





**UNIVERSIDAD DE INVESTIGACIÓN DE TECNOLOGÍA  
EXPERIMENTAL YACHAY**

**Escuela de Ciencias Físicas y Nanotecnología**

**TÍTULO: In Silico prediction of antibacterial activity of  
sesquiterpene lactones using density-functional theory and  
quantitative structure-activity relationship methods**

Trabajo de integración curricular presentado como requisito para la  
obtención  
del título de Físico

**Autor:**

Fabián Aníbal Puga Montesdeoca

**Tutor:**

Ph.D. Pinto Esperanza Henry Paul

Urcuquí, Enero 2021

Urququí, 20 de noviembre de 2020

**SECRETARÍA GENERAL**  
**(Vicerrectorado Académico/Cancillería)**  
**ESCUELA DE CIENCIAS FÍSICAS Y NANOTECNOLOGÍA**  
**CARRERA DE FÍSICA**  
**ACTA DE DEFENSA No. UITEY-PHY-2020-00023-AD**

A los 20 días del mes de noviembre de 2020, a las 10:30 horas, de manera virtual mediante videoconferencia, y ante el Tribunal Calificador, integrado por los docentes:

<b>Presidente Tribunal de Defensa</b>	Dr. MEDINA DAGGER, ERNESTO ANTONIO , Ph.D.
<b>Miembro No Tutor</b>	Dr. MOWBRAY , DUNCAN JOHN , Ph.D.
<b>Tutor</b>	Dr. PINTO ESPARZA, HENRY PAUL , Ph.D.

El(la) señor(ita) estudiante **PUGA MONTESDEOCA, FABIAN ANIBAL**, con cédula de identidad No. **1718349473**, de la **ESCUELA DE CIENCIAS FÍSICAS Y NANOTECNOLOGÍA**, de la Carrera de **FÍSICA**, aprobada por el Consejo de Educación Superior (CES), mediante Resolución **RPC-SO-39-No.456-2014**, realiza a través de videoconferencia, la sustentación de su trabajo de titulación denominado: **IN SILICO PREDICTION OF ANTIBACTERIAL ACTIVITY OF SESQUITERPENE LACTONES USING DENSITY FUNCTIONAL THEORY AND QUANTITATIVE STRUCTURE ACTIVITY RELATIONSHIP METHODS**, previa a la obtención del título de **FÍSICO/A**.

El citado trabajo de titulación, fue debidamente aprobado por el(los) docente(s):

<b>Tutor</b>	Dr. PINTO ESPARZA, HENRY PAUL , Ph.D.
--------------	---------------------------------------

Y recibió las observaciones de los otros miembros del Tribunal Calificador, las mismas que han sido incorporadas por el(la) estudiante.

Previamente cumplidos los requisitos legales y reglamentarios, el trabajo de titulación fue sustentado por el(la) estudiante y examinado por los miembros del Tribunal Calificador. Escuchada la sustentación del trabajo de titulación a través de videoconferencia, que integró la exposición de el(la) estudiante sobre el contenido de la misma y las preguntas formuladas por los miembros del Tribunal, se califica la sustentación del trabajo de titulación con las siguientes calificaciones:

Tipo	Docente	Calificación
Tutor	Dr. PINTO ESPARZA, HENRY PAUL , Ph.D.	10,0
Miembro Tribunal De Defensa	Dr. MOWBRAY , DUNCAN JOHN , Ph.D.	9,7
Presidente Tribunal De Defensa	Dr. MEDINA DAGGER, ERNESTO ANTONIO , Ph.D.	10,0

Lo que da un promedio de: **9.9 (Nueve punto Nueve)**, sobre 10 (diez), equivalente a: **APROBADO**

Para constancia de lo actuado, firman los miembros del Tribunal Calificador, el/la estudiante y el/la secretario ad-hoc.

