread vibrational and rotational spectroscopy experiments (VCD, ROA, etc.) can be used, often in combination. However, sophisticated synthetic strategies or difficult isolation of the natural compound often leads to a small amount of product available, making some of the previous techniques unpractical; in addition to the potential structural complexity of the molecule, this can make the experimental assignment of the AC problematic. For this reason, a computational approach, aimed at calculating observable properties of the products, generating spectra and assigning the AC through comparison between the calculated and the experimental spectra, has proven useful in many cases.

Formicamycin is a natural product, isolated from a new member of *Streptomyces* bacteria, which has shown great activity against pathogenic drug-resistant bacteria and fungi, without developing antimicrobial resistance. This dissertation shows that the chiral axis of Formicamycin can be assigned as R, through the calculation of electronic circular dichroism (ECD) spectra and comparison to the experimentally determined spectrum in methanol. ECD spectroscopy is very sensitive to the chiral environment of chromophores and can be used to distinguish between different isomers. The computational procedure has been broadly defined in previous studies and involves three general steps: 1) generation of an ensemble of structures, 2) optimization of the structures and calculation of the rotational strengths of each and 3) generation of the Boltzmann-weighted spectrum. Here, two different force fields (OPLS3 and MMFFs) were used for generating the ensemble of conformers, followed by PBE0 DFT calculations to determine the optimal geometry and finally, TDDFT calculations to compute the rotational strengths of each conformer. Furthermore, the spectra were calculated in four different solvents, using the implicit SMD method, in order to inform future studies

about “variable solvent circular dichroism”. Different conformations of a molecule can be controlled by the choice of solvent and it is hypothesised that a change in solvent will result in a “fingerprint” shift in the ECD spectra that could permit assignment of the stereochemistry. The entire process was automated using a module written in Python.

1. Introduction

1.1 Natural Products

Natural products are defined as chemical compounds produced by a living organism, or more generally any substance produced in a life cycle. Usually, natural products are divided into two classes: the primary metabolites, which are vital components of living organisms, and secondary metabolites, which are not involved in primary processes of life and hence are dispensable. Molecules in the first class are associated with essential cellular functions such as nutrient assimilation, energy production and growth/development; these kinds of metabolites are for instance carbohydrates, lipids, amino acids, nucleic acids, essential cofactors for enzymes, etc. The second class, in contrast with the primary metabolites, are not strictly necessary for the life of the organism, as for example pheromones, or competitive weapons like venoms, toxins, etc.

The pharmacological features of some natural products, especially the secondary metabolites, have inspired the medical sciences in the development of new drugs. The misuse of antibiotics over the last 50 years, which has led to a rise in antimicrobial resistance (AMR) and a rise in inefficacy of established drugs, has caused an increase in the studies of these compounds in order to discover and produce new alternative, more efficient medicines.

Since the main biological processes in the human body involves chiral moieties, the study of molecular configuration is of paramount importance in order to guarantee the efficient interaction of a substrate or drug with an enzyme or other biomolecule, in order to achieve the desired effect. Moreover, these compounds can have very complex structures and critical stereochemistry features, due to the enzymatic synthesis in the organism; this makes laboratory synthesis and characterization some of the main challenges of modern medical and organic chemistry. Furthermore, the synthesis of these products can involve several reaction steps, often the final product is obtained in very small amount and the assignment of the absolute configuration (AC) through classical techniques is hindered. In these cases, a computational approach to establish conformational preferences can be very useful.

In the next sections, the main techniques used in order to define the absolute configuration, the advantages and disadvantages of each in the field of natural products, as well as their synergistic use, will be briefly discussed. A particular focus will be placed on the electronic circular dichroism (ECD) technique, through a quantum mechanical description in chapter 3.[1]

1.2 Assignment of the Absolute Configuration

The absolute configuration describes the spatial arrangement of the atoms and functional groups in a chiral molecule. A chiral molecule is defined as a molecule not superimposable on its specular image or, in other words, a molecule not possessing an improper rotation axis, including reflection planes and a centre of inversion (Figure 1).

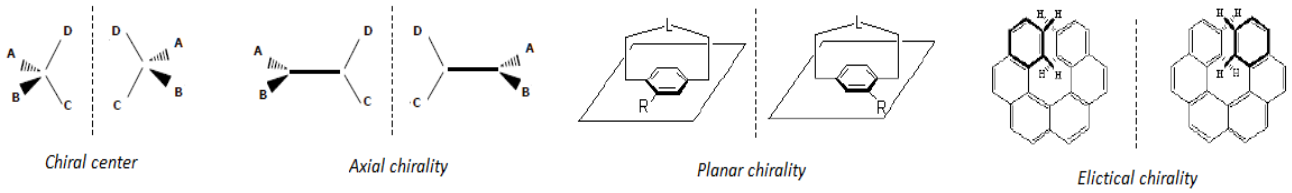


Figure 1. Different types of chirality.

The knowledge of the absolute configuration of a molecule is very important to understand biological processes (enzymatic catalysis, biological synthesis, etc.) as well as to predict the biological mechanism and activity of medicines (e.g. the R isomers of thalidomide are anti-nausea drugs, typically given to pregnant women, while the racemic, consisting of the R and S isomer, mixture of the same molecule can generate mutations in newborns).

The main techniques used to defined the absolute configuration can be classified into direct or indirect methods, depending on the need of a reference to interpret the results.[1] Each technique has its own advantages and disadvantages and can be used to deduce different information. Usually, to be able to completely define structure and configuration, a combination of techniques is used.

In the following sections, some of the most common tools are briefly presented, i.e. X-Ray diffraction, Nuclear magnetic resonance (NMR) and chiroptical techniques (see Figure 2).

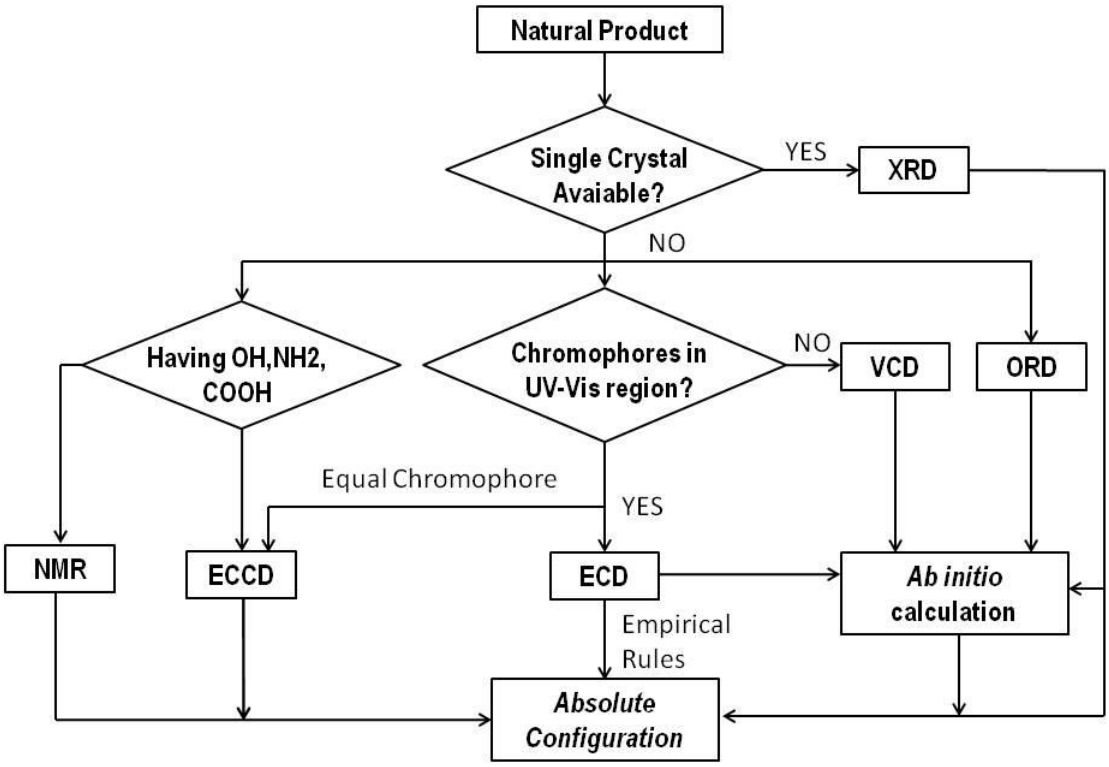


Figure 2. General approach in the assignment of the absolute configuration for natural products.

1.2.1 Single Crystal X-Ray Diffraction (XRD)[2,3]

XRD is a direct technique for obtaining chemical information such as the molecular geometry, structural data (e.g. bond and angles), and the packing of the molecules in the crystal; moreover it has the capacity to distinguish between the enantiomorphs of a chiral crystal structure and the enantiomers of a chiral molecule. All these features make this method arguably the best to assign both the structure and absolute configuration.

XRD is a passive technique (it does not involve absorption of radiation) and all the information above is obtained through analysis of the resulting scattering data of X-Ray radiation with the crystal. In particular, information about the absolute configuration is obtained by looking at a small diffraction intensity difference between the structures with opposite chirality. The main drawback of this technique is the need to have good-quality single crystals of the molecule of interest. This can be an issue for natural products, due to the complex structure and numerous steps that typically are involved in the synthesis or purification. Very often, the amount of final product obtained is of the order of

micrograms, making this technique completely impracticable. Finally, the intrinsic nature of these compounds, that often contain only light atoms, make the scattering intensity difference used to distinguish between the two enantiomers very small and the assignment of the absolute configuration is not guaranteed.[1]

1.2.2 Nuclear Magnetic Resonance (NMR)

To explain the basic mechanism of NMR spectroscopy, one has to first discuss some quantum mechanical aspects of the nuclei. Depending on the number of protons and neutrons, each nucleus has a nuclear spin quantum number (I). If the number of protons and neutrons are both even, the nucleus has no net spin and $I = 0$. However, if any one or both are odd, it will have either a half integer or integer nuclear spin quantum number, respectively. Only nuclei with $I > 0$, e.g. ^{1}H, ^{13}C or ^{31}P, are NMR-active.

If an external magnetic field is applied, the magnetic moment of the nucleus becomes aligned with the field and different spin states are generated (Zeeman splitting), depending on the value of m where $m = -I, -I+1, -1+2, \ldots, +I$ (see Figure 3).

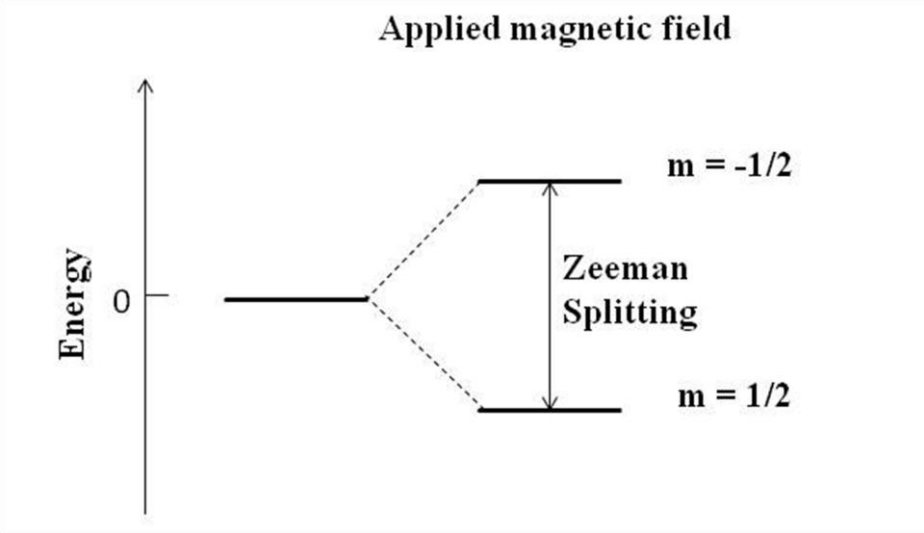


Figure 3. Generation of two spin states after the application of an external magnetic field. The energy difference between the states is defined by the Zeeman slpitting.

The frequency (ν) of the transition between these two states falls in the radiofrequency region:

$$\nu = \frac{B\gamma}{2\pi} \tag{1.1}$$

where γ is the gyromagnetic ratio (unique to each nucleus) and B is the intensity of the external magnetic field. The frequency of the transition depends on the specific environment experienced by the nucleus. In other words, the nucleus experiences a local field given by the external magnetic field and a contribution given by the specific environment (related to the electron density). Each magnetically non-equivalent nucleus will be affected by a different local magnetic field and will have a slightly different resonant frequency. Furthermore, the number of non-equivalent magnetically active nuclei within three-bond distance from a particular nucleus is responsible for the signal multiplicity, and the distance between the split peaks (in Hz) is named the coupling constant (J).[4]

By changing the experimental conditions (e.g. the sequence of pulses and delay periods in-between), it is possible to obtain different levels of information about the structure and the configuration of a molecule. In addition to the classical monodimensional ^{1}H-NMR and ^{13}C-NMR experiments, there are other more complex types of experiments that are able to give a more complete overview of the studied molecule, for example[5]:

- *COSY-90* (Correlation Spectoscopy) is a bidimensional, homo-correlated experiment. It gives information about the geminal ($^{2}J_{HH}$) and vicinal ($^{3}J_{HH}$) coupling constants.

- *HMQC* (Heteronuclear Multiple-Quantum Correlation) is a bidimensional, hetero-correlated experiment. It gives information about the coupling constants between two different spins (e.g. H and C) that are directly linked ($^{1}J_{CH}$).

- *HMBC* (Heteronuclear Multiple Bond Correlation) is the same as HMQC, but it gives information about hetero-correlated nuclei through two ($^{2}J_{CH}$) or three ($^{3}J_{CH}$) bonds.

- *NOESY* (Nuclear Overhauser Effect Spectroscopy) is a bidimensional, homo-correlated experiment. It gives information about the correlation between two spins (usually proton) through space and not through bonds. This is a very useful technique to assign the absolute configuration.

The NMR approach is very popular in order to define the absolute configuration, due to the widespread availability of the instrumentation in most laboratories, the very small amount of sample needed and the passive nature of the technique. The possibility of working with liquid/solvated samples is also a big asset of this technique.

Unfortunately, if the molecule contains a large number of NMR-active centers, the analysis of the NMR spectrum can be very difficult. Moreover, for two enantiomeric structures, a particular signal could be split or shifted in relation to the exchange velocity between the two nuclei and these variations are proportional to the enantiomeric composition of the solution. In this case, without an enantiomerically pure sample as reference, the discrimination between the two forms can be a difficult task.[6]

1.2.3 Electronic Circular Dichroism (ECD)

Another way to approach a conformational study involves the use of chiroptical techniques, which are methods based on the active (absorption) or passive (dispersion) interaction of the molecules with polarized radiation. Some of these techniques include Vibrational circular dichroism (VCD), Raman optical activity (ROA), Optical rotatory dispersion (ORD) and Electronic circular dichroism (ECD).[7]

In the following, only the latter will be discussed in detail.

1.2.3.1 Phenomenological Description[8]

To understand the interaction of an ECD-active molecule with linearly polarized radiation, it is convenient to describe the latter as the combination of two circularly polarized vectors, having opposite rotation directions, the same frequency and half the intensity of the original radiation (Figure 4).

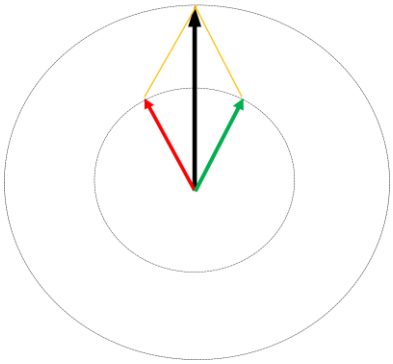


Figure 4. Vector representation of linearly polarized radiation (black) as the sum of its circularly polarized components (red and green).

A non-racemic solution, active to this technique, is able to absorb the two circularly polarized radiations in a different way and the resultant radiation will have an elliptical pattern (Figure 5).

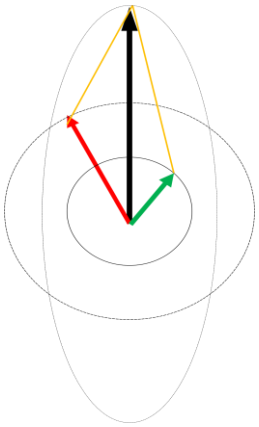


Figure 5. The classical elliptical pattern of the radiation in the CD experiment due to the different absorption of the circulrly polarized components.

The intensity of the transition is proportional to:

$$CD = A_l - A_r \tag{1.2}$$

where CD is the circular dichroism's signal intensity, A_l is the absorption of the left rotating radiation and A_r is the absorption of the right rotating radiation. Applying the Lambert-Beer equation, one can express the signal as a concentration independent quantity:

$$\Delta\epsilon = \epsilon_l - \epsilon_r = \frac{CD}{C \cdot d} \tag{1.3}$$

where ϵ_l and ϵ_r are the extinction coefficients for the left and right rotating radiations respectively, C is the concentration and d is the length of the cell. Historically, the output of the ECD instrument is expressed as molar ellipticity (Ψ, historically in $degree\ cm^2\ mol^{-1}$).[9]

$$\Psi = \frac{1}{4}(\epsilon_l - \epsilon_r)d \tag{1.4}$$

1.2.3.2 ECD Spectra

Since the resultant signal is obtained from a difference of absorptions, the signal will have a classic Cotton effect (CE, see Figure 6 for a brief explanation of this effect) pattern and it is well know that peaks in the ECD spectra of a pair of enantiomers will have opposite sign.[10] This allows for obtaining direct information about the AC.

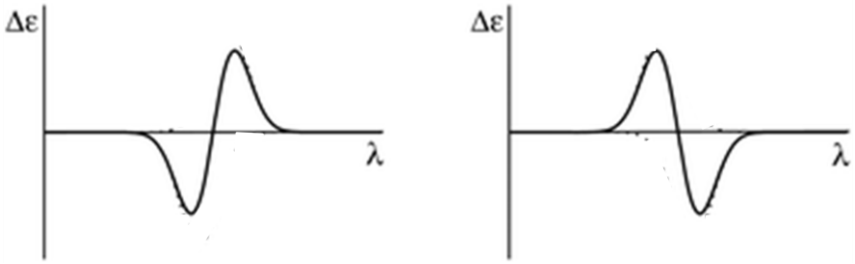


Figure 6. An example of the Cotton Effect (CE), obtained by plotting the difference of absorptions between the left and the right component of linearly polarized light (Δϵ) against the wavelength (λ). The Cotton Effect is the characteristic change in band-shape characterized by the inversion of sign at the wavelength corresponding to the maximum of absorption of the non-polarized wavelength. Moving from high to low wavelength, it is possible to define a positive (left) and negative (right) CE through the sign of the first band.

A necessary condition in order to use this technique is the presence of a chromophoric group in the chiral environment of the molecule, able to absorb the polarized radiation. The ECD signal, in fact, is strictly correlated to the UV-vis spectrum. Specifically, the zero ECD value in between a positive and a

negative peak is located at the frequency of maximum absorption in the corresponding spectrum.

A great advantage of the technique is its high sensitivity to conformational changes that generates a different intensity of the signal for different conformations. Each Cotton peak is the result of non-negligible rotational strength (R_{fg}) that depends on both the magnitude of the transition electric dipole moment (μ_{fg}) and of the magnetic transition dipole moment (m_{fg}) and the angle between them (a more detailed description of the phenomenon and these quantities are given in Chapter 3),

$$R_{fg} = \langle \mu_{fg} \rangle \cdot \langle m_{fg} \rangle \tag{1.5}$$

where g and f define, respectively, the ground state and the final state of the transition.

Often the conversion energy between two different conformations is very low (comparable to *kT*, where *k* is the Boltzmann constant and *T* the temperature) and so, in experimental conditions, the sample could be composed of a manifold of structures, each with a short lifetime. The fast nature of the technique (transitions on the order of picoseconds) makes ECD able to "recognize" all these structures and generates a spectrum that is the superimposition of the spectra of all the conformations, weighted according to their Boltzmann distribution. Different experimental conditions, such as temperature, pH and solvent, can have big effects on the final spectrum, since all these conditions can change the conformer distribution in the sample.

In order to understand ECD spectra, one can distinguish between three different cases on the basis of the number of different chromophoric groups in the molecule:

1) The presence of only one chromophoric group in the molecule that generates only one clearly defined CE signal (e.g. ketones).[7]
2) The presence of two different chromophoric groups involved in an exciton coupling system that generates a CE signal (described in the following section and in Chapter 3).
3) The presence of multiple bands when the chromophore has multiple electronic transitions in the investigated spectral region, as for example with transition metals (not discussed in this thesis).

With a molecule that has many chromophoric groups (as the one studied in this work), the final spectrum will be composed of multiple bands associated with all the different transitions. In this case, it is hard to assign each peak to its own transition and usually the AC assignment is made through comparison with a well-defined structure's spectrum.

1.2.3.3 Exciton Coupling

When two non-conjugated chromophores have strong electronic transitions of equal or similar energy and they are near one another in space, the two transitions are coupled (it is not possible to excite one of the two chromophores independently of the other). This is called "exciton coupling". As a consequence of the coupling, the two exited states split by an energy quantity, called the Davydov splitting, equal to $2V_{12}$, where V_{12} is the dipole-dipole interaction energy between the chromophores, with permanent dipole moments μ_1 and μ_2, separated by a distance r_{12}. (Figure 7).

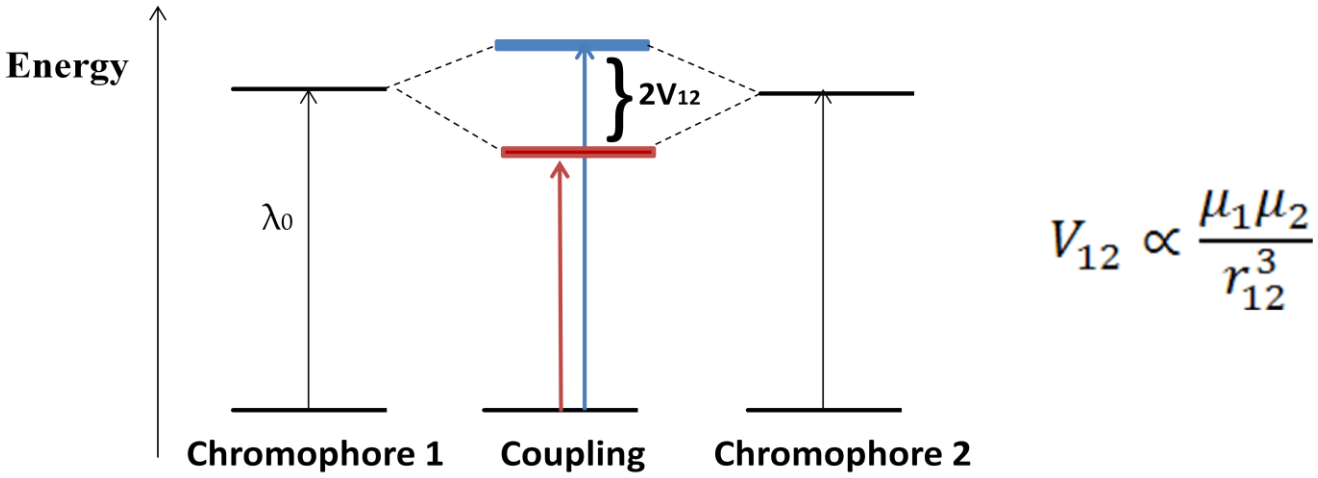


Figure 7 . Electronic transitions for the isolated chromophores ($\boldsymbol{\lambda_0}$) and for the exciton coupling system. The two transitions have energies equal to $\boldsymbol{\lambda_0 - V_{12}}$ (red) and $\boldsymbol{\lambda_0 + V_{12}}$ (blue) where $\boldsymbol{2V_{12}}$ represents the Davydov splitting.

This splitting is responsible for the appearance of two different absorption bands, in the absorption spectrum, located at shifted wavelengths (one blue-shifted, the other red-shifted) with respect to the isolated chromophore (Figure 8).

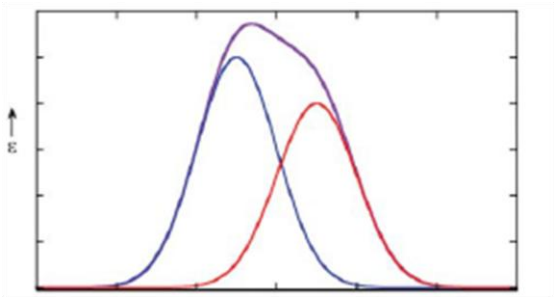
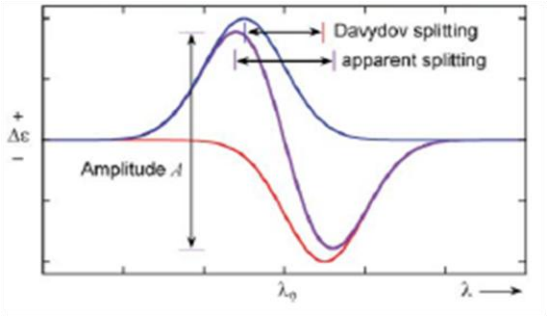


Figure 8. The absorption bands (red and blue) associated with two transitions in a coupled system at $\boldsymbol{\lambda_0 \pm V_{12}}$ *. The final absorption signal (violet) is equal to the sum of the two transitions.*

The interaction between the transition dipole moments can generate a large rotational strength. When the two electric dipoles are not coplanar, the electric dipole moment of one will generate a magnetic moment that is not orthogonal to the second and vice versa. This generates two opposite, strong rotational strengths and the resultant CD signal (called an exciton couplet) is formed by two different bands, opposite in sign, having similar areas and located at shifted wavelengths (Figure 9).



$$R_{12} \propto \pm \overrightarrow{r_{12}} \cdot \overrightarrow{\mu_1} \times \overrightarrow{\mu_2}$$

Figure 9. CD signal due to the difference in absorption between the blue and the red shifted transitions in the coupled system. Also highlighted is the Davydov splitting and the apparent splitting defined by the distance (in terms of wavelength) between the maximum and minimum absorption of the Cotton effect signal.

The intensity of the CD signal generated via exciton coupling is thus directly proportional to the product of the squared dipole moments, inversely proportional to the second power of the distance between the two dipole moments and also depends on a geometric factor ($\Omega(\alpha, \beta, \gamma)$) that takes into account the angles between the two dipoles and the distance vector in three dimensions,

$$\Delta\epsilon(\lambda) \propto V_{12}\, R_{12} \propto \pm \frac{\mu_1^2 \mu_2^2}{r_{12}^2}\, \Omega(\alpha, \beta, \gamma) \qquad (1.6)$$

In the past few decades, the method has been established as one of the most sensitive and convenient spectroscopic tools for AC analysis of chiral synthetic compounds or natural products of great structural diversity.[7,8]

1.2.3.4 Biarylic Systems

This section focuses on the direct relation between the ECD spectrum and the conformation of biarylic compounds. This class of compounds has two arylic groups linked together through a single bond. The two aromatic groups generate two different planes and the system is described in terms of one single parameter, the dihedral angle (Figure 10).

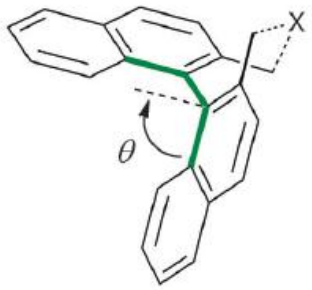


Figure 10. Simple example of a biarylic system. The dihedral angle is highlighted in green.

Depending on the experimental conditions (T, pH, solvent, etc.), the molecule can adopt different conformations having different dihedral angles. Furthermore, these systems show strong exciton coupling between the $\pi - \pi^*$ transitions of the two aromatic systems. The ECD spectrum therefore strongly depends on the adopted dihedral angle and on the strength of the exciton coupling (Figure 11).[7,11]

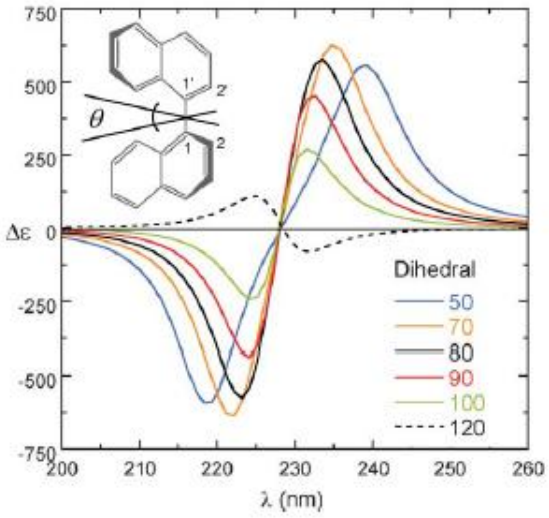


Figure 11. Relation between the dihedral angle and CD signal. Due to the nature of the signal in a coupled system, all the Cotton bands have absorbance value equal to zero at the λ_0 *value.*

All these features allow one to define not only the absolute configuration of the molecule, but also the preferred dihedral angle (and so the most stable conformation) under selected conditions. Furthermore, for flexible molecules, the final spectrum will arise from all the accessible conformations and if the system is characterized by several transitions, a typical CE peak will be associated with each chromophore as well. To avoid complexity, the "single-conformer approximation" is a common practice: this approximation assumes that the conformation having the ECD spectrum best fitting the experimental spectrum, is dominant in solution. However, one can still make an incorrect assignment, because overlap of spectra of all the conformers in solution can generate a final spectrum very similar to another conformation that is, in fact, not the main one.[7]

In conclusion, the advantages of this technique are: i) the ability to directly assign the AC of the molecule; ii) obtaining information about molecular conformation (shape of the spectrum); iii) the very high sensitivity of the technique (micrograms required) and also the high sensitivity to conformational changes. The disadvantages are related to the intrinsic nature of the molecule, which must have one or more chromophoric groups able to absorb in the UV-vis region (functionalization of the molecule with appropriate groups can also help), as well as the several factors that can influence the shape of the spectrum and make the interpretation harder and thus, assignment without a reference, difficult.

In this thesis, a computational approach has been applied with the aim of finding a significant pattern in the fingerprint area of the spectrum, which can be used to assign the configuration through comparison.

In the following section, the computational tools involved in calculating ECD spectra, are briefly introduced.

1.3 Computational Approach[2,7]

As already pointed out, the identification of the AC, through the interpretation of the CD spectrum, can be challenging for large molecules. In the same way, other techniques, such as advanced NMR, can fail in the assignment due to particular molecular symmetries in the molecule. In these cases, a computational approach is a very useful option, allowing the assignment of the AC through comparison between the experimental spectrum and the calculated one. In order to calculate chiroptical properties, several approaches have been followed in the past and different methodologies have been designed. These methodologies can involve static quantum mechanics (QM) or molecular mechanics (MM) calculations, time-dependent molecular dynamics (MD) calculations or a mixture of these.

The increase in computational power in the last twenty years, makes calculations very practicable, even for relatively large molecules. In the specific case of the calculation of an ECD spectrum, the final aim is to obtain a list of rotational strengths, R_i, at specific transition frequencies, which will then be converted into a spectrum.

Since ECD spectra depend on all the conformers present in solution, a library of conformers must be generated; all of these structures have to be geometry optimized and the corresponding rotational strength must be obtained. One can therefore define some general steps (Figure 12) of the entire process:

i) generation of a library of conformers
ii) optimization of the geometries
iii) calculation of the zero-point energy (ZPE)
iv) calculation of the rotational strengths
v) generation of the overall spectrum

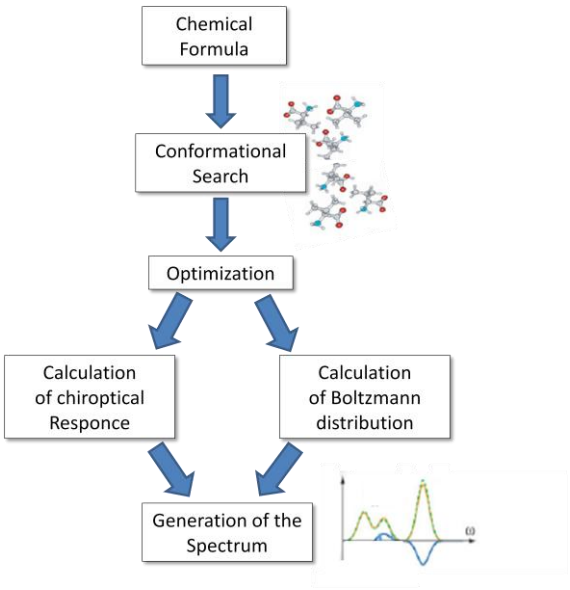


Figure 12. General flowchart of the computational procedure in order to obtain a Boltzmann-weighted CD spectrum of a natural product.

QM calculations are arguably the most accurate for conformational energies and is the only approach able to calculate the rotational strengths, but can be computationally expensive for large molecules with many conformations. These type of calculations can be divided into two main classes (see Chapter 2): *ab initio* (based on first principles using no experimental quantities) and semi-empirical (a similar framework to *ab initio* methods, but with additional simplification by approximation and with some quantities parameterized). MM methods, on the other hand, are fully parameterized and use computationally tractable mathematical functions to describe conformational energies in a fraction of the time of a QM calculation.

In order to find the best compromise between computational cost and accuracy, the different steps can be achieved using different levels of accuracy, using both MM and QM theories. In Chapter 2, a brief description of these methods will be given.

1.4 Formicamycin

African plant-ants are a family of ants that have established a protective mutualism with a guest plant. The plant offers the ants a *domatia* where they can live, while the ants offer protection to the plant

against herbivore animals. The African *Tetraponera* plant-ant is a fungus-growing insect and grows fungi inside the *domatia*. In order to keep the fungi population under control, the ants rely on actinomycete bacteria, which are able to produce antibiotics.

Researchers at the University of East Anglia discovered a new member of the *Streptomyces* bacteria family isolated from the head of the African plant-ant *Tetraponera penzigi*.[12] This new species, named *S.formicae*, has shown a great resistance against pathogenic drug-resistant bacteria and fungi, in particular *Staphylococcus aureus* (MRSA) and other vancomycin-resistant enterococci (VRE). Moreover, these bacteria do not easily develop antimicrobial resistance (AMR). The bacteria produce a class of pentacyclic structures as secondary metabolites. A previous study[12] has defined two main structures with a pharmacological interest, named Fasamycin and Formicamycin (Figure 13). In particular, the latter has shown a higher resistance against the pathogens.

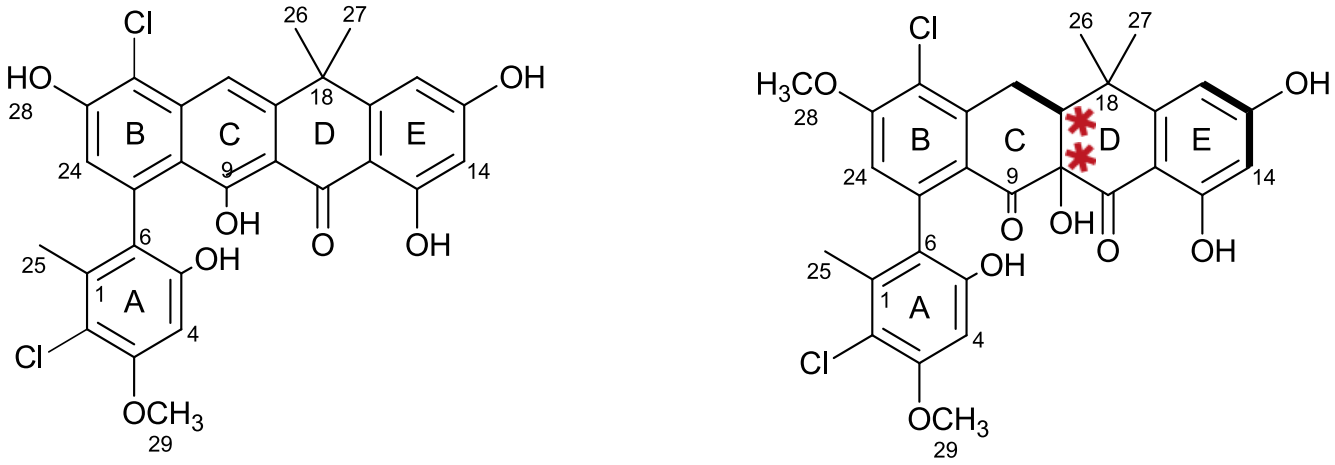


Figure 13. On the left: Chemical structure of Fasamycin. The configuration of the chiral axis between the A and B ring has been assigned as the S-configuration. On the right: Chemical structure of Formicamycin. The two chiral centers between the C and D rings have been assigned as 10R-19R.

The structures of both have been defined as in figure through ^{1}H-NMR, ^{13}C-NMR and high-resolution ESI-MS analysis. Both have a chiral axis between the A and B rings, whereas Formicamycin also has two chiral centers (C10 and C19; starred in Figure 13).

The optical activity shown by these molecules highlight a strong preference for a particular isomer. The comparison between the calculated and experimental ECD spectra of Fasamycin strongly suggests that the preferred orientation of the A ring has the hydroxyl moiety above the plane, giving an S-configuration of the chiral axis. The configurations of the two carbon chiral centers of Formicamycin

have been identified as 10R-19R through NOESY experiments and through comparison between the calculated and experimental ECD spectra. Due to the *cis* configuration of the two chiral centers, Formicamycin also presents an L-shaped scaffold (Figure 14).

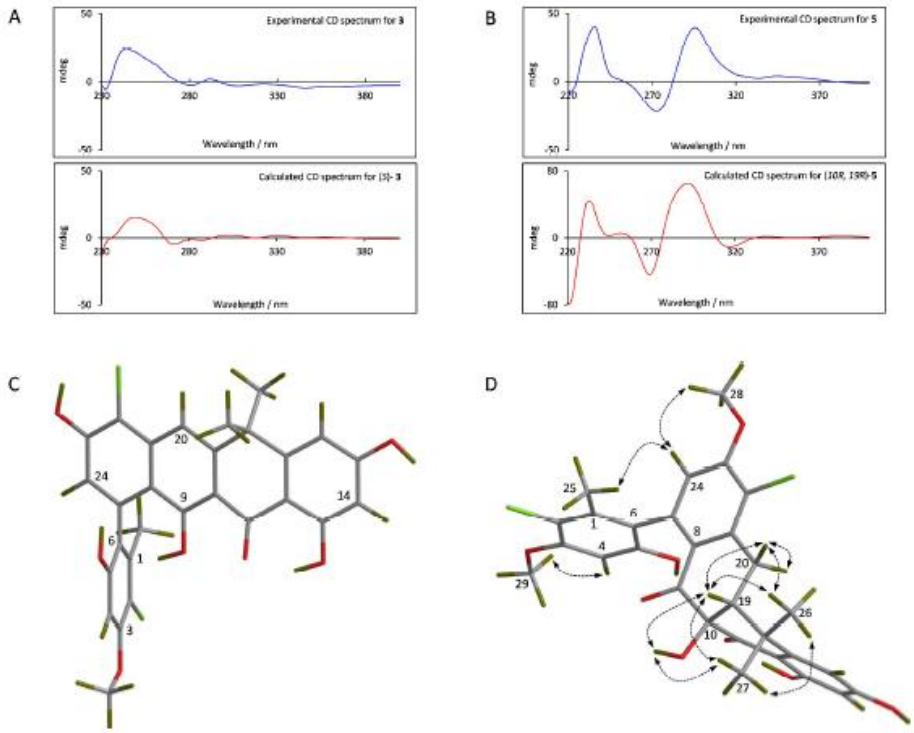


Figure 14. Experimental (blue) and calculated (red) CD spectra of Fasamycin (A) and Formicamycin (B). C and D show the 3D structures of the two, respectively.[12]

Nevertheless, the assignment of the configuration of the C6-C7 axis for Formicamycin has not been done. Further biological experiments on *S. formicae* gene expression indicates a biosynthetic mechanism for Formicamycin, where Fasamycin plays the role of intermediary. This suggests that the configuration of the chiral axis does not change between the two structures, so that Formicamycin should also have an S-configuration. A detailed study about the biosynthetic mechanism is still in progress.[12] The present study is focused on the identification of the configuration of the C6-C7 axis of Formicamycin.

1.6 Aims

The main aims of this work are:

1. *Development of an optimized computational methodology to automate the calculation of ECD spectra.*

   Normally, the calculation of ECD spectra is achieved by the generation of a library of conformers, geometry optimization and calculation of the vibrational frequencies for each structure, calculation of rotational strengths through a TDDFT calculation and finally the generation of the spectra. In this work, combined QM and MM methods will be tested in order to obtain the best compromise between accuracy of the results and computational cost of the process. A completely automated process, to manage the steps outlined above, will be developed using the Python scripting language.

2. *Refinement of the stereochemistry of Formicamycin.*

   A recent study[12] has used the methods outlined above to identify the absolute stereochemistry of the natural product Formicamycin. However, the stereochemistry about the chiral axis between the A and B rings of Formicamycin is unresolved. This project aims to elucidate the conformation about the chiral axis through comparison of the calculated ECD spectrum with the experimental spectrum, in methanol. Furthermore, the spectrum will also be calculated in several solvents in order to investigate the viability of a future approach that make use of "variable solvent circular dichroism". The basic ideas are that the balance between the different conformations of a molecule may be controlled by the choice of the solvent and that the ECD spectra are highly sensitive to conformation. As a consequence, it is expected that a change in solvent will result in a "fingerprint" shift in the ECD spectra, which can be both measured experimentally and calculated computationally, thus allowing the assignment of the stereochemistry about the chiral axis.

2. Theoretical Background

In this chapter, the theoretical concepts on which this work relies, will be briefly discussed. The discussion is divided into three main parts: Quantum Mechanics, Molecular Mechanics and Semi-empirical calculations.

2.1 Quantum Mechanics[13]

Quantum Mechanics (QM) is the most accurate approach to describe a system, since it is based on the explicit description of the nuclear and electronic structure. It is built on the solution of the Schrödinger equation, the mathematical expression used in quantum mechanics to describe the evolution in time of the relevant physical system.

2.1.1 The Schrödinger Equation

The time dependent Schrödinger equation is:

$$\widehat{H}\Psi(r,R,t) = i\hbar\frac{\delta\Psi(r,R,t)}{\delta t} \tag{2.1}$$

where Ψ is the wave function that describes the system, depending on the coordinates of both nuclei (R) and electrons (r), and $\widehat{H}$ is the Hamiltonian operator of the system. Planck's constant is given by h, $\hbar = h/2\pi$ and i is the imaginary number, with $i^2 = -1$. It is possible to separate the spatial and temporal contributions to the wave function, i.e. $\Psi(r,R,t) = \Psi(r,R)\tau(t)$. The spatial wave function, $\Psi(r,R)$, represents a stationary state satisfying the time-independent Schrodinger equation:

$$\widehat{H}\ \Psi(r,R) = E\Psi(r,R) \tag{2.2}$$

where E is the expectation value of the Hamiltonian and represents the energy associated with the wave function $\Psi(r,R)$. The Hamiltonian operator is defined as:

$$\hat{H} = \hat{T}_N + \hat{T}_e + \hat{V} \tag{2.3}$$

where $\hat{T}_N$ and $\hat{T}_e$ are the operators associated with the nuclear and electronic kinetic energy, respectively,

$$\hat{T}_N = -\hbar^2 \sum_I^N \frac{\nabla_i^2}{2M_I} \qquad \hat{T}_e = -\frac{\hbar^2}{2m} \sum_i^e \nabla_i^2 \tag{2.4a)(2.4b}$$

and $\hat{V}$ contains all the potential contributions due to the electron-electron, nucleus-nucleus and electron-nucleus Coulomb interactions,

$$\hat{V} = \hat{V}_{ee} + \hat{V}_{nn} + \hat{V}_{ne} = \frac{e^2}{4\pi\epsilon_0}\left(\sum_{i>j}^e \frac{1}{|r_i - r_j|} + \sum_i^e \sum_I^N \frac{Z_I}{|r_i - R_I|} + \sum_{I>J}^N \frac{Z_I Z_J}{|R_I - R_J|} \right) \tag{2.5}$$

2.1.2 The Born-Oppenheimer Approximation

Due to the large difference between the momentum of nuclei and electrons, it is possible to consider the latter as moving in a fixed-nuclei environment. This approximation, called the Born-Oppenheimer (BO) approximation, allows to define an electronic Hamiltonian neglecting the kinetic energy of the nuclei, whose associated Schrödinger equation is

$$\hat{H}^{elec}\Psi^{elec}(r, \boldsymbol{R}) = E(\boldsymbol{R})\Psi^{elec}(r, \boldsymbol{R}) \tag{2.6}$$

where $\boldsymbol{R}$ is the (parametric) position of the nuclei, $\Psi^{elec}(r, \boldsymbol{R})$ is the electronic wave function and E($\boldsymbol{R}$) is the corresponding electronic energy. In order to computationally resolve the above equation and obtain information about the equilibrium geometry and energy of the system, two main (and related) approaches have been developed: Hartree-Fock (HF) theory and Density Functional Theory (DFT).
All the calculations performed in this work are done using DFT, thus only this method will be briefly discussed in the following sections.

2.1.3 Slater Determinants

A Slater determinant provides a simple way to represent an antisymmetrized electronic wave function as a direct product of monoelectronic wave functions. The Slater determinant for a system composed of *N* electrons is:

$$\Psi(x_1, x_2..x_N) = \frac{1}{\sqrt{N!}} det \begin{bmatrix} \psi_1(x_1) & \cdots & \psi_N(x_1) \\ \vdots & \ddots & \vdots \\ \psi_1(x_N) & \cdots & \psi_N(x_N) \end{bmatrix} \tag{2.7}$$

or in the contracted form:

$$\Psi(x_1, x_2..x_N) = \frac{1}{\sqrt{N!}} det|\psi_1(x_1), \psi_2(x_2) \ldots \psi_N(x_N)| \tag{2.8}$$

where each electron is in a different spin orbital, ψ_i which is the product of a spatial function and a spin function:

$$\psi_i = \phi_i(r)\sigma_i(s) \qquad\qquad \langle\phi_i|\phi_j\rangle = \delta_{ij} \tag{2.9}$$

In this representation, two electrons having the same spin momentum cannot occupy the same spin orbital, in accordance with the Pauli exclusion principle.

2.1.4 Hartree-Fock (HF) Theory

Within the BO approximation, the HF method allows to calculate the ground state energy (E_0), through the variational determination of the spin orbitals that minimize the expectation value of the electronic Hamiltonian.

$$E_0[\Psi_{HF}] = \min_{[\Psi_{HF}]} \langle \Psi_{HF} | \hat{H}_{el} | \Psi_{HF} \rangle \tag{2.10}$$

where the trial function Ψ_{HF} is, in the mono-determinantal approximation, represented by a single Slater determinant. The HF energy is:

$$E_{HF} = \langle \Psi_{HF} | H_{el} | \Psi_{HF} \rangle = \sum_{i}^{N} h_i + \frac{1}{2} \sum_{i,j}^{N} (J_{ij} + K_{ij}) + V_{NN} \tag{2.11}$$

where h_i are the one-electron integrals, which take into account the kinetic energy contribution and the nucleus-electron potential for the i^{th} electron, while J_{ij} and K_{ij} are two-electron integrals, representing the classic Coulombian electron-electron repulsion and the exchange contribution, respectively.

$$\begin{aligned} h_i &= \langle \psi_i | \hat{T} + \hat{V}_{ne} | \psi_i \rangle \\ J_{ij} &= \langle \psi_i(1) \psi_i(1) \left| \frac{1}{r_i - r_j} \right| \psi_j(2) \psi_j(2) \rangle \\ K_{ij} &= \langle \psi_i(1) \psi_j(1) \left| \frac{1}{r_i - r_j} \right| \psi_i(2) \psi_j(2) \rangle \end{aligned} \tag{2.12}$$

It is now possible to split the single *N*-electron problem into *N* different monoelectronic problems, defining the set of Fock equations as:

$$\hat{F} \psi_i(x_i) = \epsilon_i \psi_i(x_i) \tag{2.13}$$

where $\hat{F}$ is the one-electron Fock operator and ϵ_i, formed by the same energy contributions as above, is the energy associated with the i^{th} spin-orbital.

It is possible to expand the spatial part of each spin orbital as a linear combination of basis functions χ_μ, called the basis set,

$$\phi_i(r) = \sum_{\mu=1}^{k} c_{\mu i} \chi_\mu(r) \tag{2.14}$$

where $c_{\mu i}$ is the expansion coefficient of the μ^{th} basis function. The expansion coefficients are

unknown and it is possible to rewrite the variational condition in terms of them,

$$E_0\left[\Psi_{HF}[c_i]\right] = \min_{[\Psi_{HF}[c_i]]} \langle \Psi_{HF}[c_i] \left| \widehat{H}_{el} \right| \Psi_{HF}[c_i] \rangle \tag{2.15}$$

Thus, the variational condition is satisfied by the expansion coefficients that resolve the matrix Roothaan-Hall (RH) equations,

$$FC_i = \mathrm{E}_i \mathrm{S} \mathrm{C}_i \tag{2.16}$$

where *F* is the Fock matrix, C_i are the vectors whose elements are the expansion coefficients of the i^{th} spin orbital and E_i is the vector containing the respective orbital energies.

The RH equations can be solved iteratively through the Self Consistent Field (SCF) method. The main disadvantage of the HF method is the neglect of the correlation energy, which results from treating each electron as moving in an average field of all other electrons and is defined as the difference between the HF energy and the true ground state energy. In fact, while the inclusion of the exchange term K_{ij} excludes the presence of two electrons with equal spin in the same position in space, due to the monodeterminantal approximation, there exist a non-null probability of finding two electrons with opposite spin in the same position in space. In order to calculate this contribution, it is necessary to go beyond the monodeterminantal approximation, describing the system as a linear combination of Slater determinants and involving other, computationally expensive, methods, e.g. Configuration Interaction (CI), Møller-Plesset perturbation theory (MPn), etc.

2.1.5 Density Functional Theory (DFT)

DFT is a theory for the ground state electronic structure, formulated in terms of the electronic density distribution. It is an alternative approach to the classical HF theory, which rather relies on wave functions. This approach is able to describe both the electronic interaction and correlation (even if in a non-exact way) that can be included in a classical HF method only through highly demanding (in terms of computational cost) methods.