Certifico que *en cumplimiento del Decreto Ejecutivo 1017 de 16 de marzo de 2020, la defensa de trabajo de titulación (o examen de grado modalidad teórico práctica) se realizó vía virtual, por lo que las firmas de los miembros del Tribunal de Defensa de Grado, constan en forma digital.*

PUGA MONTESDEOCA, FABIAN ANIBAL  
**Estudiante**



Dr. MEDINA DAGGER, ERNESTO ANTONIO , Ph.D.  
**Presidente Tribunal de Defensa**

Dr. PINTO ESPARZA, HENRY PAUL , Ph.D.  
**Tutor**

Dr. MOWBRAY , DUNCAN JOHN , Ph.D.  
**Miembro No Tutor**

ALARCON FELIX, KARLA ESTEFANIA  
**Secretario Ad-hoc**

## AUTORÍA

Yo, **Fabián Aníbal Puga Montesdeoca**, con cédula de identidad 1718349473, declaro que las ideas, juicios, valoraciones, interpretaciones, consultas bibliográficas, definiciones y conceptualizaciones expuestas en el presente trabajo; así cómo, los procedimientos y herramientas utilizadas en la investigación, son de absoluta responsabilidad de el/la autora (a) del trabajo de integración curricular. Así mismo, me acojo a los reglamentos internos de la Universidad de Investigación de Tecnología Experimental Yachay.

Urcuquí, diciembre 2020.



---

Fabián Aníbal Puga Montesdeoca

CI: 1718349473

## AUTORIZACIÓN DE PUBLICACIÓN

Yo, **Fabián Aníbal Puga Montesdeoca**, con cédula de identidad 1718349473, cedo a la Universidad de Tecnología Experimental Yachay, los derechos de publicación de la presente obra, sin que deba haber un reconocimiento económico por este concepto. Declaro además que el texto del presente trabajo de titulación no podrá ser cedido a ninguna empresa editorial para su publicación u otros fines, sin contar previamente con la autorización escrita de la Universidad.

Asimismo, autorizo a la Universidad que realice la digitalización y publicación de este trabajo de integración curricular en el repositorio virtual, de conformidad a lo dispuesto en el Art. 144 de la Ley Orgánica de Educación Superior

Urcuquí, Diciembre 2020.



---

Fabián Aníbal Puga Montesdeoca  
CI: 1718349473

## **Dedication**

To my family and friends, especially to my mom, father and brother, who have been my support and my happiness during all these years.

Fabián Aníbal Puga Montesdeoca

## Acknowledgements

Firstly, I would like to thank my family: my parents and brother for supporting me spiritually throughout writing this thesis and my life in general.

I would like to express my sincere gratitude to my advisor PhD. Henry Pinto for the continuous support of my thesis project and related research, for his patience, motivation, and immense knowledge. His guidance helped me in all the time of research and writing of this thesis.

Besides my advisor, I would like to thank to PhD. Alicja Mikołajczyk and PhD. Paola Ordoñez for their valuable help, advice and guide in the various areas of my thesis, without their help this thesis would not have been possible

My sincere thanks also go to the university of Gdansk, who provided me access to they research facilities. Without they generous support it would not be possible to conduct this research. My gratitude to Yachay Tech University who gave me the opportunity of taste the real research labor.

I thank my physicists partners for the stimulating discussions, for the sleepless nights we were working together before deadlines, and for all the fun and experiences we had in the last five years.

Finally, to the family I chose, my friends. Thank you all for every moment that I had the opportunity to live and share with you, all the adventures, love and friendship. You will always be in my hearth.

Fabián Aníbal Puga Montesdeoca



## Resumen

La creciente resistencia que desarrollan las bacterias a los antibióticos es un problema que afecta a todos los estratos sociales. Por tanto, el desarrollo de componentes antibacterianos nuevos y eficaces es de vital importancia para nuestra sociedad. Las Lactonas sesquiterpénicas (STL) son un grupo de metabolitos secundarios aislados de plantas que han mostrado un amplio espectro de actividades biológicas, especialmente actividad antibacteriana contra *Staphylococcus aureus* resistente a la meticilina. Desafortunadamente, los métodos experimentales para estudiar la efectividad de los antibióticos a base de plantas son costosos y requieren mucho tiempo. Para sobrepasar estas limitaciones, se pueden aplicar estudios computacionales para acelerar el desarrollo de antibióticos más eficientes. En este estudio, se realizaron cálculos de estructura electrónica en 21 STL para desarrollar un modelo capaz de predecir la actividad antibacteriana de nuevas moléculas de STL. Mediante el uso de una combinación óptima del método de funcional de densidad y tight-binding (DFTB) y cálculos de la teoría funcional de densidad ab initio (DFT), pudimos calcular los conformeros más favorables energéticamente, su estructura atómica y propiedades físico-químicas. Los valores calculados usando mecánica cuántica se combinaron utilizando modelos Quantitative Structure-Activity Relationship (QSAR) considerando la actividad antibacteriana obtenida experimentalmente. El modelo QSAR desarrollado utilizó diferentes combinaciones de dos descriptores. Los resultados preliminares sugieren que los modelos que incluyen el HOMO y la energía electrónica correlacionan mejor la actividad antibacteriana. Estos resultados podrían permitir una predicción confiable de la actividad antibacteriana para nuevos compuestos que pertenecen a la familia STL basándose en las propiedades calculadas por DFT.

### Palabras Clave:

Conformeros, DFT, DFTB, *Staphylococcus aureus* resistente a la meticilina, xTB, CREST, ORCA, Descriptores.

## Abstract

The growing resistance developed by bacteria to antibiotics is a problem that involves every social stratum. Therefore, the development of new and effective anti-bacterial components is of vital importance for our society. Sesquiterpene Lactones (STL) are a group of secondary metabolites isolated from plants that have shown a wide spectrum of biological activities especially antibacterial activity against methicillin-resistant staphylococcus aureus (MRSA). Unfortunately, the experimental methods to study the effectiveness of plant-based antibiotics are expensive and time-consuming. In order to tackle these limitations in silico studies can be applied to accelerate the development of more efficient antibiotics. In this study, electronic structure calculations on 21 STL were performed to develop a model capable to predicting the antibacterial activity of new STL molecules. By using an optimal combination of density-functional tight-binding (DFTB) method and ab initio density-functional theory (DFT) calculations, we were able to calculate the most energetically favorable conformers, their atomic structure and physical-chemical properties. The quantum mechanically computed values were then combined using Quantitative Structure-Activity Relationship models considering experimental antibacterial activity. The developed QSAR model used different combinations of two descriptors. Preliminary results suggest that models that includes the HOMO and electronic energy correlates better the antibacterial activity. These results could allow reliable prediction of antibacterial activity for new compounds that belong to the STL family based on the DFT computed properties.

### **Keywords:**

Conformers, DFT, DFTB, Methicillin-Resistant Staphylococcus Aureus, xTB, CREST, ORCA, Descriptors.

# Contents

<b>List of Figures</b>	<b>xii</b>
<b>List of Tables</b>	<b>xiii</b>
<b>List of Papers</b>	<b>xv</b>
<b>1 Introduction</b>	<b>1</b>
1.1 Problem Statement . . . . .	6
1.2 General and Specific Objectives . . . . .	7
1.2.1 General Objective . . . . .	7
1.2.2 Specific Objectives . . . . .	7
<b>2 Theoretical Background</b>	<b>9</b>
2.1 Quantum mechanics . . . . .	9
2.1.1 Principles of quantum mechanics . . . . .	9
2.1.2 Electronic structure problem . . . . .	10
2.1.3 Born-Oppenheimer approximation . . . . .	11
2.2 Density Functional Theory . . . . .	12
2.2.1 Thomas-Fermi Theory . . . . .	12
2.2.2 Hartree-Fock Approximation . . . . .	13
2.2.3 Hohenberg-Kohn Theorem . . . . .	14
2.2.4 The Kohn-Sham Equations . . . . .	16
2.2.5 Exchange Correlation Functional . . . . .	17
2.2.6 Local Density Approximation . . . . .	17
2.2.7 Generalized Gradient Approximations . . . . .	19
2.2.8 Meta-Generalized Gradient Approximations . . . . .	19
2.2.9 Hybrid functional . . . . .	20
2.3 Basis Sets . . . . .	20
2.4 Tight Binding . . . . .	22