DFT is based on the two Hohenberg-Kohn (HK) theorems:

1) The total energy is a unique functional of the electronic density

$$E[0] = \langle \Psi_0 | \widehat{H_{el}} | \Psi_0 \rangle = E[\rho] \tag{2.17}$$

2) The electronic density that minimizes the total energy is the exact electronic energy of the ground state and can be variationally calculated.

$$E_o \leq E[\tilde{\rho}] \tag{2.18}$$

The two theorems, thus, define a one-to-one correspondence between the total potential of the system and its own electronic density:

$$v(r) \leftrightarrow \rho(r) \tag{2.19}$$

The functional that describes the energy of the system is:

$$E[\rho] = V_{ne}[\rho] + F_{HK}[\rho] \tag{2.20}$$

where $V_{ne}[\rho]$ is the functional that describes the nucleus-electron Coulomb interaction and is easily calculable through the integral,

$$V_{ne}[\rho] = -\sum_{\alpha=1}^{M} \int \frac{Z_\alpha \rho(r)}{|R_\alpha - r|}\, dr \tag{2.21}$$

and F_{HK} is the HK functional that is formed by the sum of the kinetic term as well as the electron-electron interaction and correlation term.

$$F_{HK}[\rho] = T[\rho] + V_{ee}[\rho] \tag{2.22}$$

The explicit form of the HK functional is unknown and so the Kohn-Sham (KS) method is used to estimate it: the real system is approximated by an independent electron system (described by a single Slater determinant), subject to a single-electron effective potential such that the electronic density of the reference system is the same as that of the real system. Doing this, it is possible to describe the system through the KS one-electron operator, $\hat{F}_{KS}$, whose eigenfunctions are the spin orbitals of the reference system.

$$\hat{F}_{KS}\phi_i = \epsilon_i \phi_i \tag{2.23}$$

where

$$\hat{F}_{KS} = \hat{T}_s + \hat{V}_{KS} \tag{2.24}$$

Although not equal to the kinetic energy of the real system, it is possible to calculate the exact kinetic energy for the non-interacting reference system using wave function theory as:

$$T_{KS} = \langle \Psi_s | \hat{T}_s | \Psi_s \rangle \tag{2.25}$$

In addition, the full potential energy contribution (including electron correlation) of the real, interacting system is unknown, and must be approximated as well. The functional for the real system can then be written as:

$$E[\rho] = T_{KS}[\rho] + V_{ne}[\rho] + J[\rho] + E_{XC}[\rho] \tag{2.26}$$

where $T_{KS}[\rho]$ is the kinetic energy of the reference system, $J[\rho]$ is the classical Coulomb repulsion between the electrons in the real system and $E_{XC}[\rho]$ describes both the exchange contribution and the Coulomb correlation plus a kinetic contribution due to the inclusion of electron-electron interaction. This last term contains all the unknown factors and can be variationally calculated, knowing that the electronic density that minimizes the total energy (and so the minimum of the functional) corresponds to spin orbital functions that are eigenfunctions of the KS operator and therefore can resolve the Kohn-Sham equation (2.23). Finally, one can rewrite the Kohn-Shan operator (2.24) as:

$$\hat{F}_{KS} = \hat{T}_s + \hat{V}_{ne} + \sum_{i=1}^{N} \hat{J}_i + \hat{V}_{XC} \tag{2.27}$$

where $V_{XC} = \frac{dE_{XC}[\rho]}{d\rho}$ and E_{XC} is approximated as a functional of the electronic density and its own spatial derivatives.

2.1.6 Time-Dependent DFT (TDDFT)

DFT calculations offer a way of obtaining the ground state energy through a direct correlation between the energy of the system and the ground state electronic density (HK theorem). The latter is then obtained through a single-electron effective potential applied to a reference system of non-interacting electrons (KS). This method, thus, is a stationary method and is not able to describe very accurately the excited states' energies and properties, such as optical and dielectric properties. Whereas the HK theorems are the basis of the DFT method in the ground state, the Runge-Gross theorem represents the core of the time-dependent version of the method. The Runge-Gross theorem defines a direct and unique correlation between the external time-dependent potential applied to the system and the electronic density of that system, evolving from a fixed initial state Ψ_0,[14]

$$v_{ext}(r,t) \leftrightarrow \rho(r,t) \tag{2.28}$$

where the external potential contains both the static potential of the system in the ground state and a potential associated with an external perturbation (e.g. an electric field). The time-dependent electronic density is related to the time-dependent electronic wave function in the following way:

$$\rho(r,t) = N \int \psi_{el}^{*}(x_1 \dots, x_n, t)\psi_{el}(x_1, \dots x_n, t) \tag{2.29}$$

The theorem implies that knowing how the electronic density changes in time, starting from an initial state, it is possible to describe the Hamiltonian of the system through the application of an external field that reproduces the electronic density. The time-dependent Schrödinger equation for the system will become

$$\frac{i\delta}{\delta t}\Psi_i = \left(\hat{T} + \hat{v}_{ext}\right)\Psi_i \tag{2.30}$$

where $\hat{T}$ is the kinetic energy operator and a different electronic density corresponds to different external potentials and thus is associated with different systems, Ψ_i, evolving from the same initial state, Ψ_0. The time-dependent KS potential is:

$$V_{ks}(r,t) = V_{ne} + \int dr' \frac{\rho(r,t)\rho(r',t)}{|r-r'|} + \frac{\delta A_{xc}[\rho]}{\delta\rho(r,t)} \tag{2.31}$$

Where V_{ne} is the potential due to the nucleus-electron interaction, the integral in the second term takes into account the electron-electron Coulomb potential, and the last term contains the exchange and correlation terms. The variational condition is represented by:

$$\frac{\delta A[\rho]}{\delta\rho(r,t)} = 0 \tag{2.32}$$

where $A[\rho]$ is the action functional and describes the dynamics of the system in time under the effect of an external potential. Resolving the equation above, the exact values of the time-dependent electronic density are obtained.[14]

2.1.7 Linear Response Time-Dependent DFT (LR-TDDFT)

The usual approach involved in the calculation of excited-state properties is LR-TDDFT. In this method, a weak perturbation is applied to the system and a linear response is assumed. In this case, one can write for the external potential:

$$\nu_{ext}(r,t) = \nu_{ext,0}(r) + \delta\nu_{ext}(r,t) \tag{2.33}$$

where the first term is the ground state potential obtained through a classical DFT calculation and the $\delta\nu_{ext}$ term is the variation due to the external perturbation. From the Runge-Gross theorem, one can

then write:

$$\nu_{ext,0}(r) + \delta\nu_{ext}(r,t) \leftrightarrow \rho_0(\mathrm{r}) + \delta\rho(r,t) \tag{2.34}$$

and thus, a direct correlation exists between the perturbation applied to the ground state and the induced electronic density change. LR-TDDFT can be explained using the density matrix formalism,[15] starting from the Liouville equation:

$$\dot{P} = -\frac{i}{\hbar}[H;P] = -\frac{i}{\hbar}\hat{L}P \tag{2.35}$$

where $\hat{L}$ and P are the Liouville superoperator and the density matrix, respectively, and $\dot{P}$ is the first derivative of the density matrix with respect to time, representing the evolution in time of the system. The solution of the Liouville equation for a time independent Hamiltonian is given by:

$$\rho(t) = e^{-iHt/\hbar}\,\rho(0)e^{-iHt/\hbar} \tag{2.36}$$

If the system in the ground state is considered, before applying the perturbation, $\dot{P} = 0$ and the Hamiltonian and the density matrix commute, so that one can write:

$$F^{(0)}P^{(0)} - P^{(0)}F^{(0)} = 0 \tag{2.37}$$

where $F^{(0)}$ and $P^{(0)}$ are the Fock matrix and the density matrix associated with the unperturbed state. In this case, the eigenstates of F and P are the same, thus $P^{(0)}$ is diagonal and describes the probability of the n^{th} state to be populated. If an oscillating perturbation is applied, the off-diagonal elements (coherence) appear in the density matrix and the Liouville equation becomes:

$$FP - PF = -\frac{i}{\hbar}\frac{\delta P}{\delta t} \tag{2.38}$$

with, $P = P^{(0)} + P^{(1)}$ and $F = F^{(0)} + F^{(1)}$ so that:

$$F^{(0)}P^{(1)} - P^{(1)}F^{(0)} + F^{(1)}P^{(0)} - P^{(0)}F^{(1)} = -\frac{i}{\hbar}\frac{\delta P^{(1)}}{\delta t} \tag{2.39}$$

where $F^{(1)}$ contains the contribution due to the external perturbation and the contribution due to the variation of the density matrix:

$$F^{(1)} = g(\omega) + \frac{\delta F}{\delta P}P^{(1)} \tag{2.40}$$

$P^{(1)}$ is the variation of the density matrix due to the perturbation:

$$P^{(1)} = \frac{1}{2}(d_n e^{-i\omega_n t} + d_n^* e^{i\omega_n t}) \tag{2.41}$$

with d_n representing the density of the perturbation associated with the n^{th} transition. After substituting the above equations into eq.(2.39), it is possible to summarize the TDDFT response in the following matrix eigenvalue problem:[16]

$$\begin{pmatrix} A & B \\ B^* & A^* \end{pmatrix}\begin{pmatrix} X_n \\ Y_n \end{pmatrix} = \omega_n \begin{pmatrix} 1 & 0 \\ 0 & -1 \end{pmatrix}\begin{pmatrix} X_n \\ Y_n \end{pmatrix} \tag{2.42}$$

where the two vectors X_n and Y_n are correlated to the transition density matrix for the n^{th} transition at frequency ω_n and the elements of the A and B matrices are:

$$A_{ia,jb} = \delta_{ij}\delta_{ab}(\epsilon_a - \epsilon_i) + 2(ia|jb) + (ia|g_{XC}|jb) \tag{2.43}$$

and

$$B_{ia,jb} = 2(ia|jb) + (ia|g_{XC}|jb) \tag{2.44}$$

where *i,j* and *a,b* are the occupied and virtual KS orbitals, respectively, and g_{XC} contains the second derivates of the exchange-correlation functional,

$$g_{XC}(r,r') = \frac{\delta^2 E_{XC}[\rho]}{\delta\rho(r)\delta\,\rho(r')} \tag{2.45}$$

where *r* and *r'* are the electronic density coordinates of the two states involved in the transition.[17]

2.1.8 Density Functionals

Density functionals (DFs) differ according to the fundamental input variables used in the construction of the exchange-correlation potential and the values of the various coefficients (that may either be theoretically derived or empirically fit to experimental data) that make up the resultant expressions. A useful classification of DFs was put forward by Perdew and Schmidt and is referred to as "Jacob's ladder",[18] consisting of five rungs. The idea is that the sophistication (and presumably the associated accuracy) increases as one moves up the ladder, with each additional approximation building on those below. In addition, it is a common practice to separate the exchange and correlation parts so that the exchange-correlation energy is given as

$$E_{XC}[\rho(r)] = \int \rho(r)\varepsilon_X[\rho(r)]dr + \int \rho(r)\varepsilon_C[\rho(r)]dr \tag{2.46}$$

where the functionals are now written in terms of an energy density, ε. The functionals can then be constructed separately (or together, when fit to experimental data) and it is even possible to mix different exchange and correlation functionals, designed independently. The five steps in Jacob's ladder are discussed below.

2.1.8.1 LDA (Local Density Approximation)

LDA functionals depend only on local values of $\rho(r)$ and assume a uniform electron gas or, alternatively, that the electronic density varies very slowly,

$$E_{X/C}^{LDA} = \int dr\, \rho(r) \tag{2.47}$$

Consequently, the LDA is not very useful for describing chemical problems involving molecular

systems and has mostly been superseded by more sophisticated functionals. Furthermore, for systems with unequal α and β spin densities, each is described separately; this is called the local spin density approximation (LSDA).

2.1.8.2 GGA (General Gradient Approximation)

GGA functionals show significant improvement over the LDA and depend on both $\rho(r)$ and its first derivative, $\nabla\rho(r)$. They represent the first class of functionals that were generally applicable to problems in chemistry. Compared to LDA, GGAs are better at describing systems with inhomogeneous electron densities, e.g. molecules,

$$E_{X/C}^{GGA} = \int dr\ \rho(r)\epsilon_{X/C}^{GGA}(\rho(r), \nabla\rho(r)) \tag{2.48}$$

2.1.8.3 Meta-GGA

meta-GGA functionals depend on $\rho(r)$, as well as its first and second derivatives, $\nabla\rho(r)$ and $\nabla^2\rho(r)$.

$$E_{X/C}^{m-GGA} = \int dr\ \ \rho(r)\epsilon_{X/C}^{m-GGA}\big(\rho(r), \nabla\rho(r), \nabla^2\rho(r)\big) \tag{2.49}$$

However, the orbital kinetic energy density, $\tau(r)$,

$$\tau(r) = \frac{1}{2}\sum_{i}^{occ}|\nabla^2\phi_i(r)|^2 \tag{2.50}$$

where the sum runs over all occupied orbitals, and the second derivative (Laplacian), contain essentially the same information and since calculation of the former is numerically more stable, is often preferred for inclusion in meta-GGA functionals, giving

$$E_{X/C}^{m-GGA} = \int dr\ \ \rho(r)\epsilon_{X/C}^{m-GGA}(\rho(r), \nabla\rho(r), \tau(r)) \tag{2.51}$$

Meta-GGA functionals are usually more accurate than GGA functionals, but more computationally expensive as well. Although occurring on the third rung of Jacob's ladder, hybrid functionals (discussed next, on the fourth rung) predate meta-GGA functionals and hence, most of the latter are also hybrids functionals.

2.1.8.4 Hybrid Functionals

In this type of functionals, part of the exchange energy contribution is calculated using the HF formalism by making use of the KS orbitals. Since the HF equations are exact, if the KS orbitals were those of the exact wave function, the HF exchange would be the true, exact exchange, hence this contribution is called "exact exchange", E_X^{HF}, giving the exchange-correlation energy as

$$E_{XC}^{hybrid} = [(1-a)E_X^{DFT} + aE_X^{HF}] + E_C^{DFT} \tag{2.52}$$

where a is a scaling factor. Another important consequence of including exact exchange in hybrid functionals, is that it reduces the self-interaction error. This error is present in all DFs, due to the expression of the Coulomb interaction as a double integration over two independent electron densities.

2.1.8.5 Double Hybrid Functionals

Double hybrid functionals improve on the previous four rungs by using arguments from perturbation theory to mix in a contribution to the correlation energy obtained from the KS virtual orbitals. This is similar to the approach used in MP2 (Møller-Plesset perturbation theory of 2nd order) wave function methods, but using KS orbitals instead of HF orbitals. The exchange-correlation energy can be written as

$$E_{XC}^{double-hybrid} = [(1-a)E_X^{DFT} + aE_X^{HF}] + (1-b)E_C^{DFT} + bE_C^{MP2} \tag{2.53}$$

2.1.8.6 PBE0 (PBE1PBE)

The hybrid functional used in this work is called PBE0, also known in the literature as PBE1PBE.[19]

In PBE0, the value of a in the above equation is set to ¼ and the DFT functional depends on the electronic density $\rho(r)$, and its gradient $\nabla\rho(r)$,[20]

$$E_{X/C}^{PBE} = \int d^3r\, \rho(r)\, \epsilon_{X/C}^{PBE}(\rho(r), \nabla\rho(r) \tag{2.54}$$

2.1.9 Basis Sets

Whether wave function-based methods or DFT is used, the solution to the Schrödinger equation for molecular systems is expressed in terms of molecular orbitals, where each is assumed to consist of a linear combination of atomic orbitals (LCAO):

$$\psi_i = \sum_m c_{mi}\phi_m \tag{2.55}$$

The basis set is the ensemble of functions used to represent the atomic orbitals. The set of basis functions (called a contraction) used to describe the orbitals for each atom are conveniently defined as a linear combination of Gaussian functions (called primitives, PGTFs):

$$\chi_\mu(r) = \sum_s d_{\mu s}\, g_s(\alpha_s, r^2) \tag{2.56}$$

where χ_μ is the μ^{th} contracted function, g_s is the s^{th} primitive function, α_s is the expansion coefficient and the $d_{\mu s}$ values are fixed constants within a given basis set. Slater-type orbitals (STOs) are similar to the hydrogenic-type functions obtained from solving the Schrödinger equation exactly, for a one-electron system, and provide the best physical description of atomic orbitals. However, GTFs are used in most computational packages because integrals involving Gaussian functions are easier to calculate numerically. However, the shape of an STO cannot be properly described by a single GTF (no cusp, and falls away too quickly far from the nucleus) and hence a linear combination of PGTFs is used. The different basis sets are discriminated based on the numbers and types of GTFs that are used. The PGTFs have different shapes, depending on which orbital they describe. In principle, one can have *s*, *p*, *d* and *f*-type orbitals. The choice of the basis set is dictated by a compromise between accuracy in the

molecular description and computational cost. The larger the basis set, the more accurate the wave function/density, but also the more computationally expensive the calculation.

2.1.9.1 Minimal Basis Sets

A minimal basis set represents the most elementary basis set, in which each occupied atomic orbital is defined by one single STO. Each STO is then expanded as a linear combination of PGTFs to provide a better description of the orbitals. The general formulation for this basis set, in Pople notation[21], is STO-*n*G, where *n* represents the number of primitive functions used to describe a single STO function.

2.1.9.2 Split Valence Basis Sets

In a split valence basis set, the number of functions per atom is increased and each orbital in the valence region is expressed as a combination of two or more STO functions. The different split valence basis sets (e.g. double zeta, triple zeta, etc.) are characterized by the number and size of contractions used to describe each atom in the valence region, whereas the core orbitals are described by one large contraction. In the case of the double zeta basis set, for example, two different sizes of contractions are used. The general formulation in Pople-style is M-N1G, where M represents the number of PGTFs used to describe the core orbitals, and N and 1 represent the numbers of primitive functions used to describe each of the valence contractions. Increasing the "electron flexibility" makes the calculation of the electric dipole moments more precise.

2.1.9.3 Polarized Basis Sets (M-N1G**)

In a polarized basis set, a group of functions, which have greater angular momenta than the highest lying occupied orbitals in the valence shell, are added, such as *p*-type functions to hydrogen atoms, *d*-type functions to second and third row atoms and *f*-type functions to the metals. These functions, with different shapes, allow the system to be more flexible in the electron arrangement and get a more reliable description of the geometry.

2.1.9.4 Diffuse Basis Sets (M-N1++G)

Diffusion functions can also be added, having the same angular momentum as the set of orbitals in the valence shell, but with a larger dimension. It is useful to describe systems with electrons far from the nuclei, i.e. systems with a significant negative charge, anions or excited systems, with diffuse functions.

2.1.10 Main Types of Calculations

2.1.10.1 Calculation of the Optimal Geometry

Calculating the optimal geometry means "exploring" the potential energy surface (PES), obtained through the resolution of the electronic problem of the system with fixed nuclei, in search of the global minimum. The geometry associated with the minimum energy corresponds to the most stable geometry in terms of the chosen basis set. In order to do this, the energy gradient vector is used (partial derivative with respect to the coordinates of the nuclei):

$$g = \left(\frac{\delta E}{\delta \mathrm{X}_1}, \frac{\delta E}{\delta \mathrm{X}_2}, \dots, \frac{\delta E}{\delta \mathrm{X}_{3N}}\right) \tag{2.57}$$

The eigenvalues of the Hessian matrix (second partial derivative with respect to the coordinates of the nuclei) are then used to check the nature of the stationary points (e.g. minimum or saddle point).

$$h_{ij} = \left(\frac{\delta E}{\delta \mathrm{X}_i \delta \mathrm{X}_j}\right) \tag{2.58}$$

2.1.10.2 Calculation of Vibrational Frequencies and ZPE

Calculating the 3M-6 (and 3M-5 for linear molecules) vibrational frequencies, with M the number of nuclei, means resolving the vibrational problem. Mathematically, vibrational frequencies can be calculated through the diagonalization of the mass-weighted Hessian matrix. The numerical values obtained correspond to the squared frequencies of the 3M-6 harmonic oscillators, each one describing a

vibrational normal mode of the system.

The zero-point energy (ZPE) represents the vibrational energy of the system at 0 K temperature, which includes contributions from all the 3M-6 harmonic oscillators in their fundamental state,

$$E_{ZPE} = \frac{1}{2}\sum_{i} h\omega_i \tag{2.59}$$

where ω_i are the fundamental frequencies and the sum runs over all normal modes. The ZPE is not included in the SCF calculation of the electronic energy, however, it should be taken into account if accurate energy differences between different conformations, are required.

2.1.11 Solvent Models

Correct representation of the environment (solvent) is an important feature in order to best describe geometries, spectroscopic and chemical properties, such as electronic processes and reactivity. Regarding, spectroscopic properties, the solvent can influence the spectrum in two ways: changing the preferred conformation/geometry of the solute and through "chiral imprinting". The latter is the ability of a chiral solute to orient the solvent molecules in its solvation sphere. This condition generates a chiral environment around the solute that can strongly influence the optical rotation and thus the correlated ORD spectrum.[22]

It is possible to divide the modelling of solvent effects into two main categories: (1) an explicit model, where molecules of the solvent are added and the system is treated either quantum mechanically or through a hybrid QM/MM model (see Section 2.3) and (2) implicit (continuum) models, where the solvent contribution is implicitly treated inside the Hamiltonian. In this work, the SMD[23] (Solvation Model based on Density) solvation model was used. This model belongs to the implicit group, so that only this family of formulations will be discussed in the following. This approach is in fact the most often used, since it gives a good solvent description with a computational cost comparable to the QM calculation in the gas phase.

Generally, the Hamiltonian involved in the description of the solvent is:

$$\hat{H} = \hat{H}_0 + \hat{H}_{env} + \hat{H}_{int} \tag{2.60}$$

where the first two terms are the Hamiltonians associated with the purely QM system in the gas phase and the purely classical description of the solvent, respectively, and $\widehat{H}_{int}$ describes the interaction between the two parts. In the continuum formulation, the solvent is described by a homogeneous dielectric medium and the solute by a molecule-shaped cavity inside this medium, so that the molecules of the solvent lose their 'atomic nature' and the $\widehat{H}_{env}$ term disappears. Thus, the free energy of solvation (ΔG_s), defined as the free energy difference between the solute in the gas phase and the solute inside the medium, will be proportional to the expectation value of $\widehat{H}_{int}$. In the SMD model, the cavity surface, constructed as a Solvent Accessible Surface (SAS), is formed by a superposition of infinitesimal nuclear-centered spheres with vdW radii, R_{Z_k}, and surface tension, σ_k, for the k^{th} atom. It is possible to separate ΔG_s into two main contributions:

$$\Delta G_s = \Delta G_{ENP} + \Delta G_{CDS} \tag{2.61}$$

where the first term includes all the electrostatic contributions associated with the Electronic (E), Nuclear (N) and Polarization (P) components of the free energy, while the second represents the non-electrostatic contributions due to solvent Cavitation (C), change in Dispersion (D) energy and the energy associated with the change in the solvent (S) structure. The electrostatic part can be calculated by solving the Poisson equation,

$$\epsilon \nabla^2 \Phi = -4\pi\rho_f \tag{2.62}$$

where ϵ is the permittivity of the medium, ρ_f is the solute charge density and Φ is the electrostatic potential of the medium. The value of ΔG_{ENP} is given by:

$$\Delta G_{ENP} = \langle \Psi | H^{(0)} - \frac{e}{2}\phi | \Psi \rangle + \frac{e}{2}\sum_k Z_k \phi_k \; - \; \langle \Psi^{(0)} \big| H^{(0)} \big| \Psi \rangle \tag{2.63}$$

where $\Psi^{(0)}$ and $H^{(0)}$ are the wave function and the Hamiltonian in the gas phase, Ψ is the wave function in the solvent, Z_k is the atomic number of the k^{th} atom and $\phi = \Phi - \Phi^{(0)}$ is called the reaction field, where $\Phi^{(0)}$ is the electrostatic potential of the gas phase molecule. The reaction field at

an arbitrary position, r , on the cavity surface, can be calculated as:

$$\phi(r) = \sum_{m} \frac{q_m}{|r - r_m|} \tag{2.64}$$

where q_m is the apparent surface charge on the m^{th} cavity surface area element, at position r_m.
The CDS contribution, can be calculated as:

$$G_{CDS} = \sum_{k}^{atoms} \sigma_k A_k\left(R, \{R_{Z_k} + r_S\}\right) + \sigma^{[M]} \sum_{k}^{atoms} A_k\left(R, \{R_{K_z} + r_s\}\right) \tag{2.65}$$

where σ_k and $\sigma^{[M]}$ are the atomic surface tension due to the k^{th} atom and to the molecular surface tension respectively, A_k is the solvent accessible surface area, which depends on the position R_{Z_k}. $r_S = 0.4$Å is the solvent probe radius and is added to the vdW radii to define the position.

2.2 Semi-Empirical (SE) Methods

Semi-empirical calculations are very useful for systems that are too large for an *ab initio* QM approach. These methods follow the same formulation as the Hartree-Fock method, but are computationally much less demanding since several, usually small, integrals are neglected and others are replaced by parameters, obtained from both experimental data and simplified approximations.
In this study, the two adopted SE methods (AM1[24] and PM6[25]) belong to the larger class of Neglect of Diatomic Differential Overlap (NDDO) methods, where the one-electronic, h_i, contributions as well as the core potential of the HF energy (2.11), are completely parameterized. The two-electron integrals are expanded in simpler multipole-multipole contributions and then parameterized. Correlation is implicitly included in the latter parameters through an underestimation of the exact value.[26] The main difference between the two methods is the number and the kind of parameters inside the core potential. AM1 was a revolution in the field of SE methods, because it was the first to be sufficiently accurate to be applied to chemical problems in general. It also included Gaussian functions to describe hydrogen bonding inside the core potential. Following the development of AM1, several improvements have been made in the accuracy of the parameterization. PM6 represents an up-to-date method compared to

AM1, where all the main group element parameters have been improved with a particular focus on the description of molecules with biochemical interest.[25]

2.3 Molecular Mechanics (MM)

Molecular mechanics is the branch of computational chemistry that attempts to describe the structural and energetic properties of a system using classical mechanics principles in which each atom is approximated as a charged particle and bonds as springs. MM uses empirical force fields (FFs) to describe and model the system of interest. A FF is a mathematical equation, parameterized using experimental (X-Ray, NMR, etc.) or *ab inito*/semi-empirical data, which describes the energy a system through the relative positions of the particles and the bonds between them (see the next section).