2.5	RIJCOSX Algorithm . . . . .	23
2.6	Dispersion correction . . . . .	23
2.7	Ab Initio Method . . . . .	23
2.8	Sesquiterpene Lactones . . . . .	24
2.9	Cytotoxicity . . . . .	25
2.10	Methicillin-resistant Staphylococcus aureus . . . . .	25
2.11	Quantitative Structure-Activity Relationship . . . . .	25
2.11.1	Descriptors . . . . .	26
2.11.2	Endpoint . . . . .	27
2.11.3	Statistical variables . . . . .	27
<b>3</b>	<b>Methodology</b>	<b>29</b>
3.1	Computational Methods . . . . .	29
3.1.1	Conformer-Rotamer Ensemble Sampling Tool . . . . .	29
3.1.2	ORCA . . . . .	30
3.2	Computational procedure . . . . .	31
3.3	Quantitative structure-activity relationship model development . . . . .	34
3.3.1	Endpoint . . . . .	35
3.3.2	Descriptors . . . . .	36
3.3.3	Prediction . . . . .	37
3.3.4	Molecular orbitals and Structure activity relationship . . . . .	37
<b>4</b>	<b>Results &amp; Discussion</b>	<b>39</b>
4.1	Structures modeling and electronic properties calculation . . . . .	39
4.2	Quantitative structure-activity relationship model . . . . .	42
4.2.1	Minimum inhibitory concentration prediction . . . . .	51
<b>5</b>	<b>Conclusions &amp; Outlook</b>	<b>53</b>
	<b>Bibliography</b>	<b>55</b>
	<b>Abbreviations</b>	<b>65</b>

# List of Figures

1.1	General classification of STL . . . . .	2
2.1	Schematic diagram of Jacob's Ladder . . . . .	18
3.1	Graphical description of energy variation between conformers . . . . .	30
3.2	Thesis process diagram . . . . .	31
3.3	Scifinder user interface . . . . .	32
3.4	Comparison between reported and relaxed structure . . . . .	33
3.5	ORCA input file . . . . .	34
4.1	QSAR models graphics . . . . .	44



# List of Tables

3.1	Training and validation set molecules . . . . .	35
4.1	Selected <i>Sesquiterpene lactones</i> . . . . .	39
4.2	Calculated electronic properties of STL . . . . .	41
4.3	Most relevant combination of descriptors . . . . .	43
4.4	Statistical variables . . . . .	44
4.5	Result summary . . . . .	45
4.6	Prediction set molecules . . . . .	52





# Chapter 1

## Introduction

Antibiotic resistance has been a latent problem that has grown and evolved in the past decades affecting all human kind; therefore, the development of new and effective antibacterial components is of vital importance for our society. *Sesquiterpene Lactones* (STL) are a group of secondary metabolites isolated from plants belonging to the *Asteraceae* (*Compositae*) family<sup>1</sup>. These organic molecules are of great interest in the scientific community since they have a wide variety of chemical structures and have shown a wide spectrum of biological activities including antimicrobial, anti-fungal, anti-inflammatory, anticancer, among others. Despite these novel properties, it is also known that some STL are toxic to both human and animal parasites, which undoubtedly raises their scientific interest<sup>2</sup>. It should be noted that the *Asteraceae* family is one of the most widely distributed and abundant plant families. This means that there are various specimens from which STL are obtained. Being synthesized from different sources, it is to be expected that a single chemical structure will not be obtained. In other words, their structures are as varied as their biological diversity. Thanks to these differences, extracts that contain these molecules have been used for different purposes (without knowing their structure) in ancient medicine and in the modern pharmaceutical industry. Currently it has been possible to know the structure of several of these molecules and in the Dictionary of Natural Products (DNP), there are at least 5000 different structures for STL<sup>3</sup>.