The main advantage of MM is the cost of the calculation; MM calculations are computationally much less demanding and orders faster than *ab initio* or semi-empirical calculations. This means that bigger systems can be processed in a very short time (it is possible to run computations on millions of atoms on the order of milliseconds).[27] The limits, however, are the inability to describe the electronic structure (such as the breaking and the formation of bonds or the excitation of an electron), and the high specificity of each FF to the type of atoms and bonds it can describe. This means that the FF must be chosen depending on the kind of problem and nature of molecules one wishes to model and a FF that is able to accurately describe the properties of one specific class of systems, might be very inaccurate for other classes of systems, depending on the parameterization process. Fortunately, many FFs are available nowadays and if performance is not as expected, further parameterization is always possible. The "ideal" FF should be able to describe all the properties of a system (molecular geometry, conformational and stereoisomeric energies, intermolecular interactions, torsional barriers, vibrational frequencies, etc.) at the same level of accuracy as a corresponding QM calculation.

In the next section, a brief description of the general form of a FF is given, with a particular focus on the Merck Molecular Force Field (MMFF) and the Optimized Potential for Liquid Simulations (OPLS) FF. [27–29]

2.3.1 Force Fields

A force field defines the energy of a system through a parameterized potential energy function,

depending on the position of the particles, $U(r_0, r_1, \ldots)$. The general form of a force field is given by the following equation:

$$\begin{aligned} U = & \sum_{bonds} \frac{1}{2} k_b (r - r_0)^2 \\ & + \sum_{angle} \frac{1}{2} k_a (\theta - \theta_0)^2 \\ & + \sum_{torsion} \frac{V_n}{2} [1 + \cos(n\phi - \delta)] \\ & + \sum_{improper} \frac{1}{2} k_{imp} (r - r_o)^2 + \sum_{ij} 4\epsilon_{ij} \left(\frac{\sigma_{ij}^{12}}{r_{ij}^{12}} - \frac{\sigma_{ij}^{6}}{r_{ij}^{6}} \right) + \sum_{elec} \frac{q_i q_j}{r_{ij}} \end{aligned} \tag{2.66}$$

The potential energy has six different contributions. The first two terms in eq. (2.66) approximate bond stretching and bending as a harmonic potential. The force constants are given by k_b and k_a and the equilibrium values are r_0 and θ_0. The third term describes the torsional energy across three bonds as a Fourier series, where V_n is the potential barrier associated with the torsion around the angle ϕ, n is the periodicity and δ is the phase offset. The positive contribution associated with improper torsions, in the fourth term, is necessary to ensure the planarity of some groups, such as sp^2 carbons. The last two terms are associated with the van der Waals and electrostatic interactions between the particles, respectively. In the first of these terms, σ_{ij} is related to the diameter of particles i and j and ϵ_{ij} is the value of the minimum interaction energy between the particles. In the second term, q_i and q_j are the partial charges of particles i and j. Furthermore, the vdW parameters for an interacting pair are obtained from combining rules that are typically (but not limited to) either the arithmetic or geometric mean of the individual particles, e.g. for the σ_{ij} parameters, this can be either

$$\sigma_{ij} = \frac{1}{2}(\sigma_i + \sigma_j) \qquad \text{or} \qquad \sigma_{ij} = \sqrt{\sigma_i \sigma_j} \tag{2.67}$$

FFs are continuously changed and improved, so it is possible to find different parameterizations of the terms in literature or even new contributions to best describe different systems, e.g. the addition of hydrogen bonds, polarization effects, etc. Based on the results of previous work,[12] here two FFs have been used: MMFF and OPLS. [27,29]

2.3.1.1 MMFF and MMFFs

The Merck molecular force field is a second generation FF, parameterized using data obtained from *ab initio* calculations. The main improvements are the inclusion of a stretch-bend term to improve the geometry description and a more sophisticated way to describe the vdW and electrostatic interactions. Equation (2.66) takes into account the vdW interactions using the classical 12-6 repulsion-attraction Lennard-Jones potential. The MMFF, however, uses a different "buffered 14-7" form and a buffered form of electrostatic interaction as well, as shown in Equations (2.68) and (2.69). This provides a better description of the non-bonded interaction and electrostatics.

$$E_{vdW\,ij} = \epsilon_{ij}\left(\frac{1.07R_{ij}^{*}}{R_{ij}+0.07R_{ij}^{*}}\right)^{7}\left(\frac{1.12R_{ij}^{*7}}{R_{ij}^{7}+0.12R_{ij}^{*7}}-2\right)^{7} \tag{2.68}$$

$$E_{Q_{ij}} = \frac{332.0716q_iq_j}{D\left(R_{ij}+\delta\right)^{N}} \tag{2.69}$$

In Equations (2.68) and (2.69), R_{ij}^{*} is the minimum energy distance between atoms *i* and *j*, obtained using an augmented arithmetic mean expression, *D* and δ are the dielectric constant and the electrostatic buffering constant (0.05Å), respectively, and *N* is usually set to 1 or 2 when a distance dependent dielectric constant is used. The well depth, ϵ_{ij}, is obtained using a Slater-Kirkwood based formula. Finally, MMFFs is the "static" (s) version of the regular MMFF. Energies given by the two force fields are mostly the same, but when an sp^2 nitrogen is present the static version gives a better description of the planarity of the bond.[30,28,31]

2.3.1.2 OPLS3

OPLS3 is a new generation FF that has been refined in order to have an improved description of liquid state thermodynamic properties. In comparison to the first version of the OPLS model, it has a very accurate parameterization, based on data two orders of magnitude more than the previous version. In particular, all the parameters regarding bond stretches, angles and torsions have been obtained from QM calculations, while the vdW and Coulomb interactions have been refined to better reproduce solvation free energies. The general functional form is given below:

$$E = \sum_{i<j} \left[q_i \, q_j e^2 / r_{ij} + 4\epsilon_{ij} \left(\sigma_{ij}^{12} / r_{ij}^{12} - \sigma_{ij}^{6} / r_{ij}^{6} \right) \right] f_{ij} + \sum_{bonds} K_r \left(r - r_{eq} \right)^2 + \sum_{angles} K_\theta \left(\theta - \theta_{eq} \right)^2 + \sum_{dihedrals} [\frac{V_1}{2}(1 + cos\varphi) + \frac{V_2}{2}(1 - cos2\varphi) + \frac{V_3}{2}(1 + cos3\varphi) + \frac{V_4}{2}(1 - cos4\varphi)] \quad (2.70)$$

The Coulomb and vdW interactions are calculated for each *ij* couple, where *i* and *j* can also be virtual sites that are not coincident with the position of the atoms. f_{ij} is a scale factor, applied to every couple of atoms separated by three or less bonds. The main improvements are the description of the energy barrier for bond torsions, bond charge increment corrections and a better sigma hole description through the incorporation of off-atom charge sites. All these features make this FF one of the most accurate to describe structures in a liquid matrix, interactions between proteins and ligands as well as conformational studies of organic molecules.[29]

2.4 Conformational Analysis

The generation of an ensemble of conformers is the first crucial step in the calculation of the ECD spectrum of a molecule. A conformational analysis must correctly explore the potential energy surface and detect the structures that correspond to local minima in the chosen energy range without generating unreasonable geometries or undesirable high-energy structures. This study can be performed at different levels of theory (both QM and MM) but, since this calculation can involve the generation of a large number of conformers, the QM approach is computationally very expensive for molecules that present a high number of flexible bonds and therefore impractical. The general approach of searching for conformers can be either systematic or stochastic; the first is based on the exploration of 3D space through a systematic increment of the torsion angle of each rotatable bond (very expensive for large, flexible molecules), while the second is a random process that can involve different algorithms (genetic algorithms, Monte Carlo simulation, distance geometry, etc.).[32]

In this work, the conformational search has been performed using implementations in two different

software packages. Firstly, based on previous studies,[12] using the MacroModel [33] implementation in the Maestro software package[34] and secondly, the open source RDKit package.[35] Both use a stochastic approach: MacroModel's conformational search uses a mixed Monte Carlo/low-frequency mode search method, while RDKit uses a distance geometry method.

2.4.1 MacroModel and the Mixed Monte Carlo Multiple Minima/Low-Mode Conformational Search Method

This mixed method was borne out of the idea of joining the strengths of two methods: Monte Carlo simulation (MC) and the Low-MODe conformational search (LMOD). The Monte Carlo Multiple Minimum (MCMM) search is a very efficient method to explore both the close and far regions of the PES. For each step, the MCMM method generates random changes in dihedral angles of rotatable bonds; then samples the rings by pretending to break a bond and treating all the other bonds as rotatable, if the atoms in the broken bond end up lying close to each other this bond is reformed, otherwise, another set of bond rotations is attempted; it then performs several steps of minimization using the selected FF and finally does a stereochemical check. If the energy of the new conformer is within the selected energy window, it will be added to the conformer ensemble. The main disadvantages of this method are the high computational cost if the number of rotatable bonds is large and the bad description of the changes when explicit atoms fill most of the space.

The LMOD, instead, is a local method based on a saddle point search. This method explores the low-frequency vibrations and through an amplification of the normal mode, generates different structures. It can be applied to both small and big molecules, but the individual search is more expensive than the MCMM method. The mixed method then performs a search in which some steps are performed via the MCMM method and others involving the LMOD method.[36,37]

2.4.2 RDKit and Distance Geometry

In the distance geometry approach, the first step is creating a matrix containing, for each couple of atoms (*i* and *j*), the upper and lower limit distance (see Figure 15).

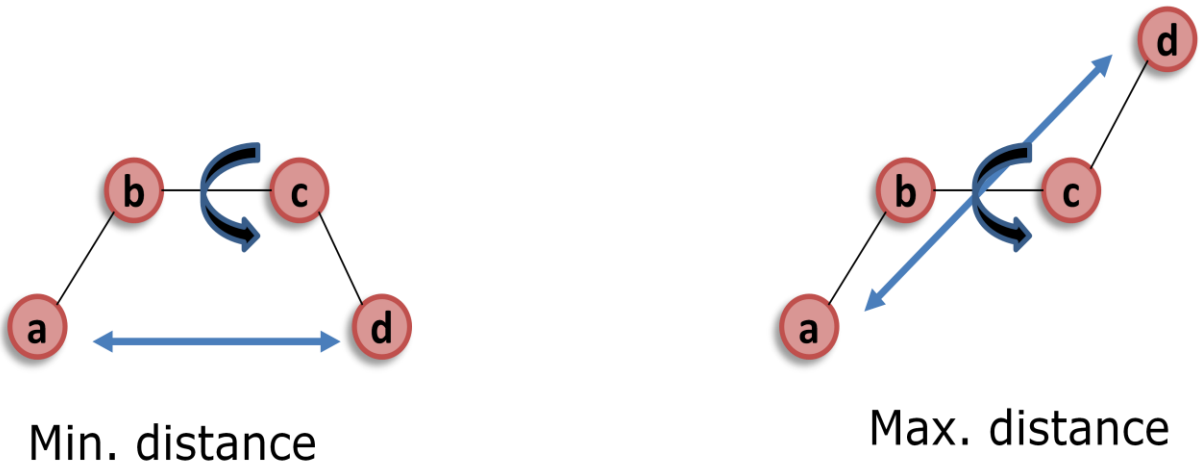


Figure 15. Minimum and maximum distance, calculated through rotation around the b-c bond.

This can be expressed in the following equation:

$$l_{i,j} \leq |x_i - x_j| \leq u_{i,j} \tag{2.71}$$

where x represents the positions of the atoms while l and u are, respectively, the minimum and the maximum distance between the atoms. Then, a random distance matrix is generated, in which all the distances are bounded by the previous values. The random distance matrix will be composed of several x values $\{x_1 ... x_m\}$ that satisfy the condition:

$$|x_i - x_j| = \delta_{i,j} \tag{2.72}$$

where

$$\delta_{i,j} \in [l_{i,j}, u_{i,j}] \tag{2.73}$$

The 3D coordinates are then generated and an ensemble of structures is obtained. RDKit supports the use of either UFF or MMFF94 to perform a MM optimization of the structures and remove all the redundant geometries.[32,38]

3. Electronic Circular Dichroism

This chapter will focus on the description of circular dichroism from a quantum mechanical point of view, in order to understand the origin of the signals in ECD spectra. Firstly, a brief introduction to the basic concepts regarding radiation will be given, secondly, Linear Response Theory (LRT) will be presented in order to describe the matter-radiation interaction and finally, Circular Dichroism (CD) and Exciton coupling CD will be described.

3.1 Radiation[9]

An electromagnetic wave can be described as the sum of an oscillating electric and magnetic field having equal amplitude, that propagate in one direction (e.g., z), mutually orthogonal and orthogonal to the propagation direction as well. If one considers radiation propagating in a medium, having index of refraction n, the two fields can be expressed as:

$$E = Re\{E_0 e^{i\psi}\} \qquad B = Re\{B_0 e^{i\psi}\} \qquad \text{(3.1a), (3.1b)}$$

where $E_0$and B_0 are the amplitudes of the electric and magnetic oscillation, $Re\{\}$ indicates that only the real part is take into account and

$$\psi = (2\pi\nu(t - n\vec{k}\vec{r}/c)) \qquad \text{(3.2)}$$

which represents the phase of the wave having frequency ν at time t and position $\vec{r}$. The vector $\vec{k}$ defines the direction of propagation so that, in this case, one can rewrite $\vec{k}\vec{r} = \vec{z}$. It is possible to write expressions (3.1a and 3.1b) using the Euler equation as:

$$E = \overrightarrow{v_e} E_0 cos\psi \qquad B = \overrightarrow{v_B} B_0 cos\psi \qquad \text{(3.3a)(3.3b)}$$

with $\overrightarrow{v_B} = \overrightarrow{v_e} R$ and R is the rotation matrix that produces a 90° rotation.

$$R(90°) = \begin{bmatrix} 0 & -1 \\ 1 & 0 \end{bmatrix} \tag{3.4}$$

If one now considers linearly polarized radiation, it is useful to describe this as the sum of two circularly polarized waves having half intensity of the initial radiation and opposite polarization, e.g. for the electric field:

$$E_+ = \frac{E_0}{2}(i\cos\psi - j\,sin\psi) \qquad E_- = \frac{E_0}{2}(i\cos\psi + jsin\,\psi) \tag{3.5a)(3.5b}$$

where *i* and *j* are the orthogonal unit vectors. These two expressions sum together to give equation (3.3a).

3.2 Linear Response Theory[39]

Liner response theory (LRT) offers a way to describe a spectroscopic experiment through the macroscopic description of the matter-radiation system, giving back direct information of the macroscopic property of interest. If $\langle\hat{B}\rangle_0$ is defined as the expectation value associated with a macroscopic property in thermodynamic equilibrium, then, for weak perturbation, it is possible to assume a linear response of the system to an external force $F(t)$ and calculate the expectation value associated with the deviation from the equilibrium, $\langle\tilde{B}(t)\rangle$. The Hamiltonian that describes the interaction between the system and the applied force, is given by

$$H' = -AF(t) \tag{3.6}$$

where *A* is the observable conjugate to the force $F(t)$. Assuming a linear response,

$$\langle\widetilde{B}\,(t)\rangle = \langle\hat{B}\rangle - \langle\hat{B}\,\rangle_0 = \int_0^t \phi_{BA}(t-t')F(t)dt' \tag{3.7}$$

where ϕ_{BA} is the response function of the system that measures the variation of the observable *B* under a perturbation, *A*, applied at time t'. ϕ_{BA} is a real function, since both the force and the observable are

real, with $\phi_{BA}(t) = 0$ if $t < t'$ and $\phi_{BA}(t) = 0$ if $t \to \infty$. Applying a monochromatic force, written as $F(t) = \frac{1}{2} F_\omega(e^{-i\omega t} + e^{i\omega t})$, gives

$$\langle \widetilde{B}(t) \rangle = \frac{1}{2} F_\omega e^{-i\omega t} \int_0^\infty \phi_{BA}(\tau) e^{i\omega\tau} d\tau + \frac{1}{2} F_\omega e^{i\omega t} \int_0^\infty \phi_{BA}(\tau) e^{-i\omega\tau} d\tau \tag{3.8}$$

where $\tau = (t - t')$ and ω is the frequency of the radiation. The susceptibility function is now introduced as the Fourier transform (FT) of the response function.

$$\begin{aligned} \chi_{BA}(\omega) &= \int_0^\infty \phi_{BA}(\tau) e^{i\omega t} d\tau \\ &= \int_0^\infty d\tau \ cos(\omega t)\, \phi_{BA}(\tau) + i \int_0^\infty d\tau \ sin(\omega t)\, \phi_{BA}(\tau) \\ &= \chi'_{BA}(\omega) + i\chi''_{BA}(\omega) \end{aligned} \tag{3.9}$$

where in the second line, the Euler equation was used. As is shown in (3.9), χ_{BA} is a function formed by an even real part and an odd imaginary part, thus, after substituting (3.9) into (3.8), $\langle \widetilde{B}(t) \rangle$ becomes:

$$\langle \widetilde{B}(t) \rangle = F_\omega \left[\chi'_{BA}(\omega)\, cos(\omega t) + \chi''_{BA}(\omega) sin(\omega t) \right] \tag{3.10}$$

where $\langle \widetilde{B}(t) \rangle$ is still real. This highlights that a system perturbed by a monochromatic field gives back two different responses: one in phase with the perturbation, proportional to the real part of the susceptibility $\chi'_{BA}(\omega)$ and the other one out of phase by $\pi/2$ to the imaginary part, $\chi''_{BA}(\omega)$.

To calculate the exchange energy between radiation and matter, the variation of the energy of the matter in time must be considered. This can be done by considering the electric power, *W*,

$$W = -\frac{1}{\tau} \int_0^\tau \frac{\partial E}{\partial t} \, dt = \frac{1}{2} F_\omega^2 \, \omega \, \chi''_{BA}(\omega) \tag{3.11}$$

where τ is an interval of time that contains several periods of the radiation. From this equation, it is

possible to see that the active phenomenon, which involves energy exchange between matter and radiation, depends only on the imaginary term of the susceptibility function, while the real part of the latter is involved in the passive phenomena.

3.3 Circular Dichroism

In order to describe the physical origin of circular dichroism, one must go beyond the dipole approximation that is normally applied in the description of the interaction between radiation and matter. In particular, it is necessary to take into account at least the linear term of the series expansion for the radiation expression:

$$e^{\pm i\vec{k}\vec{r}} = 1 \pm i\vec{k}\vec{r} \pm \cdots \tag{3.12}$$

In this way, the interaction Hamiltonian between the sample and the electromagnetic field, H_{int}, will also contain the contributions due to the magnetic dipole and the electric quadrupole. The latter can be safely neglected, since its contribution is orders of magnitude smaller than the other contributions and is not relevant to the description of the CD phenomenon.

$$H_{int} = -\hat{\mu}E_0\, cos(\omega t) - \hat{m}_B B_0\, cos(wt) \tag{3.13}$$

In (3.13), $\hat{m}_B$ represents the projection of the magnetic dipole moment operator in the direction of the applied magnetic field, *B*, and B_0 is the amplitude of the latter.

An easy way to describe CD is to consider the passive phenomena of rotation of polarized light and then highlight the connection to circular dichroism. Consider radiation travelling in an isotropic sample along the *z* direction, with the electric field (*E*) polarized along the *x* direction. This, obviously, is able to generate an oscillating electric dipole moment (and so a response) only in the *x* direction. However, in this way it is not possible to explain the phenomena that involve the rotation of the polarization of the radiation. In order to do that, one has to realize that the perturbation due to the magnetic field of the radiation, that is orthogonal to *E*, i.e. directed along *y*, is able to induce an oscillating electric dipole moment in the *y* direction, thus inducing the rotation of the polarization direction of the radiation. If one considers a weak perturbation, according to LRT, the observable rotatory power, φ, is associated

with the response function, $\Phi_{\mu_y m_y}$, in the following way:

$$\varphi = \int_{-\infty}^{t} \Phi_{\mu_y m_y}(t - t')F(t')dt' \tag{3.14}$$

where $F(t') = \frac{1}{2}B_0\left(e^{-i\omega t'} + e^{i\omega t'}\right)$. The response function is:

$$\begin{aligned}\Phi_{\mu_y m_y} \propto -\frac{i}{\hbar}\langle g|\{\hat{m}_y\ e^{iH_0\tau/\hbar}\ \hat{\mu}_y e^{-iH_0\tau/\hbar} \\ - e^{iH_0\tau/\hbar}\ \hat{m}_y\ e^{-iH_0\tau/\hbar}\ \hat{\mu}_y\}|g\rangle\end{aligned} \tag{3.15}$$

where $\tau = (t - t')$ is the time after the perturbation, μ_y is the observable that one measures, m_y is the observable associated with the perturbation and $|g>$ is the ground state wavefunction of the unperturbed Hamiltonian. Knowing that

$$\sum_f |f\rangle\langle f|\ = 1 \tag{3.16}$$

and

$$e^{iH_0\tau/\hbar}|f\rangle = e^{i\epsilon_f\,\tau/\hbar}|f\rangle \tag{3.17}$$

where $|f\rangle$ are the eigenstates of H_0, with associated energies ϵ_f, (3.15) becomes

$$\Phi_{\mu_y m_y} \propto -\frac{i}{\hbar}\sum_f\{e^{i\omega_{fg}\tau}\,\langle g|\hat{m}_y|f\rangle\,\langle f|\hat{\mu}_y|g\rangle - \ e^{-i\omega_{fg}\tau}\langle g|\hat{\mu}_y|f\rangle\,\langle f|\hat{\mathrm{m}}_y|g\rangle\} \tag{3.18}$$

with $\omega_{fg} = (\epsilon_f - \epsilon_g)/\hbar$. Since the two operators m_y and μ_y commute, (3.18) becomes:

$$\Phi_{\mu_y m_y} \propto -\frac{i}{\hbar}\sum_f\langle g|\hat{\mu}_y|f\rangle\,\langle f|\hat{\mathrm{m}}_y|g\rangle\left(e^{(i\omega_{fg}-\gamma/2)\tau} - e^{(-i\omega_{fg}-\gamma/2)\tau}\right) \tag{3.19}$$

where the relaxation effect has been introduced via the $e^{-\gamma\tau/2}$ term. Eq. (3.19) shows that a system

perturbed by a magnetic field directed along *y*, oscillating at a frequency equal to the transition frequency, responds with an induced electric dipole moment oscillating along *y* at the same frequency. This latter generates the rotation of the electric field *E*, originally polarized along *x*, which is related to the passive phenomenon of optical rotatory dispersion (ORD) and to the active phenomenon of circular dichroism (CD).

At this point, to explain the active phenomenon, one has to consider the susceptibility function and in particular, the Fourier transform of the product between (3.18) and the Heaviside function $\Gamma(\tau)$, which is needed in order to make explicit that there is no response before the perturbation: $\Gamma(\tau) = 0$ for $\tau < 0$ and $\Gamma(\tau) = 1$ for $\tau \geq 0$. The Fourier Transform of $\Gamma(\tau)$ has a real and an imaginary term, while the FT of the response function is purely imaginary. The convolution of the two FTs will thus give

$$\chi_{\mu_y m_y} \propto \frac{1}{\hbar} \sum_f R_{fg} \left[\frac{\omega_{fg} + \omega - i\gamma/2}{\left(\omega_{fg} + \omega\right)^2 + \gamma^2/4} + \frac{\omega_{fg} - \omega + i\gamma/2}{\left(\omega_{fg} - \omega\right)^2 + \gamma^2/4} \right] \tag{3.20}$$

where ω_{fg} is the frequency of the transition from the state $|g\rangle$ to $|f\rangle$, ω is the frequency of the oscillating field and R_{fg} is named the rotational strength and represents the probability of the transition:

$$R_{fg} = \langle g|\hat{\mu}_y|f\rangle \langle f|\hat{m}_y|g\rangle \tag{3.21}$$

This rotational strength is the value responsible for the intensity of the CD signal.

It is now possible to take into account some considerations about the rotational strength based on group theory. In the hypothesis of a totally symmetric ground state $|g\rangle$, the rotational strength can be non-vanishing only if both the electric and the magnetic dipole moment operators, $\hat{\mu}$ and $\hat{m}$, transform as the final state $|f\rangle$. As a consequence, μ and m must transform as the same irreducible representation. The vector μ is polar and its components transform as *x*, *y* and *z*, respectively, i.e. they change sign upon reflection with respect to an orthogonal plane, while staying unvaried under reflection with respect to a plane containing the component itself. On the other hand, m is an axial vector whose components have the same symmetry properties as the (R_x, R_y, R_z) vector, i.e. they stay unvaried upon reflection with respect to orthogonal planes but change sign upon reflection with respect to a plane containing the vector. The conclusion is that μ and m can transform as the same irreducible

representation only if the molecule belongs to a symmetry point group without any improper axis of rotation, including any reflection plane and the inversion center. In other words, only chiral molecules can have optical activity.

As explained in the previous section, the real part of the susceptibility function is associated with the passive phenomena, while the imaginary part is linked to the active phenomena. In the present case, the response function associated with the rotatory power (3.14) was calculated and, since it is a passive property, it is linked to the real part of the susceptibility. In addition, the CD signal is described by the difference between the extinction coefficient of the left and the right circularly polarized light and the associated property is the molar ellipticity,

$$\varphi = \frac{\pi}{\lambda}(n_l - n_r) \tag{3.22}$$

$$\Psi = \frac{1}{4}(\epsilon_l - \epsilon_r)d \tag{3.23}$$

where φ and Ψ are the rotatory power and the molar ellipticity, respectively, n_l and n_r are the index of refraction of the left and right component of the circularly polarized radiation and ϵ_l and ϵ_r are the extinction coefficients of the two components. The correlation between the rotator power and the molar ellipticity can now be demonstrated. To do this, one has to introduce the complex index of refraction:

$$\boldsymbol{n} = n(1 - i\kappa) \tag{3.24}$$

The intensity of the light travelling in the medium will diminish exponentially with the factor $e^{-2\pi\nu n\kappa\frac{z}{c}}$:

$$I = I_0 e^{-2\pi\nu n\kappa\frac{z}{c}} \tag{3.25}$$

Now, if one introduces the extinction coefficient, ϵ,

$$\epsilon = \frac{2\pi\nu n\kappa}{c} = \frac{4\pi n\kappa}{\lambda} \tag{3.26}$$

from equations (3.22), (3.23), (3.24) and (3.26), the following equation is obtained:

$$(\varphi - i\varphi') = \frac{\pi}{\lambda}(\boldsymbol{n_l} - \boldsymbol{n_r}) \tag{3.27}$$

where $\frac{\Psi}{d} = \varphi'$ and $\varphi - i\varphi'$is called the complex rotatory power. The conclusion is that the complex rotatory power is correlated to the complex index of refraction, in the same way that ordinary rotatory power is correlated to the real index, and the imaginary part of the rotatory power is proportional to the difference between the extinction coefficient of the left and the right component of the circularly polarized light. Thus, the active phenomenon of CD is linked to the imaginary part of the susceptibility function (eq.(3.20)) through the molar ellipticity,

$$\frac{\Psi}{d} = \varphi' \propto \sum_f R_{fg} \frac{\omega^3 \gamma}{\left[\left(\omega_{fg}^2 - \omega^2\right)^2 + \omega^2\gamma^2\right]} \tag{3.28}$$

and the CD signal, $\Delta\epsilon$, will be:

$$\Delta\epsilon(\omega) \propto R_{fg} \frac{\omega^2\gamma}{\left(\omega_{fg} - \omega\right)^2 + \gamma^2} \tag{3.29}$$

where the strength of each transition (the area under each peak) depends on the value of R_{fg} and so on the product between the matrix elements of the transition electric and magnetic dipole moment of the $g \rightarrow f$ transition. The equation above describes a Lorentzian band shape, centered at ω_{fg} with full width at half maximum (FWHM) equal to γ.[9]

3.4 Exciton Coupled CD[40]

Consider a system composed of two nonchiral chromophores having equal or similar frequency of absorption, ω_{fg}. If the two chromophores are not coupled, then the absorption spectrum of the molecule will simply be the sum of the absorption spectra of the two isolated chromophores. On the other hand, if the chromophores are close enough in space, the system's exited states will be mixed states of the original chromophores and the absorption spectra will show a blue or a red shift, based on

the combination between the latter, so that

$$\omega_{abs} = \omega_{fg} \pm J \tag{3.30}$$

where J represents the contribution due to the interaction of the dipole moment of the two chromophores:

$$J = < C_1^* C_2 |\, H_{int} | C_1 C_2^* > \tag{3.31}$$

Here C_1 and C_2 refer to the two chromophores and the terms inside the bra-ket are the two mixed states in which only one chromophore is exited at a given time (*). Depending on the orientation of the two moieties and thus on the symmetry of the system, this can present chiral behaviour and thus optical activity.

Now, consider the origin of the CD spectrum of a molecule like Formicamycin, where two similar nonchiral aromatic moieties are close in space and bound together, generating a chiral axis. The form of H_{int} will be the same as given in (3.13), but this time, the μ and m_B values will be the sum of the components belonging to the two chromophores:

$$\mu = \mu_1 + \mu_2 \qquad\qquad m_B = m_1 + m_2 \tag{3.32a), (3.32b}$$

Remembering that the rotational strengths depend on the product between the electric and the magnetic transition dipole moments and that the two isolated chromophores do not show optical activity, R_{fg} will be proportional to the mixed states, $\mu_1 m_2$ and $\mu_2 m_1$. Setting the origin of the reference Cartesian system on one chromophore, it is possible to express R_{fg} as proportional to:

$$R_{fg} \propto \langle g|m_1|f\rangle\langle f|\mu_2|g\rangle \tag{3.33}$$

The definition of magnetic moment gives $\vec{m}_1 = e\vec{R} \times \vec{p}_1$, where $\vec{p}_1$ is the momentum associated with the electron of chromophore 1 , $\vec{R}$ is the distance between the two chromophores and e is the charge of the electron. One can then rewrite (3.33) to give:

$$R_{fg} \propto \vec{R} \cdot \langle g|\vec{p}_1|f\rangle \times \langle f|\vec{\mu}_2|g\rangle \tag{3.34}$$

where the triple product property $\vec{R} \times \vec{p}_1 \cdot \vec{\mu}_2 = \vec{R} \cdot \vec{p}_1 \times \vec{\mu}_2$ has been applied. Finally, from the interaction between radiation and matter, the momentum is correlated to the electric dipole through the commutator:

$$\vec{p} = i\frac{m}{e\hbar}[H, \vec{\mu}_e] \tag{3.35}$$

R_{fg} becomes:

$$R_{fg} \propto \vec{R} \cdot \langle g|H\vec{\mu}_1 - \vec{\mu}_1 H|\, f\rangle \times \langle f|\vec{\mu}_2|g\rangle \tag{3.36}$$

and

$$R_{fg} \propto \vec{R} \cdot \langle g|\vec{\mu}_1|f\rangle \times \langle f|\vec{\mu}_2|g\rangle \tag{3.37}$$

where the eigenvalue equation, H|g> = ϵ_g |g>, has been used.

From this result, it is possible say that in a coupled system, the CD signal depends on the imaginary part of the susceptibility function, as in the previous case, (3.29), but now, R_{fg}depends only on the interaction of the transition dipole moments of the two chromophores, thus on the angle and distance between them. In particular, when the angle is equal to zero or 90°, the molecule will not be chiral and so no CD signal will be present. Furthermore, opposite enantiomers will have opposite sign for the Cotton effect band; Δϵ(ω) will be equal to zero at ω_{fg} and the positive and negative maxima will have a frequency equal to $\omega_{abs} = \omega_{fg} \pm J$, where $2J$ is called the Davydov splitting.

4. Applied Methodology and Results

As introduced in Chapter 1, in order to calculate an ECD spectrum, four main steps are needed: the generation of an ensemble of conformers, geometry optimization and the calculation of the vibrational frequencies for each conformer, the calculation of the rotational strengths, and the generation of the spectrum. In this chapter, the computational details applied in this work and the results, are reported.

The work can be divided into two parts: in the first part, the process was applied to Formicamycin using setup and software according to a previous study,[12] in order to verify the presence of a fingerprint area that can be used to assign the absolute configuration of the chiral axis; in the second part, the entire process was computationally optimized involving semi-empirical and DFT calculations, automated using a Python script.[41,42] The two processes will be described separately and the automated script will be described in the next chapter.

4.1 Assignment of the Chiral Axis Configuration

4.1.1 Conformational Analysis

The conformational analysis of Formicamycin was carried out using the MacroModel[33] ConformationalSearch option available in the Maestro11 software package (described in Chapter 2).[34] After drawing the molecule with the correct stereochemistry (C10-C19 = R, R) of the two chiral centers, two different force fields, namely MMFFs[28] and OPLS3,[29] were tested in the generation of the conformers in three different solvents (dielectric constant model), hexane ($\varepsilon = 1.9$), 2-methoxyethanol ($\varepsilon = 17.2$) and water ($\varepsilon = 80.1$). The ECD spectrum in methanol ($\varepsilon = 33.0$) were calculated, using the same methodology, by others[43]. The choice of the solvents was done to have an increasing scale of polarity, in order to be able to evaluate the solvent effect in the final spectrum. In all the calculations, the conformers were obtained exploring an energy window of 21 kcal/mol, in accordance with previous research.[12]

Through the Maestro software it is possible to export each conformation as a *.pdb* file that contains, in addition to the 3D coordinates of each atom, also the potential energy of the conformation expressed in kcal/mol and other useful chemical information. The number of structures obtained varied in the 6 combinations of FF and solvent, but in every case direct visualization of the molecules showed that the

main structural differences involved the orientation of the A ring and the ‘up’ or ‘down’ orientation of the hydroxyl group, the ‘up’ or ‘down’ orientation of both the hydroxyl groups on C10 and C15, and the angle of the L shape of the aromatic scaffold (see Figure 16).

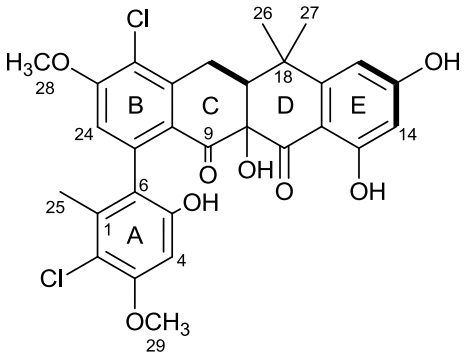


Figure 16. Chemical structure of Formicamycin.

The two limiting configurations of the chiral axis between the A and B rings can now be defined. Based on the orientation of the A ring, one can define a ‘hydroxyl configuration’ (R), where the hydroxyl group on C5 points inward and therefore ‘inside’ the scaffold’s L shape, otherwise a ‘methyl configuration’ (S), when the methyl group points inward (see Figure 17).

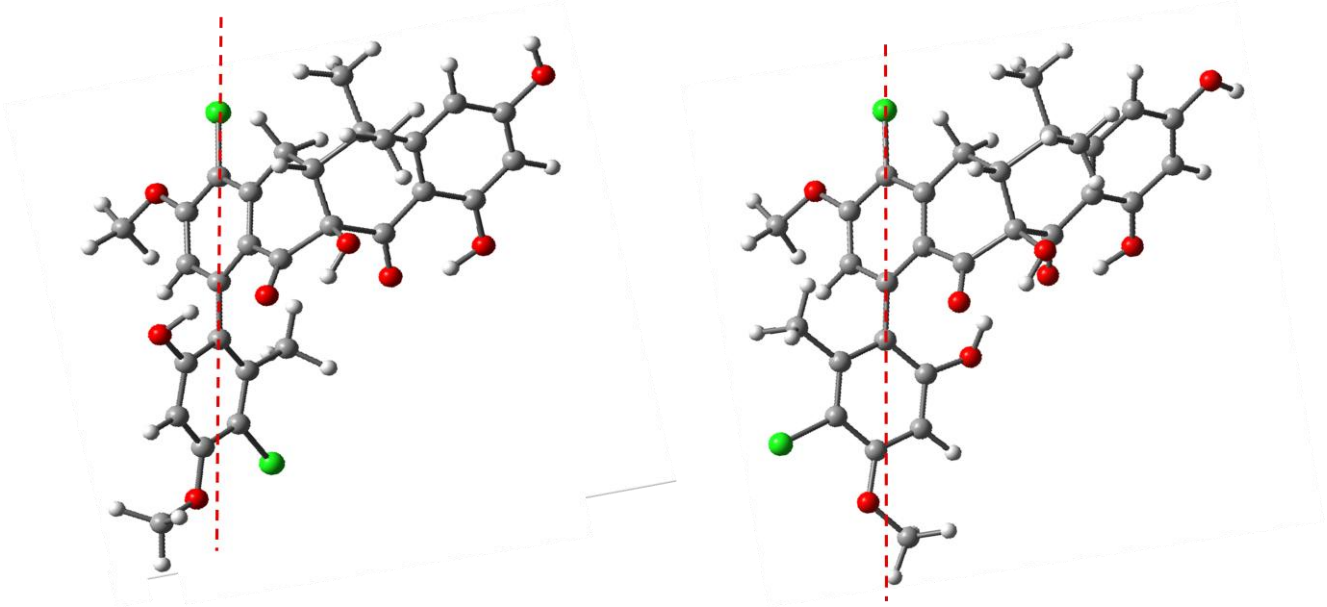


Figure 17. Chiral axis of Formicamycin in the two limiting configurations. On the left: the methyl configuration (S); on the right: the hydroxyl configuration (R).

The output ensembles of structures showed a good balance between the two configurations (see Table 1 and Table 3).

Table 1. Number of structures obtain from the conformational search in the three different solvent and involving the two different force fields.

n-Hexane (1.88)		2-Methoxyethanol (17.2)		Water (80.1)	
MMFFs	OPLS3	MMFFs	OPLS3	MMFFs	OPLS3
12	**147**	**38**	**348**	**39**	**126**

4.1.2 Optimization and Frequency Calculation

The next steps of the process are the optimization of the geometries, using a QM potential, and the calculation of the vibrational frequencies for all the conformers obtain from the conformational analysis. In order to do this, the Gaussian09 software[44] was used. This software accepts *.com* files as input, so all the *.pdb* files were converted into *.com* files using a Python script (see Appendix A). The two calculations (optimization and frequency) were performed separately and not using the "*opt+freq"* option available in Gaussian. This is due to the great number of structures obtained from the previous step that would make the *opt+freq* calculation on each, prohibitively expensive. Rather, after the optimization, an energy-based selection was applied in order to exclude the conformers with a low Boltzmann weighted contribution to the entire ensemble and exclude the structures that were energetically equivalent. In particular, all the conformers contributing less than 1% (with a difference from the most stable conformer of more than 4×10^{-3} Hartree, or 2.5 kcal/mol) have been excluded from the ensemble and those with an energy difference less than 1×10^{-6} Hartree were defined as equivalent (see Appendix A). The frequency calculation, in order to calculate the ZPE, was then performed only on unique and statistically relevant conformations, making the process less computationally demanding (Table 2). All QM calculations were done using the PBE1PBE (PBE0) hybrid exchange-correlation functional and the def2-TZVP basis set.[45] The different solvents were taken into account using the SMD implicit solvent model.[46]

Table 2. Number of structures before and after the selection of the conformers in the three solvent, done using the two different force fields.

n-Hexane (1.88)		2-Methoxyethanol (17.2)		Water (80.1)	
MMFFs	OPLS3	MMFFs	OPLS3	MMFFs	OPLS3
12	**147**	**38**	**348**	**39**	**126**
		Conformer Selection			
3	**6**	**10**	**19**	**9**	**11**

4.1.3 TDDFT Calculations

The transition energies were calculated using Gaussian09 with the same setup as above. The number of lowest electronic transitions to calculate was set to 30, in agreement with previous studies,[12] and the rotational strength of each transition was calculated. The output files of a Gaussian calculation are in the *.log* format, so after each calculation, a conversion from *.log* to *.com* was done, in order to facilitate the next step in the calculation.

4.1.4 Generation of the Spectrum

The Boltzmann-weighted spectrum was generated using the GaussSum software.[47] GaussSum is free and open source software that is able to generate UV-vis and ECD spectra of a single molecule starting from the electronic transition energies and the associated transition dipole moments and/or rotational strengths, calculating the maximum absorption for each transition (see Appendix A). In order to generate a Boltzmann-weighted spectrum, the Boltzmann factor of each conformer was calculated based on the potential energy, as shown in (4.1)

$$\frac{N_i}{N} = \frac{e^{\frac{-(\epsilon_i - \epsilon_0)}{RT}}}{\sum_i e^{\frac{-(\epsilon_i - \epsilon_0)}{RT}}} \tag{4.1}$$

where ϵ_i is the energy of the i-th conformer in cal mol^{-1}, R is the gas constant (1.9872 cal mol^{-1} K^{-1}) and T is the absolute temperature (set to 298.15 K). All 30 electronic transition energies and the relative rotational strengths were extracted. With each conformer, one can now associate a 30-element vector in which each element corresponds to the energy of a specific transition. Each element in the vector can now be multiplied by the Boltzmann factor of the relevant conformer and then summed so that the result is a single, weighted 30-element vector (Figure 18).

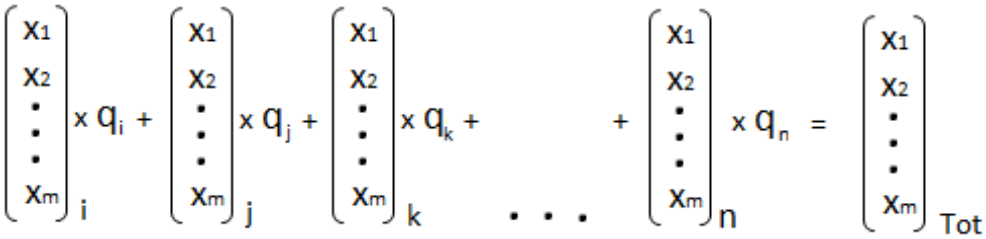


Figure 18. Vector representation of the weighting process. The indexes i,j…n represent the different conformers; $x_1 \ldots x_m$ are the energies associated with each transition and q is the Boltzmann coefficient for a specific conformer.

The final vector represents a "virtual molecule" describing the entire weighted ensemble. This "virtual molecule" was used as input for the GaussSum software and the absorption values plotted against the wavelength using the matplotlib package[48] in order to visualize the spectrum. The spectra were calculated from 200 nm to 500 nm in order to highlight the fingerprint area and the half-bandwidth was set to 0.25 eV (Figure 19). All these processes were managed using Python and, in addition to the final spectrum, a superimposition of all the single spectra was also generated by the final script (see Figure 19).

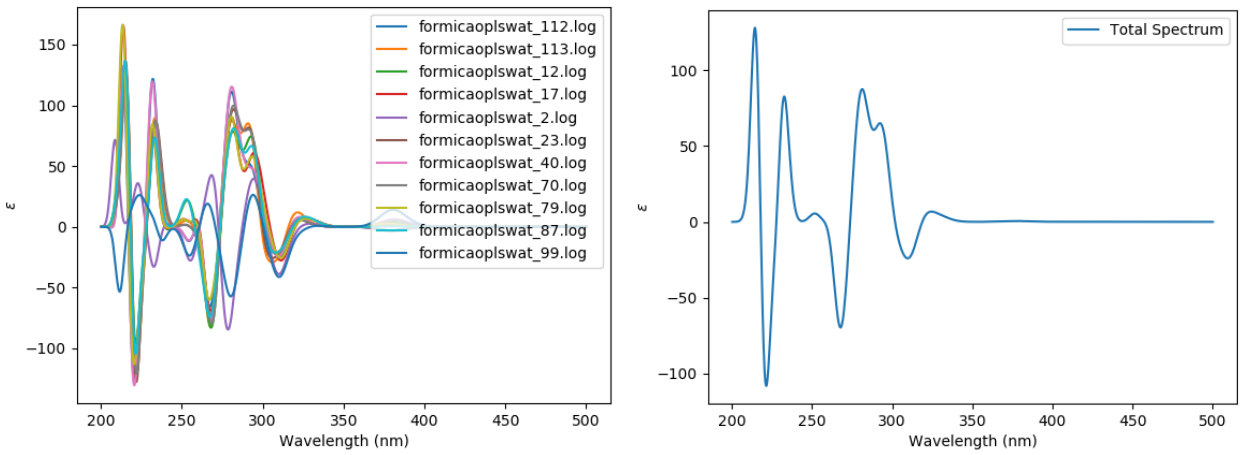


Figure 19. On the left: superposition of all the single conformer ECD spectra in water, using the OPLS3 force field for conformer generation. On the right: the final ECD spectrum, obtained by summing over all the spectra.

The partial spectra generated only by structures with a specific configuration (hydroxyl or methyl), were also calculated. Table 3 shows the distribution of the two configurations is reported.

Table 3. Number of conformers involved in the generation of the spectra and the distribution of configurations.

	Total Unique Conformers	Methyl	Hydroxyl
n-Hexane (1.88)			
MMFFs	3	1	2
OPLS3	6	4	2
2-Methoxyethanol (17.2)			
MMFFs	10	4	6
OPLS3	19	7	12
Water (80.1)			
MMFFs	9	4	5
OPLS3	11	6	5

4.2 Results

In this section, a summary of the obtained spectra is reported. For each FF/solvent, three spectra were calculated: a total spectrum of all the conformers, a spectrum generated only by the conformers having the hydroxyl configuration and one with only the methyl configuration. The energetic distribution of the conformations giving the spectra is also reported.

4.2.1 Hexane/MMFFs

The Boltzmann distributions of the conformers obtained with MMFFs in hexane are reported in Table 4. The associated spectra are shown in Figure 20.

Table 4., Boltzmann factor and distribution of the structures with the same specific configuration, obtained in hexane and using the MMFFs force field.

Conformer	Boltzmann factor	Distribution OH	Distribution Methyl
3	0,666	0,686	
9	0,305	0,314	
8	0,029		1,00

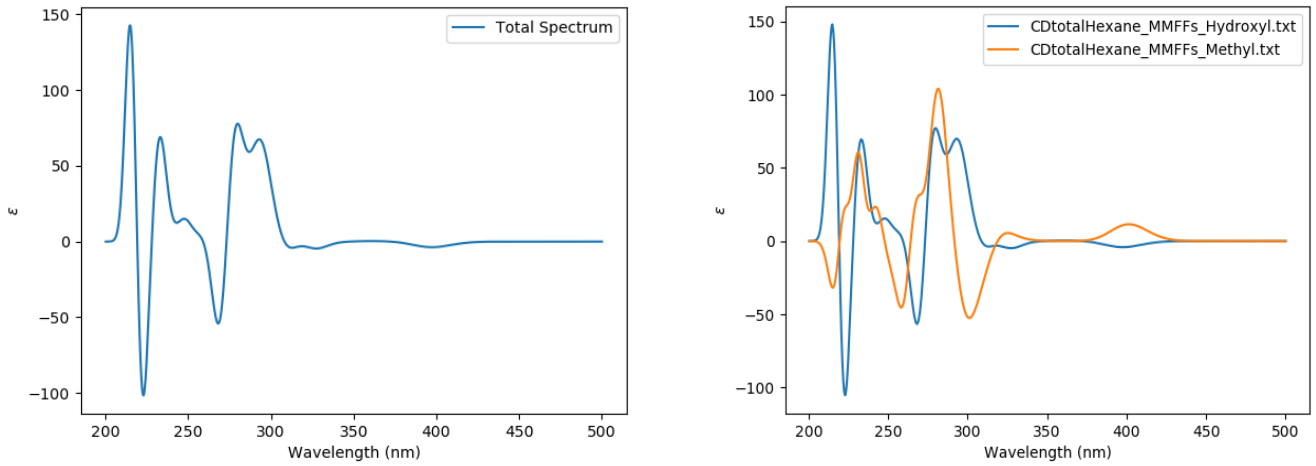


Figure 20. On the left: CD spectrum generated by all the conformers in Table 4. On the Right: CD spectrum generated by the structures in Table 4 with a hydroxyl (blue) configuration and the single structure with a methyl (orange) configuration.

4.2.2 Hexane/OPLS3

The Boltzmann distributions of the conformers obtained with OPLS3 in hexane are reported in Table 5. The associated spectra are shown in Figure 21.

Table 5. Boltzmann factor and distribution of the structures with the same specific configuration, obtained in hexane and using the OPLS3 force field.

Conformer	Boltzmann factor	Distribution OH	Distribution Methyl
34	0,499	0,686	
12	0,228	0,314	
5	0,100		0,468
21	0,0481		0,255
28	0,0440		0,206
38	0,0215		0,100

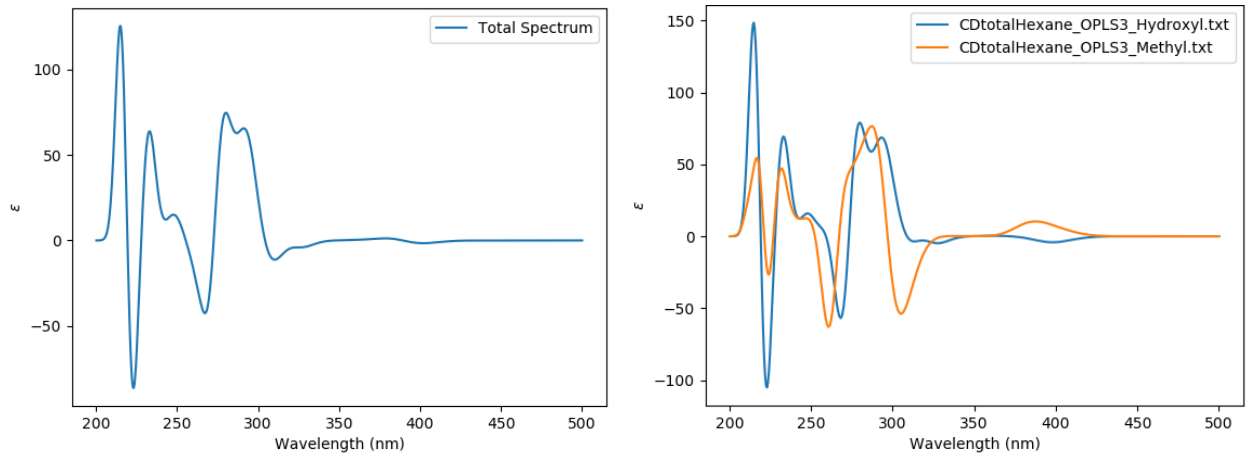


Figure 21. On the left: CD spectrum generated by all the conformers in Table 5. On the Right: CD spectrum generated by the structures in Table 5 with a hydroxyl (blue) configuration and methyl (orange) configuration

4.2.3 2-Methoxyethanol/MMFFs

The Boltzmann distributions of the conformers obtained with MMFFs in 2-methoxyethanol are reported in Table 6. The associated spectra are shown in Figure 22.

Table 6. Boltzmann factor and distribution of the structures with the same specific configuration, obtained in methoxyethanol and using the MMFFs force field.