The structural diversity of STL is a little unusual because among all the structures, 87% are formed by only 7 of the more than 100 types of STL. To get a general idea of how the main groups are structured a summary—without going into detail about the possible functional groups—is provided in Figure 1.1. But it is expected that the naturally formed structures that can be found in plant extracts are more complex and with many more extra elements than the structures previously described.

In Figure 1.1 the basic structure for STL is sketched based on the bond-line notation, here every carbon atom is located in each corner or terminal, oxygen atoms are represented by "O", single lines represent single bonds, double lines represent double bonds, finally all the incomplete bonds are filled with Hydrogen atoms which are suppressed for simplicity. This notation is use through the entire thesis. In recent decades, STL have been the object

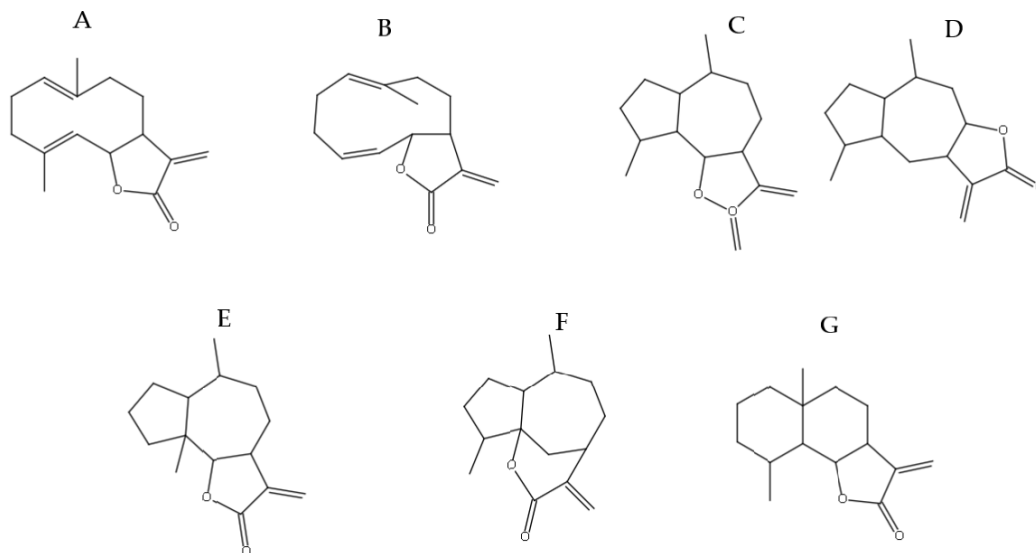


Figure 1.1: General classification of STL. A: Germacranolides, B: Heliangolides, C+D: Guaianolides, E: Pseudoguaianolides, F: Hypocretenolides, G: Eudesmanolides

of great attention thanks to the fact that they have presented a wide variety of properties that can be used mainly in medicine, agriculture, and nutraceuticals. Those properties that can have an impact on medicine are the ones that get the most attention. Some STL can inhibit the growth of cancer cells and therefore have anti-tumor and cytotoxic activity. Others can inhibit microbial growth, better known as antibiotics. STL can also work as insect feeding deterrents for specific types of plagues. Phytotoxins or plant-growth inhibitors are other properties of STL, growth of plants or seed germination regulation<sup>4</sup>. STL is not only useful for humans but also plants where it is used as chemoprophylaxis in schistosomiasis. Not all the properties of STL are good for humans, some of them can cause allergic reactions depending on the quantity that gets in contact with skin. Some other properties are the anti-fungal and anti-inflammatory activity<sup>3</sup>.

Although STL have shown a wide spectrum of biological activities including antimicrobial; unfortunately, the experimental methods to study the effectiveness of plant-based antibiotics are expensive and time-consuming. An alternative to tackle these limitations and accelerating the research on these compounds is performing *in silico* studies of these molecules to predict their antibacterial activity.

### Importance of modeling

The importance of *in silico* research (or computational modeling) lies in the versatility of this type of studies that can be easily applied in various research areas. It allows us to test hypothesis and explore a wide range of possibilities

(computational experiments) before carrying out an actual experiment. Appropriate computational experiments allow us to corroborate results and could provide important guidelines on what we can expect from an experiment. Computational modeling allows us to manipulate variables that are not manipulable in experiments since many of these are abstract mathematical expressions based on physical concepts. Currently, given the great development that exists in terms of the computing power, it is easy to recognize that an intelligent combination of computer modelling with targeted experiments could be reflected on lowering the total expenses involved in research and more importantly, it can save time making the research area more efficient. One advantage of simulations is that we can build a model based on already reported experimental results.

On the other hand, when carrying out this type of *in silico* studies we can place molecules in any medium and in practically any condition, as well as vary these conditions in the course of the simulation and see how our system develops. As we can arbitrarily vary these parameters, we will be able to determine what type of conditions will be optimal to achieve the best performance.

To the best of our knowledge there are scarce *in silico* studies that have been developed on these STL molecules and the information that is currently available is mostly obtained from experiments with STL. Despite the little computational information that exists, it can be said that the results obtained from the different investigations are of relevance since they have established relationships between the type of structure and the biological activity that the molecule presents, as well as determining active groups. There is a common denominator in how STL studies are carried out. These generally first present an experimental phase in which they characterize STL to later go to a computational stage. The results obtained in this phase will strongly depend on the results obtained experimentally. This dependence may limit the level at which structural relationships are established since it depends on how the molecule was synthesized or what database is used for computational data. A great advantage of STL is that the main skeleton-type has been well studied. By knowing its main structure, arrangements can be made either by adding or removing different secondary functional groups that STL may have. By achieving this we help make the system under study easier to calculate.

### **Sesquiterpene Lactones and Docking studies**

Docking method allows us to determine what will be the dominant link mode between ligand and the protein of interest<sup>5</sup>. By knowing the 3D structure of the target molecule and especially the structure responsible for the biological interaction (such as antibacterial activity). It is possible to identify what type of compound can be joined using computational modeling techniques<sup>6</sup>. This method is widely used mainly in computational physics and chemistry for the different advantages it presents when determining structures and possible operating mechanisms. The docking studies on STL are not excluded, some studies theoretically propose an interaction mechanism that has helped to increase the knowledge about STL. Generally, *in silico* studies on the biological activity of organic molecules cannot establish theories that cover all molecules. Since they are so varied, we are going to find different morphological and structural characteristics that would change their mechanisms of action, interaction, and function. This is why these studies focus on groups of not very large molecules that share similar characteristics, either structural or of the

target molecule.

A study that exemplifies this was carried out by Aliyu et al. in 2016<sup>7</sup>. They isolated STL from *Vernonia blumeoides* and carried out docking studies to determine the Quorum inhibitory potential. They were able to establish that these STL had different cognate sites which bind to the target proteins. This meant that these molecules had the potential to reduce virulence and pathogenicity of drug-resistant bacteria *in vivo*. For the Docking model part, they used the 3D coordinates available in the Protein Data Bank (PDB)<sup>8</sup> of the target molecules and the STL. Later they found the most favorable energetic conformer and optimized it to the level of density-functional theory. By taking these relaxations into account we ensure that the results obtained will be within a range of possible energies so that the processes occur naturally.

The relaxation of the molecules is an important process when calculating the total energy of the system. By knowing the energy of the system and knowing that its values are within "normal" or expected ranges, we can have a first control point for our study. This relaxation process of the compounds is generally carried out at the beginning of the *in silico* studies to ensure that the structures are consistent with their energy. Hegazy et al. in 2015<sup>9</sup> carried out a docking and pharmacophore study to determine the inhibitory power that STL from *Cyara cornigera* has on Acetylcholinesterase. They use the 3D structure of each component and perform an energy minimization to each one of them and a protonation process at 300 K and pH to 7 was applied to the protein structure before starting the docking study. The calculation of the electrostatic interactions between two atoms was needed to determine the active site of the protein. Many considerations have to be taken into account when working with molecules, not only the structure and the energy is important, other factors will affect the dynamics of the molecule.