Conformer	Boltzmann factor	Distribution OH	Distribution Methyl
9	0,131		0,288
18	0,122		0,269
29	0,108	0,199	
10	0,105	0,192	
31	0,104		0,229
4	0,0976		0,214
30	0,0889	0,163	
12	0,0860	0,158	
20	0,0856	0,157	
22	0,0716	0,131	

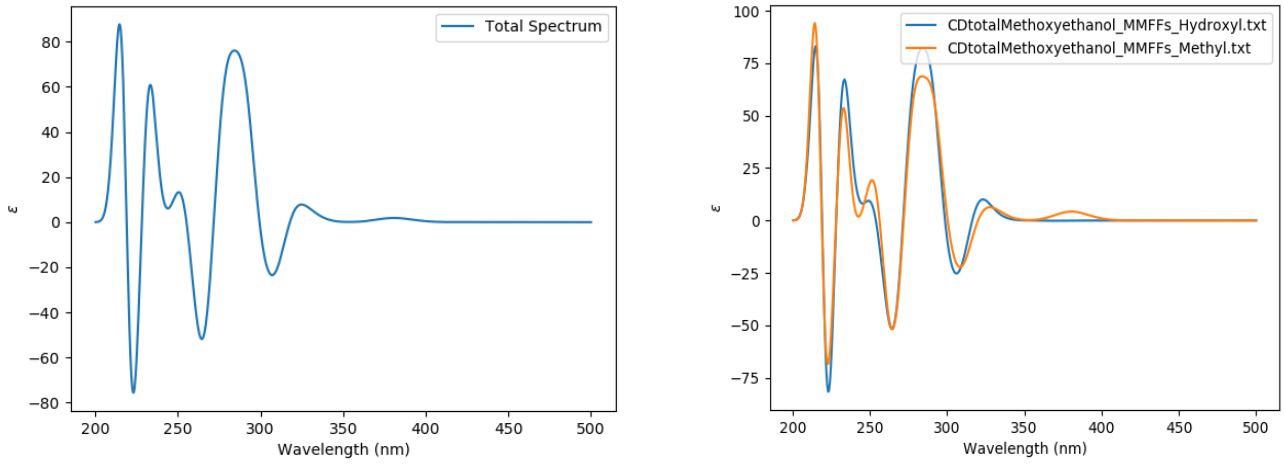


Figure 22. On the left: CD spectrum generated by all the conformers in Table 6. On the Right: CD spectrum generated by the structures in Table 6 with a hydroxyl (blue) and methyl (orange) configuration.

4.2.4 2-Methoxyethanol/OPLS3

The Boltzmann distributions of the conformers obtained with OPLS3 in 2-methoxyethanol are reported in Table 7. The associated spectra are shown in Figure 23.

Table 7. Boltzmann factor and distribution of the structures with the same specific configuration, obtained in methoxyethanol and using the OPLS3 force field

Conformer	Boltzmann factor	Distribution OH	Distribution Methyl
318	0,121		0,359
55	0,100	0,151	
82	0,0967	0,146	
307	0,0965		0,286
127	0,0903		0,267
142	0,0822	0,124	
272	0,0796	0,120	
254	0,0795	0,120	
221	0,0792	0,120	
242	0,0662	0,100	
149	0,0168	0,0254	
156	0,0164	0,0247	
43	0,0135		0,0400
168	0,0132	0,0199	
151	0,0122	0,0185	
163	0,0103	0,0156	
174	0,00991	0,0150	
44	0,00899		0,0266
204	0,00716		0,0212

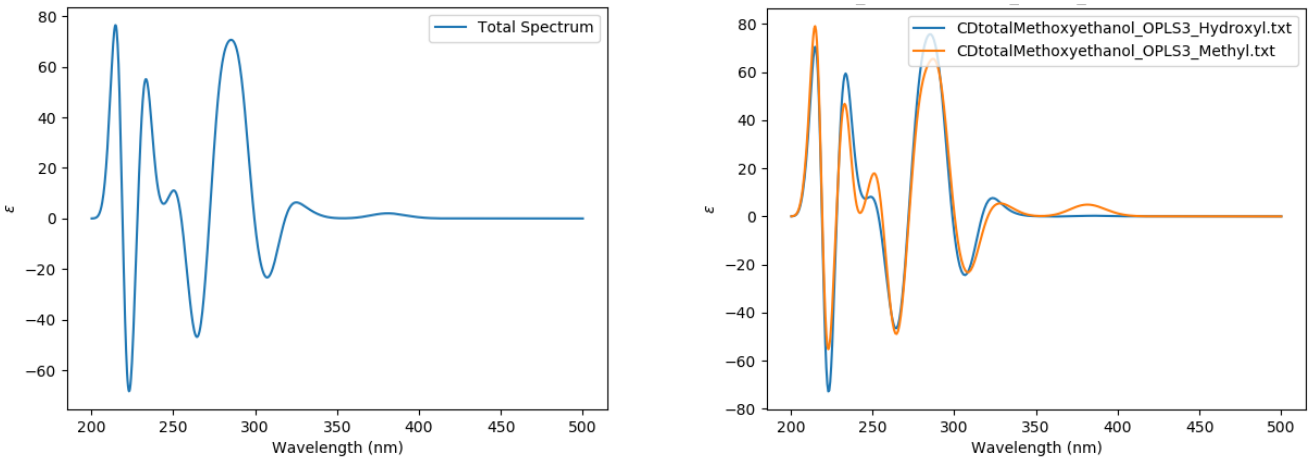


Figure 23. On the left: CD spectrum generated by all the conformers in Table 7. On the Right: CD spectrum generated by the structures in Table 7 with a hydroxyl (blue) and methyl (orange) configuration.

4.2.5 Water/MMFFs

The Boltzmann distributions of the conformers obtained with MMFFs in water are reported in Table 8. The associated spectra are shown in Figure 24.

Table 8. Boltzmann factor and distribution of the structures with the same specific configuration, obtained in water and using the MMFFs force field.

Conformer	Boltzmann factor	Distribution OH	Distribution Methyl
5	0,135		0,262
13	0,130		0,253
6	0,126		0,246
14	0,123		0,239
37	0,111	0,227	
16	0,104	0,214	
38	0,0929	0,191	
22	0,0918	0,189	
27	0,0871	0,179	

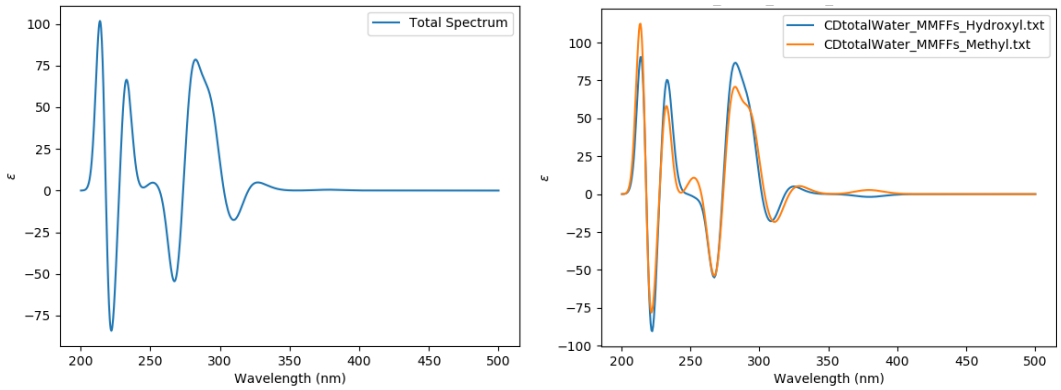


Figure 24. On the left: CD spectrum generated by all the conformers in Table 8. On the Right: CD spectrum generated by the structures in Table 8 with a hydroxyl (blue) and methyl (orange) configuration.

4.2.6 Water/OPLS3

The Boltzmann distributions of the conformers obtained with MMFFs in hexane are reported in Table 9. The associated spectra are shown in Figure 25.

Table 9. Boltzmann factor and distribution of the structures with the same specific configuration, obtained in water and using the OPLS3 force field

Conformer	Boltzmann factor	Distribution OH	Distribution Methyl
79	0,129		0,245
87	0,125		0,237
17	0,121		0,230
12	0,118		0,224
23	0,106	0,227	
70	0,100	0,214	
113	0,0893	0,191	
112	0,0883	0,189	
40	0,0837	0,179	
99	0,0181		0,0343
2	0,0152		0,0289

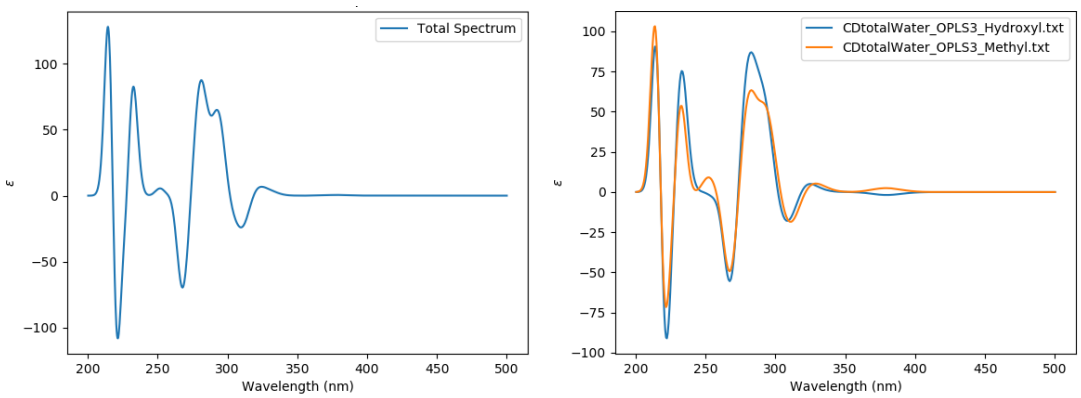


Figure 25. On the left: CD spectrum generated by all the conformers in Table 9. On the Right: CD spectrum generated by the structures in Table 9 with a hydroxyl (blue) and methyl (orange) configuration.

4.2.7 Methanol/MMFFs

The Boltzmann distributions of the conformers obtained previously[43], with MMFFs in methanol, are reported in Table 10. The associated spectra are shown in Figure 26.

Table 10. Boltzmann factor and distribution of the structures with the same specific configuration, obtained in methanol and using the MMFFs Force Field.

Conformer	Boltzmann factor	Distribution OH	Distribution Methyl
13	0,120		0,260
13	0,115		0,249
32	0,114		0,247
38	0,113		0,245
37	0,110	0,222	
4	0,110	0,222	
6	0,097	0,196	
12	0,095	0,192	
14	0,083	0,168	

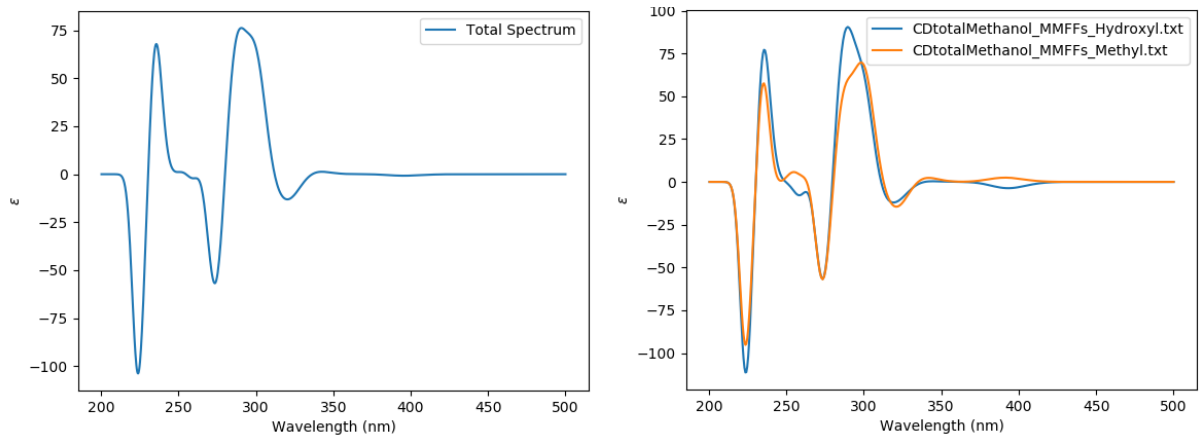


Figure 26. On the left: CD spectrum generated by all the conformers in Table 10. On the Right: CD spectrum generated by the structures in Table 10 with a hydroxyl (blue) and methyl (orange) configuration.

4.3 Discussion

4.3.1 Chiral Axis

In all the spectra, two relevant areas can be identified in the fingerprint region, to distinguish between the two configurations. Between 250 and 300 nm, a shoulder is seen when the molecule is in the methyl configuration, which is not present for the hydroxyl configuration; in addition, the sign of the broad peak around 380 nm is opposite for the two configurations. It is, however, very difficult to assign each peak or pattern to a specific transition. In particular, the shoulder between 250 nm and 300 nm presents a different pattern depending on both the orientation of all the hydroxyl groups (on the A and E rings, as well as in-between the C and D rings) and the configuration of the A ring, so that the latter is hard to identify. In addition, the peak at 380 nm does not always have the same sign for a given configuration in each of the individual spectra, however, in the final weighted spectrum, the sign of the peak is unique to the configuration (Figure 19, left). This could be due to an overlap of signals associated with other moieties in the molecule. Nevertheless, through simple comparison of the overall spectra of the hydroxyl and methyl configurations, it is possible to identify the chirality without associating every individual transition to its own peak.

Finally, the experimental spectrum in methanol matches better with the calculated spectrum corresponding to the hydroxyl configuration (R), as shown in Figure 27. The shoulder at 260 nm is absent and the broad peak at 390 nm is negative. The configuration therefore, surprisingly, appears to be the opposite of that of its precursor Fasamycin (see Chapter 1). This would imply that during the proposed biosynthesis, there must be an inversion of configuration between Fasamycin and

Formicamycin, or alternatively the two structures follow two different biosynthetic pathways.

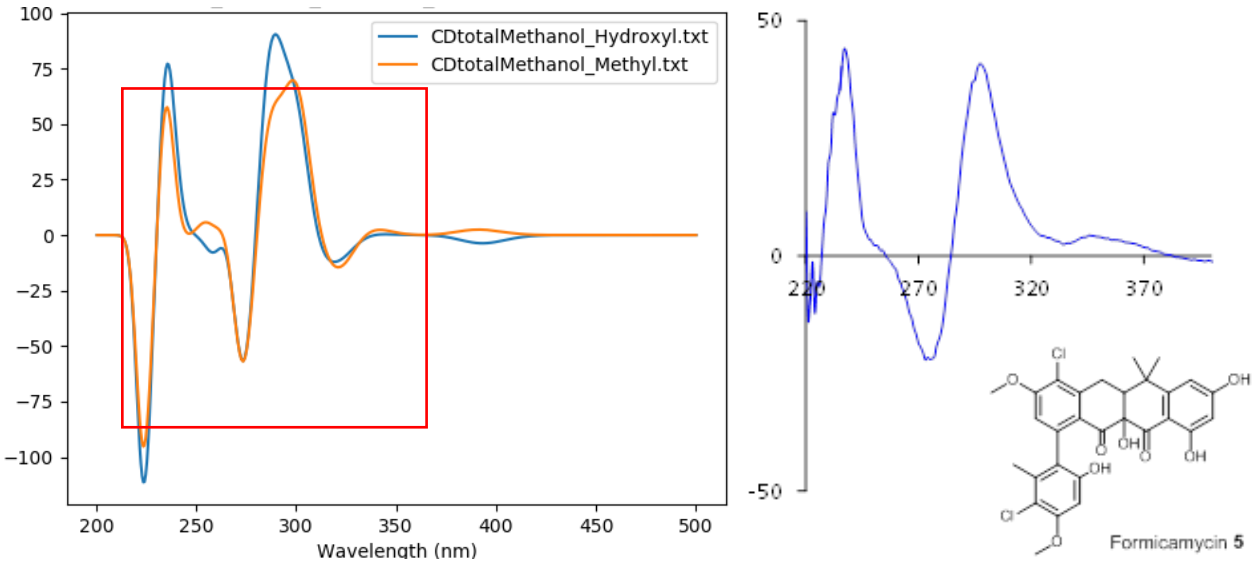


Figure 27. Comparison between the experimental CD spectrum of Formicamycin in methanol (on the right) and the calculated spectra in methanol of the two configurations: R-configuration (hydroxyl) in blue and S-configuration (methyl) in orange. The shape of the shoulder at 260 nm and the sign of the peak at 390 nm shows a better match with the R-configuration.

4.3.2 Effect of Solvent

It has already been highlighted that it is not possible to define a direct correlation between the polarity of the solvent and the shape of the ECD spectrum, which mostly depends on the geometric features of the molecule. However, the change in solvent can generate a completely different ensemble of conformers, with respect to both the geometrical features and distribution of conformations. The results show that the final ECD spectra indeed show significant differences in different solvents (Figure 28). Analysing the spectra of the two configurations calculated in different solvents, one can see that the variation in both the shift and intensity of the transitions is different for the two configurations (Figure 29).

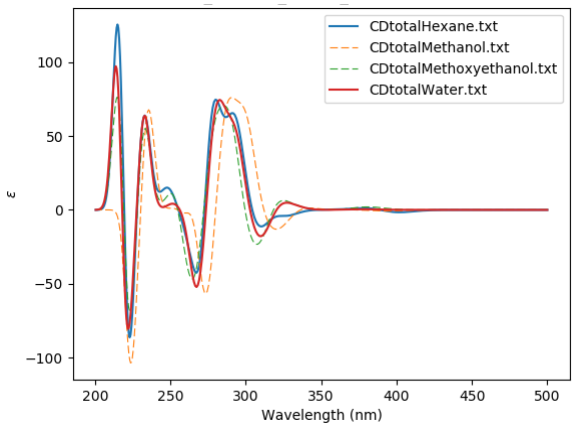


Figure 28. Comparison between the spectra calculated in the four different solvents, with conformers obtained using the OPLS3 force field.

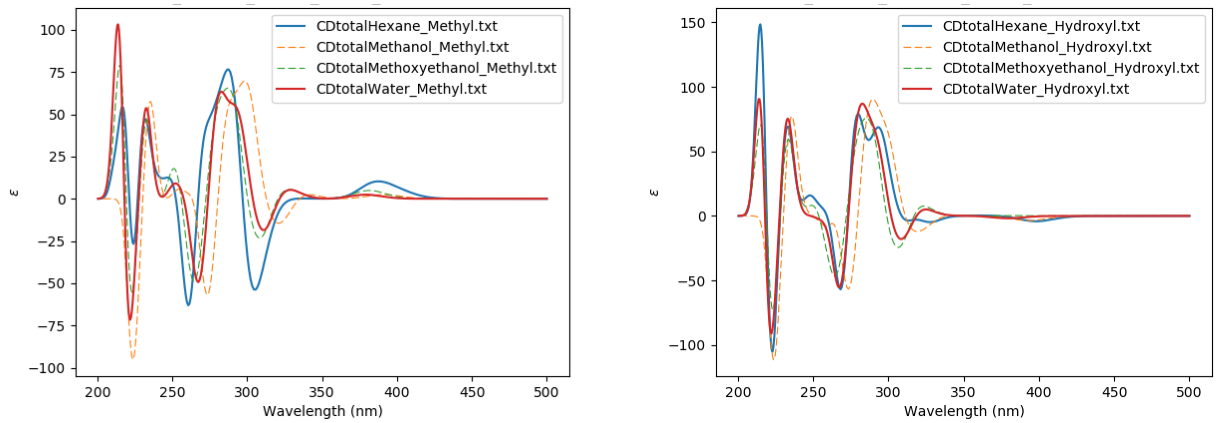


Figure 29. CD spectra of the methyl (left) and hydroxyl (right) configuration in the four different solvents, calculated with conformers from the OPLS3 force field.

For example, comparing hexane and methanol (Figure 30), the change in solvent induces greater differences for the molecule in the methyl configuration than in the hydroxyl configuration.

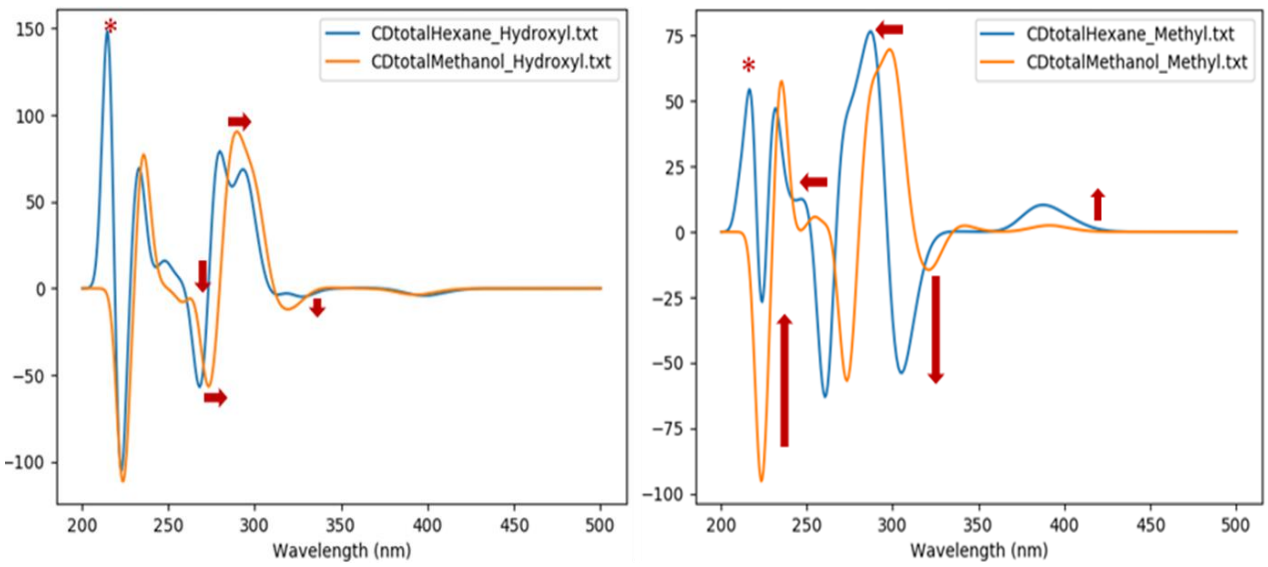


Figure 30. Comparison between the ECD spectra calculated in hexane (blue) and methanol (orange) in the two configurations (hydroxyl on the left and methyl on the right). The red arrows highlight the extent of changes in the two cases.

This could be the basis for an additional approach in the assignment of the absolute configuration (AC), in which the extent of variation due to the solvent is an important parameter. Comparison between experimentally recorded spectra in two solvents could directly give information about the AC in cases where the spectrum in a particular solvent cannot sufficiently discriminate between two configurations.

4.3.3 Effect of Force Field

Looking at the number of conformers generated by the two FFs, it is clear that OPLS3 produces more conformations than MMFFs (Table 1). This makes this procedure more computationally expensive, since in the following step the DFT optimizations will involve a greater number of molecules. The distribution of the conformer energies using the two FF was compared with that obtained after optimization at the PBE0/def2-TZVP level in SMD hexane (Figure 31). The MMFFs force field has a smaller distribution of the energies and better matches the QM optimized conformations. It also gives a lower number of false negative (i.e., to be excluded eventually, as explained below) structures, compared to OPLS3, where a large number of conformers (see the yellow area in Figure 31) with a high relative energy using the FF end up within the QM cut off region.

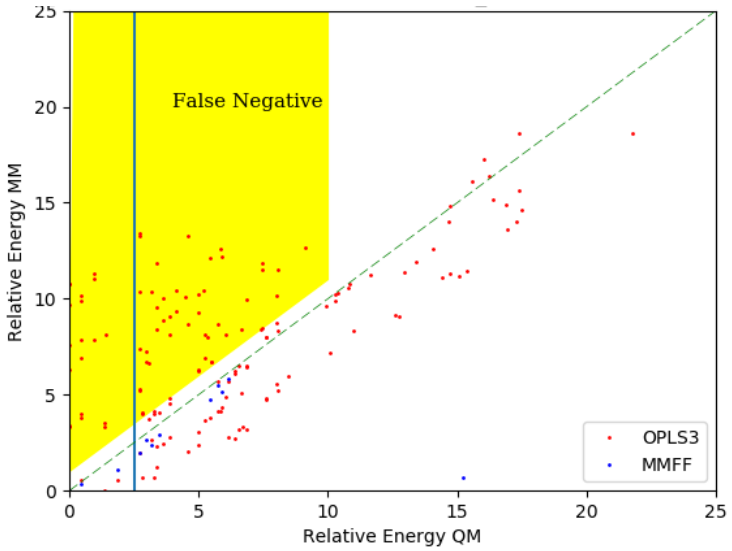


Figure 31. Comparison between the energies of the conformers obtain by the MM (y-axis) and QM (x-axis) calculation, expressed as difference from the most stable conformer in kcal/mol.

The vertical blue line in Figure 31 represents the energy limit imposed during the selection of the conformers (≈2.5 kcal/mol), so that all the structures with a QM energy higher than this value will be excluded after the optimization and represent 'wasted' computational effort. Comparing the final spectra obtained, starting from the two different force fields, a very good agreement is seen (Figure 32). The spectra of the hydroxyl and methyl configurations, generated starting from the MMFFs and OPLS3 force fields, showed a good match when water and 2-methoxyethanol were set as solvent. However, the MMFFs analysis in hexane generated only one conformer having the methyl configuration after the QM optimization (Table 3). The resultant spectrum (generated from only this one molecule) is very different from the one obtained with the OPLS3 conformational analysis used as the starting point (Figure 33).

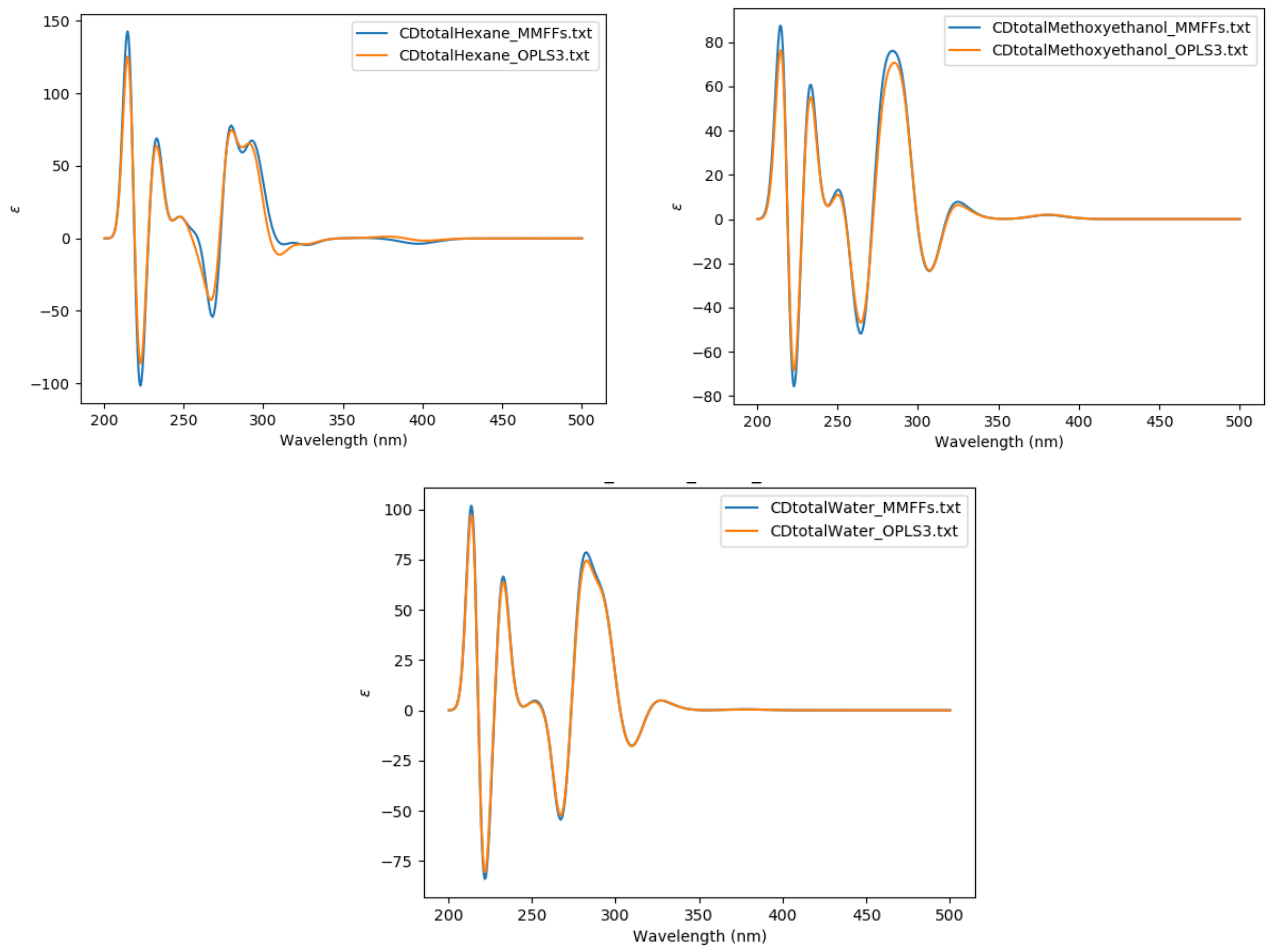


Figure 32. Comparison between the spectra of all the conformers obtained starting from the MMFFs (blue) and OPLS3 (orange) force field in hexane (upper left), methoxyethanol (upper right) and water (below).

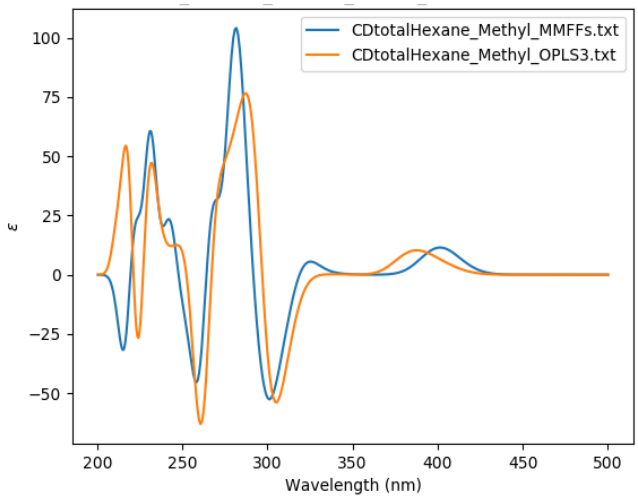


Figure 33. ECD spectra generated by the Formicamycin conformers having an S (methyl) configuration of the chiral axis, starting with MMFFs (blue) and OPLS3 (orange) in hexane.

Therefore, even though OPLS3 generates a large number of conformations, which after the QM optimization are too high in relative energy despite being within the initial cut-off, it also generates a diverse set of structures that are important in the lower energy region. It might thus not be best to use the cheaper MMFFs procedure alone, since the number of final conformations might not be enough to generate an accurate final ECD spectrum. This procedure could be improved by having a different criterion for generating conformers that is not only based on energy, but also structural diversity, such as a root-mean-square-deviation (RMSD) criterion.

4.4 Optimization of the Process

In the second part of this work, in order to reduce the computational cost and limit, as far as possible, the dependencies on commercial software, an alternative protocol was applied. In particular, the conformational analysis was performed using free and open source software packages and a semi-empirical QM method was adopted for the geometry optimization. It should be noted that Gaussian, which is a commercial package, was still retained, due to ease of use. However, there are free (either open source or with free licensing to academic institutions) packages such as Psi4[49] or ORCA[50] that provides similar functionality and can be substituted for Gaussian (with further refinement to the Python automation script).

4.4.1 Conformational Analysis

In order to substitute the Maestro software, conformer generation routines in two open source packages, OpenBabel[51] and RDKit[35], were tested , with the latter found to give the best results. The main reasons of this are a better Python implementation of the RDKit package and a more representative ensemble of structures. The OpenBabel package contains two different procedures for conformer generation: Confab[52] (a systematic approach) and a genetic algorithm (a stochastic approach). For both of these, the number of structures was not enough to obtain a good Boltzmann-weighted ECD spectrum, and the algorithms on which these methods are based do not consider the chiral axis of Formicamycin as a degree of freedom. That means that all or almost all the structures obtained from the analysis have the same hydroxyl or methyl configuration, depending on the initial configuration. Furthermore, a previous study,[32] found that RDKit's results are comparable to

commercial, proprietary software.

RDKit accepts the input molecule as a chiral SMILES string and can set the number of conformers to be generated, which depends on the number of rotatable bonds in the molecule. The starting structure and all the generated conformers were optimized using the available MMFF94 force field and structurally equivalent structures were excluded based on a RMSD criterion of 2.0 Å. The entire process was carried out three times for Formicamycin, generating 200, 100 and 50 starting structures in hexane. In all cases, the final spectrum was similar to one another as well as those obtained with the previous methodology. The ideal minimum number would depend on the overall conformational flexibility (number of rotatable bonds) of the molecule, but in this case, 50 was chosen as an acceptable number of starting structures.

It is important to emphasize that, due to the stochastic nature of the method, the ensemble of generated conformers will typically be different each time that the program is run, however, the successive optimization steps homogenize the results thanks to the good representativity of the generated ensemble. The distribution of the two configurations obtained are reported in Table 11.

Table 11. Configurational distribution associated with the generation of 50, 100 and 200 starting structures, using RDKit.

N° of generated conformers	Hydroxyl	Methyl
50	24	26
100	55	45
200	135	65

The output structures were saved as a single *.sdf* file. From this single file, each structure was extracted and converted into *.com* format before the next step. All these steps were automated using Python (see Appendix A).

4.4.2 Semi-Empirical Optimization

Since the most expensive step of the process is the QM optimization of all the structures obtained from the conformational analysis, a semi-empirical method was trialled as an alternative before the conformational selection and the full DFT calculation. The optimization was again done using Gaussian09 and the PM6 and AM1 methods (see Chapter 2) with the SMD solvent model. Selection of conformers was done using the same criteria used in the first part (4×10^{-3} Hartree as limit and 1×10^{-6} Hartree to define the same structure). The number of conformations obtained after PM6 geometry optimization are summarized in Table 12.

Table 12. Number of conformers and their distribution after the first selection, which used a PM6 geometry optimization.

N° of starting conformers	1° selection	N° conformers after 1° selection	Hydroxyl	Methyl
50		9	4	5
100		10	5	5
200		11	5	6

4.4.3 Optimization and Vibrational Frequency Calculation

All the conformers retained after the conformational selection, were then optimized again using a full QM potential and the vibrational frequencies were calculated. These calculations were performed using Gaussian09 and the same setup as before (i.e. PBE1PBE/def2-TZVP and the SMD solvent model). This was followed by another conformer selection, similar to the previous one, and the results are reported in Table 13.

Table 13. Number of conformers and their distribution after the second selection, which used a DFT geometry optimization.

N° of conformers from the SE optimization	2° selection	N° conformers after 2° selection	Hydroxyl	Methyl
9		5	2	3
10		5	2	3
11		5	2	3

Finally, a TDDFT calculation was done on the remaining structures and the generation of the spectra was done in the same way as before.

4.4.4 Results and Discussion

Comparison between the spectra of Formicamycin in hexane, calculated using the two semi-empirical (AM1 and PM6) methods (200 starting structures), and the spectrum from the previous method (using the OPLS3 force field), shows a very good agreement when PM6 is used (Figure 34), but an unacceptable match using AM1. Therefore, PM6 was tested further, starting from 200, 100 and 50 conformers and the match of the final spectra demonstrated that 50 conformers were enough to get good accuracy, while keeping the computational cost to a minimum (Figure 35).

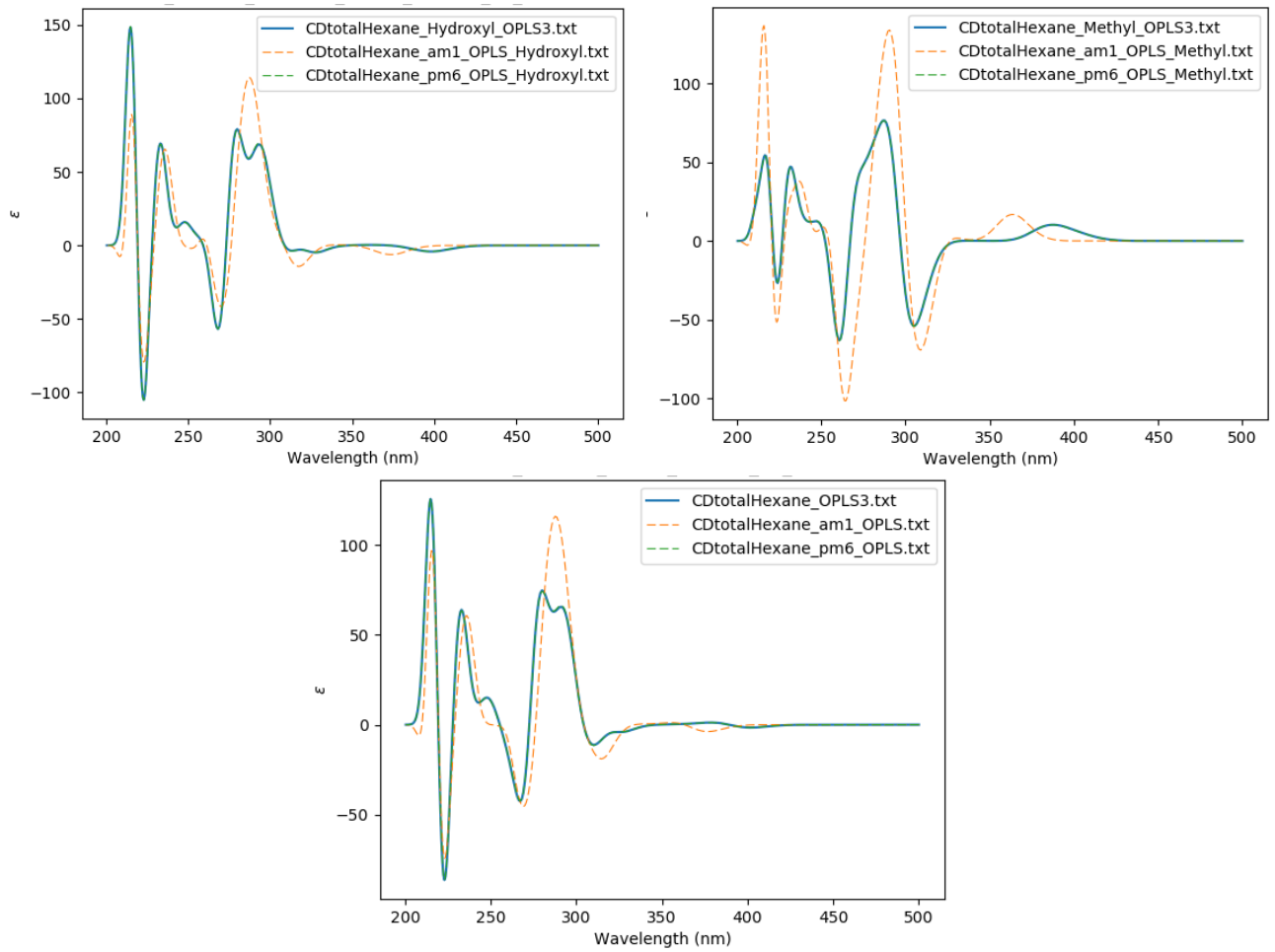


Figure 34. Comparison between the spectra obtained using AM1 (dotted orange), PM6 (dotted green) and the spectrum from the previous method, using the OPLS3 force field (blue). The upper left and right are associated to the hydroxyl and methyl configuration respectively, while the bottom spectrum is that generated by all the conformers.

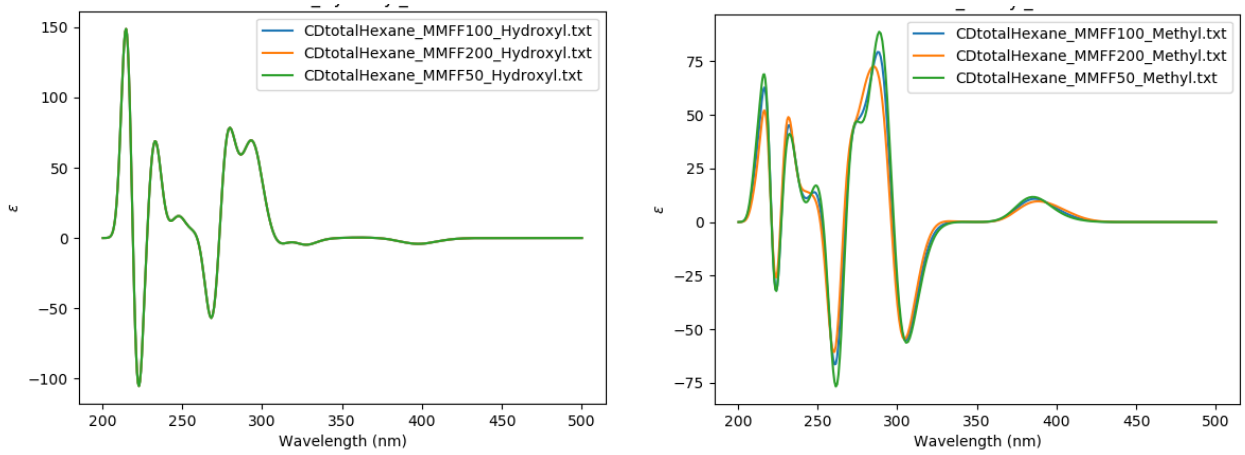


Figure 35. Comparison of the result obtained starting from an ensemble of 50 (green), 100 (blue) and 200 (orange) conformers in hexane. On the left: spectra associated with the hydroxyl configuration. On the right: spectra associated with the methyl configuration.

The relative energies of the conformers obtained from the conformational search, as calculated using the semi-empirical PM6 optimization and the DFT optimization, are plotted in Figure 36. The left side of Figure 36 compares the energies of the 200 conformations from RDKit's conformational analysis, calculated using MMFF94, to the optimized PM6 energies. As in the previous case, there is a great distribution of the energies and the majority of the conformers will be excluded after the semi-empirical calculation. In Figure 36 (right), in which the comparison between PM6 and DFT energies is shown, only the conformers with a PM6 relative energy less than 2.5 kcal/mol (on ordinate axis) have been selected from the previous step. Since the energies compare well with the DFT values, the following DFT optimization will be more efficient.

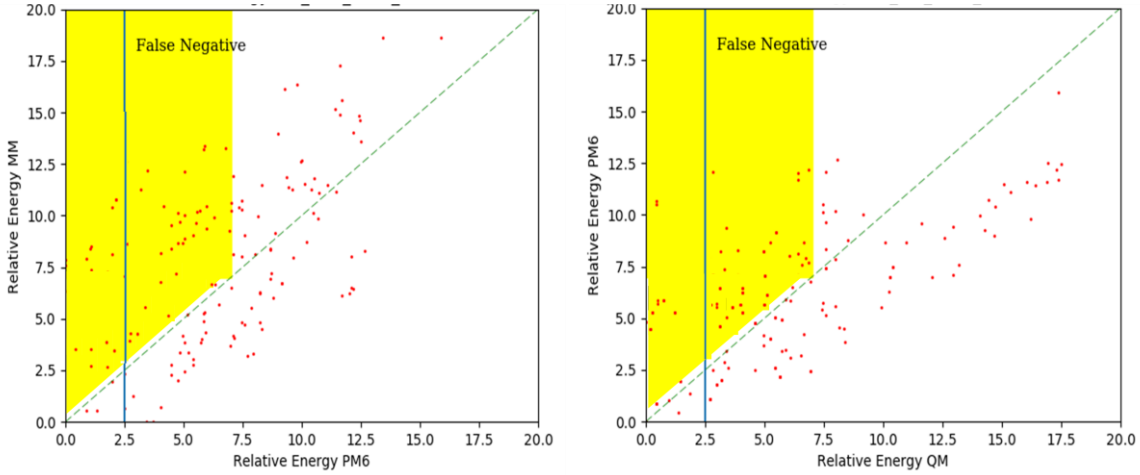


Figure 36. On the left: Comparison between the energies of the RDKit structures (y-axis) and the energies after the semi-empirical PM6 optimization (x-axis). On the right: Comparison between the energies from the SE calculation (y-axis) and the energies after the QM optimization (x-axis). All calculations were done in hexane as solvent.

5. Conclusion

The main aim of the present work was the calculation of Boltzmann weighted ECD spectra in order to identify the absolute configuration of the chiral axis of Formicamycin through comparison with the experimental spectrum. The procedure to generate the spectra can be divided into four main steps: the generation of an ensemble of conformers, the optimization of the geometry and the calculation of ZPE, a TDDFT calculation to get the rotational strengths and the generation of the Boltzmann weighted spectrum. The process was done in four different solvents (hexane, 2-methoxyethanol, water and methanol[43]) in order to analyze the effect of the solvent on the spectra.

The conformational search was first done using MacroModel's[33] conformational search function, implemented in the Maestro11[34] software package (chapter 2), setting as energy window, 21.0 kcal/mol. Two different force fields were tested (MMFFs[28] and OPLS3[29]) and both generated good spectra. The MMFFs conformational search generated less conformers with a smaller distribution of energy then OPLS3, making the overall process computationally cheaper. However, this is not necessarily ideal, as in some cases (see Section 4.2.1), MMFFs does not generate enough conformers for a good spectrum, so that OPLS3 proved to be a more reliable method.

All the calculations were carried out using the Gaussian09 software,[44] the PBE1PBE (PBE0) hybrid exchange-correlation functional with the def2-TZVP basis set[45] and the SMD solvation model,[46] to include solvent effects. This was guided by results from a previous study.[12] The optimization of the geometry, the ZPE calculation and the TDDFT calculation were performed separately, after selecting a subset of structures based on their relative energies, in order to reduce the computational cost. In particular, all the conformers contributing less than 1% (with a difference from the most stable conformer of more than 4×10-3 Hartree, or 2.5 kcal/mol) were excluded from the ensemble and those with an energy difference less than 1×10-6 Hartree were defined as equivalent.

The generation of the Boltzmann weighted spectra was done with the GaussSum software,[47] between 200 nm and 500 nm and with a sigma value of 0.25 eV. The comparison between the experimental spectrum in methanol and the calculated spectrum in the same solvent showed a very good match, allowing assignment of the absolute configuration of Formicamycin's chiral axis as R. The identification was possible thanks to the sign of the peak at 380 nm and the shoulder shape at 260 nm. Further study was done to elucidate a "fingerprint shift" when the solvent is changed. This showed that the spectra associated with the S-configuration has a greater number of changes, with respect to both the shift and intensity of the peaks, than the R-configuration. This could be the basis of a new approach

that can be investigated in future, where solvent-induced changes in the spectra provides further information to assist with the assignment of the AC.

The process was then optimized, reducing the computational cost and the dependencies on commercial software, as far as possible. The ensemble of conformers was generated using the open source RDKit software package.[35] This method allows the user to set the number of structures to generate and after the entire process was tested starting from 200, 100 and 50 conformers, 50 starting structures was chosen as the computationally most efficient approach without loss of information. Geometry optimizations were still carried out using Gaussian09, but made less demanding through employing a semi-empirical calculation. Both the AM1[24] and PM6[25] methods were tested, but only the latter proved to be a reliable alternative, showing a very good match with the results obtained through a full QM (DFT) optimization. The ZPE and the TDDFT calculation, as well as the generation of the spectra, was done in a similar way. The spectra obtained through this computationally "lighter" approach showed negligible differences with the previous "heavier" method, suggesting that this can be chosen as a new and viable alternative. Finally, the optimized process was completely automated using the Python scripting language.

Appendix A: The Python Automation Script

Main Process

In this chapter, the automated Python[53] script will be discussed, analyzing the main steps and highlighting the interconnected commands. The script uses the modules `shutil`, `numpy`, `glob`, `random`, `saga`, `matplotlib`, `time`, `random` and `RDKit`.

```
import subprocess as sub
import shutil
import glob, os
import numpy
import sys
import random
import time
import saga
import math
import matplotlib.pyplot  as plt

from subprocess import Popen, PIPE
from rdkit import Chem
from rdkit.Chem import AllChem, TorsionFingerprints, RWMol, rdmolfiles
from rdkit.ML.Cluster import Butina
```

A flowchart of the entire process is shown in Figure 37. In the flowchart, the four main parts are highlighted: the conformational search (`Conformational_Search`), remote calculation (`RemoteCalculation`), file selection (`SelectLogFile`) and generation of the spectra (`CDspectra`).

```
Conformational_Search(smilesstructure,NumConf,FileAllSDF,path,namedir,namefile)

RemoteCalculation(JobperGroups,SSHService,outputdircom,RemotePath)

SelectLogFile(os.path.join(outputdircom,SEdir),RelevantDir,Enrange,margin)

CDspectra(path,start,end,Sigma,numpts,Title,spectradir)
```

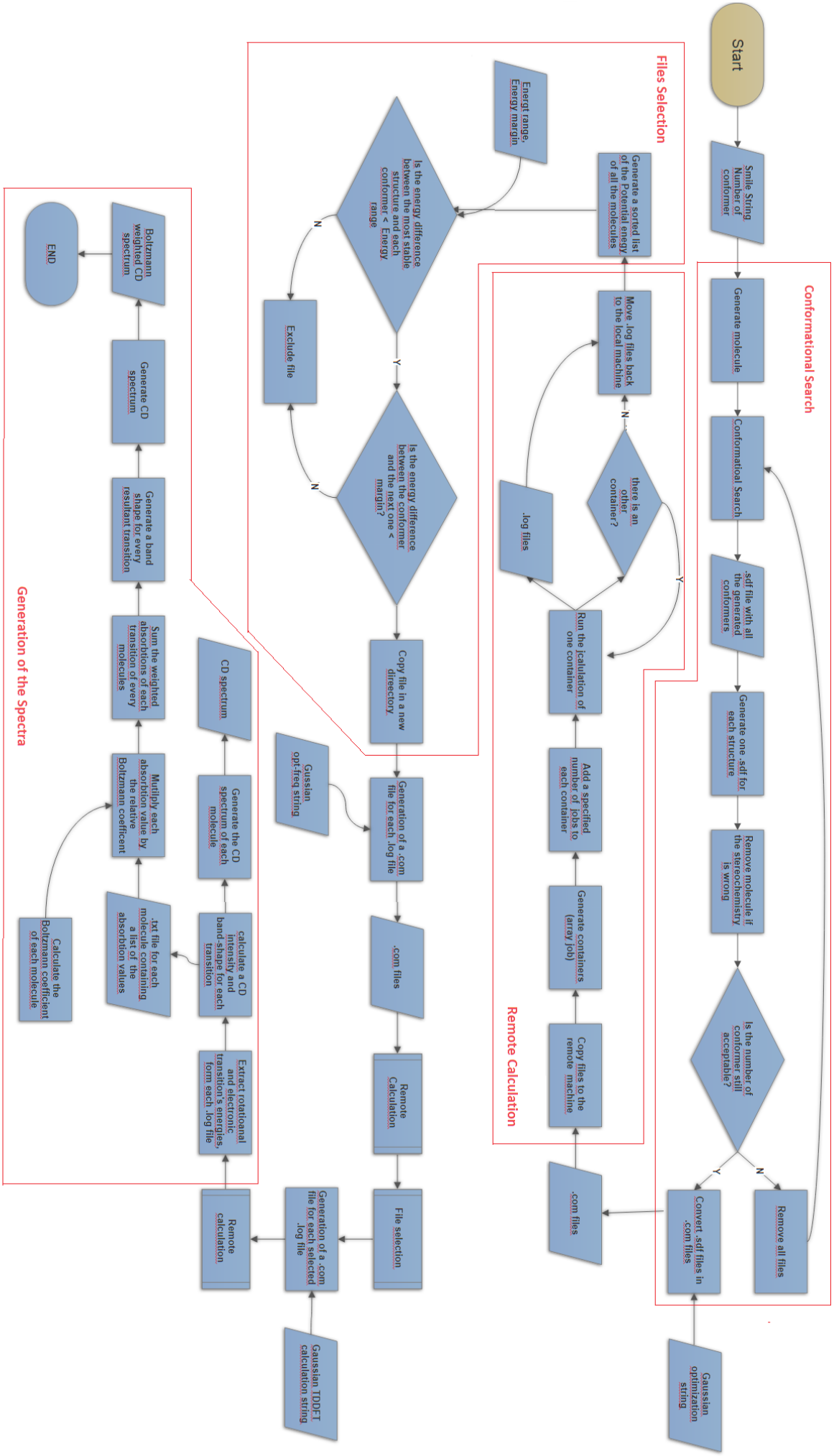


Figure 1. Flowchart of the process.

1. *Conformational_Search*

Variables must be set at the beginning of the script, in order to define the molecule to be analyzed, the number of conformations to generate and the names and paths for the output files.

```
#############Conformers_Generation#################
smilesstructure="O=C(/C=C/C=C/C=C/C=C/C(O[C@H]1[C@H](OC(C)=O)[C@@H](C)O[C@@H](C2=C(O)C([C@H](O)[C@@]34[C@@H](O)C
O)C8=C(C(Cl)=C(O)C=C8)OC7=O"
NumConf=int(20)
FileAllSDF="Conformers.sdf"  #File with all the conformers (specify .sdf)
path=' ' #Path of this script
namedir= "ConformationalSearch"   #Name of the directory that will contain the entire  project
namefile= "Simocyclinone_"      #Name of the single structure(The Conformer will have this name +IDconformes)
```

Taking as input a chiral SMILES[54] string, the script generates a molecule, adds hydrogen atoms, generates the 3D coordinates, assigns the correct stereochemistry and then performs a first energy minimization, using RDKit [35] tools.

```
m = Chem.MolFromSmiles(smilesstructure)

m = Chem.AddHs(m)   #Add Hydrogen
```

```
AllChem.EmbedMolecule(m,AllChem.ETKDG())
new = Chem.MolFromMolBlock(Chem.MolToMolBlock(m, includeStereo=True))
Chem.AssignStereochemistry(m)
```

```
AllChem.MMFFOptimizeMolecule(m)
```

Following this, the generation of conformers is performed:

```
#Generate conformers
conformerIds = AllChem.EmbedMultipleConfs(m, numConfs= NumConf, maxAttempts=1000, pruneRmsThresh=0.1, useExpTorsionAnglePrefs=True , useBasicKnowledge=True, enforceChira
lity=True, numThreads=0)
```

To ensure that all the chiral centers assigned in the starting SMILES string will be maintained in the output conformers, it is important to specify `enforceChirality=True`.

Next, all the conformers are written to a single *.sdf* file[55] as well as divided into several *.sdf* files, one for each conformer.

```
#Write all the conformation in a single .sdf file
write_conformers_to_sdf(m, FileAllSDF , rmsClusters, conformerPropsDict, minEnergy)
```

2. *RemoteCalculation*[56]

The setup needed is shown below:

```
############Remote calculation###########
SSHService = ""
RemotePath = ""
JobperGroups = int(10)                  #Number of job in each bulk calculation"
Nproc = "24"
mem = "10GB"                                    # gaussian calculation setup
String1 = "opt pm6 scrf=(smd,solvent=Methanol)"     # Semiempirical
String2 = "opt freq  pbe1pbe/def2tzvp  scrf=(smd,solvent=Methanol)"     #opt-freq
String3 = "pbe1pbe/def2tzvp scrf=(smd,solvent=Methanol) TD(NStates=30)"     #TDDFT
```

where `SSHService` and `RemotePath` define, respectively, the remote machine on which to run the calculations and the path of the input and output files on the remote machine. Using `JobperGroups`, it is possible to set the maximum number of concurrent jobs for each bulk calculation, `Nproc` and `mem` defines the number of processors and the memory to use in each calculation and the three strings (`String1` to `String3`) contain the route sections for the three different Gaussian calculations to perform. The script then initiates an SSH session on the remote machine:

```
service = saga.job.Service("pbs+ssh://"+SSHService)
```

The `pbs+ssh` specification highlights that the calculations will be run through the PBS job submission protocol. The calculations were submitted to the Centre for High Performance Computing cluster. It then generates an empty bulk job container to fill with the chosen number of groups:

```
    # create and  populate our containers
    containers = list()
    start=0
    for c in range(0, num_job_groups):
        # create containers
        containers.append(saga.job.Container())
```

Following this, the input files are copied to the remote machine.

```
infilesource = 'file://localhost'+LocalInputPath+'/'+files[j]+'.com'
infiletarget = 'sftp://'+SSHService+RemotePath

out = saga.filesystem.File(infilesource)
out.copy(infiletarget)
```

The next part describes all the information regarding the calculation, e.g. the executable needed to run the calculation (i.e. Gaussian[44]), the modules that need to be loaded to have all related software libraries available, the working directory, path for output and error files generated by the PBS software, the queue, memory allocation and maximum wall time. This information can change based on the queue management software and the hardware specifications of the remote machines and must therefore be manually filled in.

```
#add jobs to container.
jd = saga.job.Description()
jd.executable  =  ' ; '
jd.arguments   = [RemotePath+'/'+files[j]+'.com']
jd.working_directory = "."
jd.output          = ""
jd.error           = ""
jd.queue           = ""    # Using a specific queue
jd.project         = ""
jd.total_cpu_count = 24
jd.wall_time_limit = 2880  # minutes

jd.name        = ['job.%03d' % j]
j = service.create_job(jd)
containers[c].add(j)
```

The last row adds the j^{th} job to the c^{th} container and then the calculations are performed, running one container (of multiple jobs) at a time.

```
containers[c].run()
```

After the calculation has completed, the output file/s are sent back to the local machine in the same way as before, using the SSH service.

3. *SelectLogFiles*

This function selects conformers based on the specified energy criteria.

```
##################Conformer selection#######################
Enrange =  0.007                        #max Energy difference from the most stable
margin = 0.000001                       #minimun difference of energy for two different structures
```

Note that Gaussian09 writes final electronic energies in atomic units, thus the energy window must be expressed in Hartree. The script next generates a sorted list of all the conformers, by energy.

```
        for line in file:
                if (line[1:9] == "SCF Done"):
                        parts = line.strip().split()
                        E = float(parts[4])
        Energy.append(E)        #Generate a list of energies
Nfile = len (files)
ll = zip(Energy,files)      #Join the two previous lists
for i in range(Nfile):
        j = 1
        ls = sorted(ll)         #Energy sorted array
```

Following this, the criteria above is applied in order to exclude the high-energy, non-relevant conformers.

```
if abs(ls[0][0]-ls[i][0]) < Enrange:    #Exclude all the file with energy difference > range
        F = ls[i][1]

        relevant.append(ls[i][0])
        files.append(ls[i][1])

l1 = zip(relevant,files)            #Generate an new array with the files of interest
```

```
if abs(l1[i][0]-l1[i+1][0]) < margin:
        del (l1[i+1])                       #delete the element of the array
        Nfile=len(l1)
```

4. *CDspectra*

For the generation of the ECD spectrum, the setup parameters are the following:

```
#####################spectra#########################
start = 200                     # Start value in nm
end = 500                       # End value in nm
Sigma = 0.25
numpts = 500
Title = 'Simocyclinone_Methanol_50'   #Title of the spectrum
spectradir = "spectra0.25"          # Directory that will contain all the spectra
```

where `start` and `end` define the spectral range, `sigma` and `numpts` define the standard deviation value and the number of points to use when drawing the spectrum, which are correlated to the band width and resolution. A loop is then performed for each input file (*.log* files) consisting of the following steps: 1. the rotational strength and the transition energies are extracted.

```
  for line in file:  #extract the rotational strenghts from the .log file
          if (line[1:50] == "1/2[<0|r|b>*<b|rxdel|0> + (<0|rxdel|b>*<b|r|0>)*]"):
```

```
etrotats = numpy.array(etrotats, "d") #create an array of rotational strenghts
```

```
   elif (line[1:14] == "Excited State"): #extract the transition's electronic energies
```

```
etenergiescm.append(cm)       #append transition's electronic energies
```

2.For each transition, the maximum absorption value is calculated and a band width is associated to each transition with the GaussSum[47] function `GaussianSpectrum`.

```
for i in range(len(etrotats)): #calculation of the max absorption of each transition

        peakmax.append(prefactor * etrotats[i] *
                       etenergiescm[i] * 1e-40)


t = GaussianSpectrum(startwaveno,endwaveno,numpts,
                                ( etenergiescm,[peakmax] ),
                                real_FWHM)
```

3. A *.txt* file is created containing all the calculated absorption values, at each wavelength specified, and with this information the CD spectrum of a single conformer is generated, using `matplotlib`[48].

```
plt.figure(1)
plt.title(title +" "+ filename)
plt.plot(wave, Abs,label=filename)       #plot abs and wavelenth in order to generate the CD spectrum
plt.ylabel('$\epsilon$')
plt.xlabel('Wavelength (nm)')
plt.legend(loc='upper right')

plt.savefig(os.path.join(path, spectradir,'CD{}.png'.format(filename)))
plt.close('CD{}.png'.format(filename))
inputfile.close()
```

Each iteration of the loop generates a new spectrum for the current conformer, which is appended to the previous one. Thus, at the end of the process, the final spectra will be a superimposition of the CD spectra of all the individual conformers (see Chapter 4, fig.19).

The next step is to generate the Boltzmann-weighted CD spectrum. To do this, the potential energies are extracted from the *.log* files and the Boltzmann factor, at 298.15 K, is calculated for each conformer.

```
difference = []   #Calculation of the difference of energy from the stablest
for i in Toten:

        diff = i - lowest

        difference.append(diff)

        bf = []
for i in difference: # calculation of Boltzmann factor

        deno = 298.15*1.9872/1000   #Denominator and numerator of the Boltzmann distribution
        nume = -i*630                #taking acconut that the differences of energy are in Hartree
        bfact = math.exp(nume/deno)
        bf.append(bfact)
tot = reduce(lambda n,m:n+m,bf)  #Sum of all the the Boltzmann factor

boltzmann = []
for i in bf:
        x = i / tot          #Contribution of each molecule
        boltzmann.append(x)
```

Then, the *.txt* files, previously created, are read and a matrix containing the absorption values of all the conformers is generated.[57]

```
Abs = []
for line in file:
        parts = line.strip().split()
        R = float(parts[0])
        Abs.append(R)
Abs = numpy.array(Abs, "d")           #Exctract all the abs values

Abs = [Abs[x:x+numpts] for x in range(0, len(Abs),numpts)] #Create a matrix with the abs values for each molecule
```

Each row of the matrix is related to one conformer so that the next step is multiplying each row with the corresponding Boltzmann factor.

```
SAbs = []
for i in range(Nfile):
        for j in range(numpts):
                x = boltzmann[i] * Abs[i][j]
                SAbs.append(x)
SAbs = numpy.array(SAbs, "d")
SAbs = [SAbs[x:x+numpts] for x in range(0, len(SAbs),numpts)]  #Generate a matrix with all the scaled value
```

The rows are them summed in order to obtain a single row with the weighted absorption values.

```
specarr = []            # Generate a sigle array summing the saled abs of each molecule
for j in range(numpts):
        a =  SAbs[0][j] - SAbs[0][j]
        for i in range(Nfile) :
                a = a + SAbs[i][j]
        specarr.append(a)
```

Finally, a single spectrum is generated through `matplotlib`, as above.

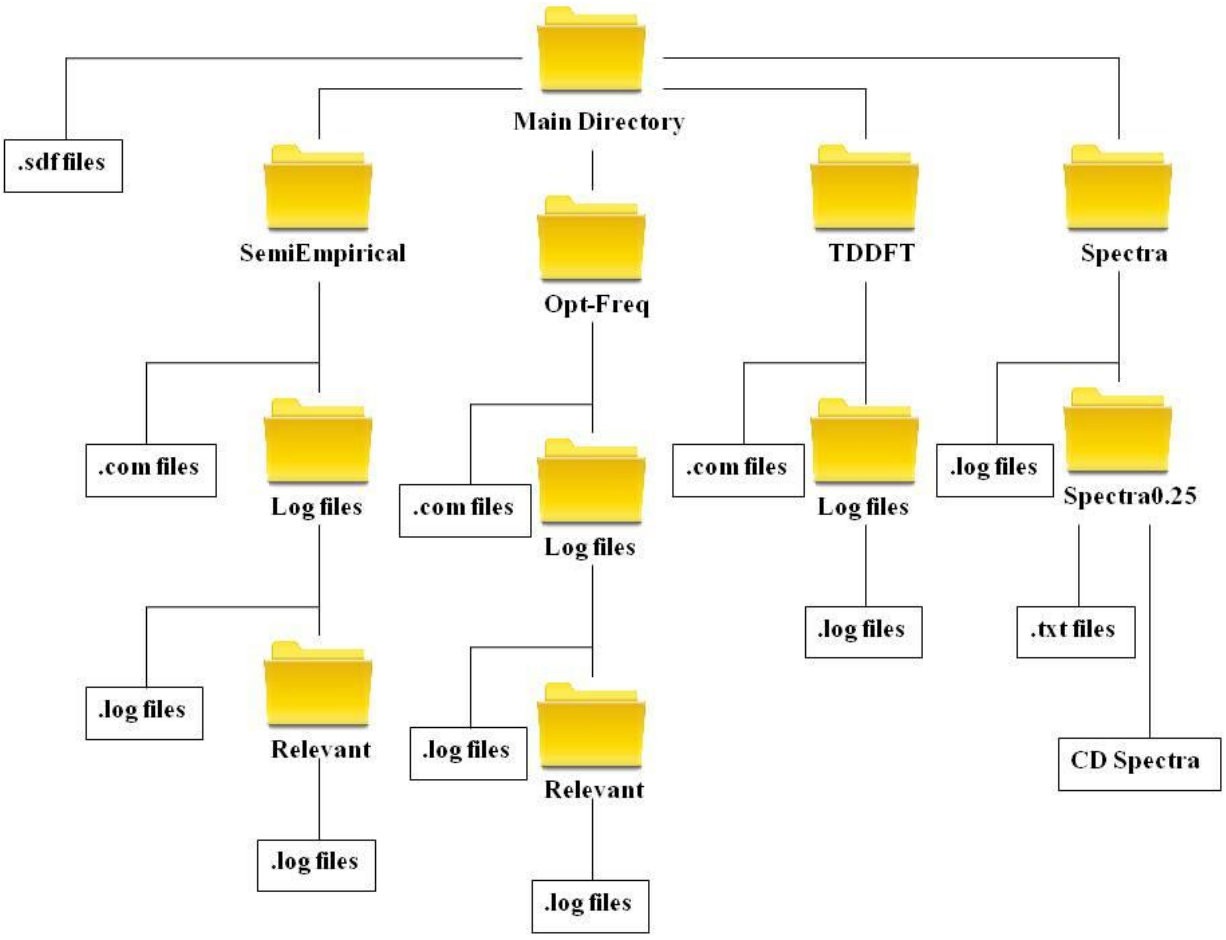
Other Commands

In addition to the above mentioned four main functions, two other operations were written. The first involves conversion of the file format from both *.sdf* (Conformational search output files) and *.log* (Gaussian output files) to *.com* (Gaussian input files), in order to perform the QM calculations. Since the conversion of the *.sdf* and *.com* files are performed only once, no function was created. On the other hand, the conversion from *.log* to *.com* is done several times using the `LogToCom` function.

```
def LogToCom(LogFilesPath,NewComFolder,String)
```

In both cases, the coordinates of the molecule are read and then rewritten in the correct format.

In addition, the information regarding the calculation (i.e. number of processors, memory and route section), chosen in the `RemoteCalculation` function (see above), are added. The second operation involves file management. This includes creating a directory, copying, moving and removing files. At the conclusion of the entire process, a specific folder structure is created where all the input, output and spectra are stored. The final structure is shown schematically as:



where the `Main` directory is allocated in the same directory as the script and can be renamed through the `ConformationalSearch` setup.

References

1. Kong, L. Y. & Wang, P. Determination of the absolute configuration of natural products. *Chin. J. Nat. Med.* **11,** 193–198 (2013).
2. Cecilia Noguez, F. H. Ab Initio Electronic Circular Dichroism of Fullerenes, Single-Walled Carbon Nanotubes, and Ligand-Protected Metal Nanoparticles. *Chirality* **26,** 553–562 (2014).
3. Flack, H. D. & Bernardinelli, G. The use of X-ray crystallography to determine absolute configuration. *Chirality* **20,** 681–690 (2008).
4. Darbeau, R. Nuclear magnetic resonance (NMR) spectroscopy: A review and a look at its use as a probative tool in deamination chemistry. *Appl. Spectrosc. Rev.* **41,** 401–425 (2006).
5. Mills, N. 150 and More Basic NMR Experiments: A Practical Course, 2nd Edition (Braun, S.; Kalinowski, H.-O.; Berger, S.). *J. Chem. Educ.* **77,** 831 (2000).
6. Seco, J. M., Quiñoá, E. & Riguera, R. The Assignment of Absolute Configuration by NMR. *Chem. Rev.* **104,** 17–117 (2004).
7. Pescitelli, G., Di Bari, L. & Berova, N. Conformational aspects in the studies of organic compounds by electronic circular dichroism. *Chem. Soc. Rev.* **40,** 4603 (2011).
8. Berova, N., Bari, L. Di & Pescitelli, G. Application of electronic circular dichroism in configurational and conformational analysis of organic compounds. *Chem. Soc. Rev.* **36,** 914 (2007).
9. Condon, E. U. Theories of optical rotatory power. *Rev. Mod. Phys.* **9,** 432–457 (1937).
10. Barron, L. D. *Molecular Light Scattering and Optical Activity*. *Vasa* (2009). doi:10.1017/CBO9780511535468
11. Mason, S. F., Seal, R. H. & Roberts, D. R. Optical activity in the biaryl series. *Tetrahedron* **30,** 1671–1682 (1974).
12. Qin, Z. *et al.* Formicamycins, antibacterial polyketides produced by Streptomyces formicae isolated from African Tetraponera plant-ants. *Chem. Sci.* **8,** 3218–3227 (2017).
13. Foresman, J. B. & Frisch, Æ. Exploring Chemistry with electronic structure methods (2nd ed.). *Pittsburgh, PA: Gaussian Inc.* 266, 278–283 (1996).
14. Marques, M. A. L., Maitra, N. T., Nogueira, F. M. S., Gross, E. K. U. & Rubio, A. *Fundamentals of Time-Dependent Density Functional Theory*. **837,** (2012).
15. Dreuw, A. & Head-Gordon, M. Single-reference ab initio methods for the

calculation of excited states of large molecules. *Chem. Rev.* **105,** 4009–4037 (2005).
16. Bannwarth, C. & Grimme, S. A simplified time-dependent density functional theory approach for electronic ultraviolet and circular dichroism spectra of very large molecules. *Comput. Theor. Chem.* **1040–1041,** 45–53 (2014).
17. Wang, F., Yam, C. Y. & Chen, G. Time-dependent density-functional theory/localized density matrix method for dynamic hyperpolarizability. *J. Chem. Phys.* **126,** 244102 (2007).
18. Perdew, J. P. Jacob's ladder of density functional approximations for the exchange-correlation energy. in *AIP Conference Proceedings* **577,** 1–20 (2001).
19. Gaussian inc. Gaussian site. *Density Functional (DFT) Methods* <http://gaussian.com/dft/> (2018).
20. Ernzerhof, M. & Scuseria, G. E. Assessment of the Perdew–Burke–Ernzerhof exchange-correlation functional. **5029,** (2005).
21. Ditchfield, R., Hehre, W. J. & Pople, J. A. Self- Consistent Molecular- Orbital Methods. IX. An Extended Gaussian- Type Basis for Molecular- Orbital Studies of Organic Molecules. *J. Chem. Phys.* **54,** 724–728 (1971).
22. Mukhopadhyay, P., Zuber, G., Wipf, P. & Beratan, D. N. Contribution of a solute's chiral solvent imprint to optical rotation. *Angew. Chemie - Int. Ed.* **46,** 6450–6452 (2007).
23. Marenich, A. V., Cramer, C. J. & Truhlar, D. G. Universal solvation model based on solute electron density and on a continuum model of the solvent defined by the bulk dielectric constant and atomic surface tensions. *J. Phys. Chem. B* **113,** 6378–6396 (2009).
24. Dewar, M. J. S., Zoebisch, E. G., Healy, E. F. & Stewart, J. J. P. Development and use of quantum mechanical molecular models. 76. AM1: a new general purpose quantum mechanical molecular model. *J. Am. Chem. Soc.* **107,** 3902–3909 (1985).
25. Stewart, J. J. P. Optimization of parameters for semiempirical methods V: Modification of NDDO approximations and application to 70 elements. *J. Mol. Model.* **13,** 1173–1213 (2007).
26. Bredow, T. & Jug, K. Theory and range of modern semiempirical molecular orbital methods. *Theor. Chem. Acc.* **113,** 1–14 (2005).
27. González, M. A. Force fields and molecular dynamics simulations. *Collect. SFN* **12,** 169–200 (2011).
28. Halgren, T. a. Merck Molecular Force Field. *J. Comput. Chem.* **17,** 490–519 (1996).
29. Harder, E. *et al.* OPLS3: A Force Field Providing Broad Coverage of Drug-like Small Molecules and Proteins. *J. Chem. Theory Comput.* **12,** 281–296 (2016).
30. Molecular, M., Field, F. & Halgren, T. A. Electrostatic Parameters for Intermolecular Interactions *. **17,** 520–552 (2000).
31. Halgren, T. A. MMFF VI. MMFF94s option for energy minimization studies. *J.*

Comput. Chem. **20,** 720–729 (1999).
32. Ebejer, J., Morris, G. M. & Deane, C. Freely Available Conformer Generation Methods : How good are they ? Freely Available Conformer Generation Methods : How good are they ? (2012). doi:10.1021/ci2004658
33. Schrödinger Release 2016–3: MacroModel (Schrödinger, LLC, New York, NY, USA, 2016). Schrödinger inc.
34. Schrödinger Release 2017-4: Maestro, Schrödinger, LLC, New York, NY, 2017. Schrödinger inc.
35. http://www.rdkit.org. RDKit: Open-source cheminformatic.
36. Schrödinger Release 2017-4: MacroModel, Schrödinger, LLC, New York, NY, 2017. *MacroModel 9.7 User manual*. (2009).
37. Shelley, J. Conformational Searching using MacroModel and ConfGen.
38. More, J. J. & Wu, Z. Distance geometry optimization for protein structures. *J. Glob. Optim.* **15,** 219–234 (1999).
39. B.J. Berne, H. Eyring, D. Henderson, W. J. *Physical Chemistry. An Advanced Treatise, Vol. 8B*. (1971).
40. Knoester, J. Optical Properties of Molecular Aggregates. in *Proceeding of the international school of physics 'Enrico Fermi', course cxlix, Organic nanostructures science and application*
41. Oliphant, T. E. Python for scientific computing. *Comput. Sci. Eng.* **9,** 10–20 (2007).
42. Millman, K. J. & Aivazis, M. Python for scientists and engineers. *Computing in Science and Engineering* **13,** 9–12 (2011).
43. Wilkinson, K. A. Unpublished work.
44. 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. Marenich, J. Bloino, B. G. Janesko, R. Gomperts, B. Mennucci, H. P. Hratchian, J. V. Ort, 2016. Gaussian 09, Revision A.02.
45. Weigend, F. & Ahlrichs, R. Balanced basis sets of split valence, triple zeta valence and quadruple zeta valence quality for H to Rn: Design and assessment of accuracy. *Phys. Chem. Chem. Phys.* **7,** 3297 (2005).
46. Marenich, A. V., Cramer, C. J. & Truhlar, D. G. Universal solvation model based on solute electron density and a contiuum model of the solvent defined by the bulk dielectric constant and atomic surface tensions. *J. Phys. Chem. B.* **113,** 6378–6396 (2009).
47. M. O'Boyle, A. L. T. and K. M. L. GaussSum. *J. Comp. Chem.* **29,** 839–845
48. Hunter, J. D. Matplotlib: A 2D Graphics Environment. *Comput. Sci. Eng.* **9,** 90–95 (2007).
49. Parrish, R. M. *et al.* Psi4 1.1: An Open-Source Electronic Structure Program Emphasizing Automation, Advanced Libraries, and Interoperability. *J. Chem. Theory Comput.* **13,** 3185–3197 (2017).

50. Neese, F. Software update: the ORCA program system, version 4.0. *Wiley Interdiscip. Rev. Comput. Mol. Sci.* **8,** 1–6 (2018).
51. Noel M. O'Boyle , Michael Banck , Craig A. James , Chris Morley, T. V. and G. R. H. Open Babel: An open chemical toolbox. *J. Cheminform.* **3:33,** (2011).
52. O'Boyle, N. M.; Vandermeersch, T.; Flynn, C. J.; Maguire, A. R.; Hutchison, G. R. Confab - Systematic generation of diverse low-energy conformers. *J. Cheminform* **3,** 1–9 (2011).
53. Rossum., G. Van. The Python Reference Manual. Network Theory Ltd.
54. Weininger, D. SMILES, a Chemical Language and Information System: 1: Introduction to Methodology and Encoding Rules. *J. Chem. Inf. Comput. Sci.* **28,** 31–36 (1988).
55. Landrum, G. RDKit Documentation. *Read. Writ.* (2011). doi:10.5281/zenodo.60510
56. The SAGA Project. SAGA-Python Documentation Release v0.29. (2015).
57. Stéfan van der Walt, S. C. C. and G. V. The NumPy Array: A Structure for Efficient Numerical Computation. *Comput. Sci. Eng.* **3,** 22–30 (2011).