In some cases, to assure that the results are reliable, it is necessary to apply more than one method that takes into account different parameters of the system. There are cases where if we just have the coordinates of the molecules, this will not be enough to achieve accurate results. We might need also to calculate the electronic structure, or to perform Quantitative Structure-Activity Relationship (QSAR) models. Luo et al. in 2011<sup>10</sup> uses a combination of the density functional theory (DFT) method to calculate the electronic structure of STL, and a QSAR method to then employ a Docking model to simulate how STL will interact with human aromatase. They took the IC<sub>50</sub> values for the STL from previous studies and optimized the geometries of the molecules using DFT. From this, they took some relevant data called descriptors such as the total energy, highest occupied molecular orbital (HOMO), lowest unoccupied molecular orbital (LUMO), bandgap energy, dipole moment, etc to generate a QSAR model. Then they perform the molecular docking to determine the pharmacophore analysis. After all these considerations they found that modifying an external double bond of one of the compounds they can eliminate the cytotoxicity of the molecule. Looking at this study we can say with certainty that combining different models is a reliable methodology to secure results and to have a better understanding of the system dynamics. It is important to mention that to the best of our knowledge there are no studies at the quantum mechanical level that includes Van der Waals interactions.

### Sesquiterpene Lactones and Quantitative Structure-Activity Relationship Studies

Thanks to the utility that these molecules have in different biological fields, it is important to be able to understand not only their operating mechanisms but also how their atomic and electronic structure is related to their biological activity. This relationship might be described by the development of appropriate mathematical model. Currently, there are some reports on the Structure-Activity Relationship (SAR) and a few studies on the Quantitative Structure-Activity Relationship (QSAR). However, these models have some limitations, i.e., you cannot make a single model that describes everything at the same time. For this reason the desired model is developed to describe a specific function of a set of molecules of interest. It should be noted that with the development of this type of study it will be possible to understand and improve the different uses that can be given to STL.

In 1971 Kupchan et al.<sup>11</sup> carried out the first study to establish the relationship between structures and anti-tumor activity. They were able to demonstrate that STL enhanced their biological activity by increasing lipophilicity and with the presence of extra  $\alpha$ ,  $\beta$ -unsaturated carbonyl groups. They also established that the presence of  $\alpha$ -methylene- $\gamma$ -lactone is essential for cytotoxicity.

Another study carried out by Schmidt in 2006<sup>3</sup> concluded that the study of the dependency between structure and activity is of great importance for the development of future fields of research. In his research, he determined that, although STL do not have the structure of a drug, they can be quite similar in their physico-chemical properties and that we can take advantage of this similarity to use them as distributors of a specific drug that works as an antitumor, anti-inflammatory or anti-protozoal. Thanks to the QSAR studies, Schmidt was able to determine that the cytotoxicity of STL depends largely on the amount of alkylant structure elements.

Reyes et al. in 2007<sup>12</sup> used the 3D-QSAR / Comparative molecular similarity indices analysis (CoMSIA) model to characterize the elements of the STL structures responsible for inhibiting P-glycoprotein (Pgp). In particular, this type of model can only be used in molecules that bind to the same type of receptor and that also has a similar mechanism. Reyes was able to determine that the most important groups are the carbonyl in Carbon number two (c-2), Carbon number three (c-3) and Carbon number eight (c-8) that function as H-bond receptors.

STL can also be used to fight diseases for which there is still no cure, such as cancer. Schomburg et al. in 2013<sup>13</sup> found that some STL can interrupt or interfere with C-Myb transcription. Excessive expression of c-Myb (a proto-oncogene) causes tumors in humans, as well as leukemia and colon cancer. The QSAR method used as main descriptors a model of pharmacophores, the alignment of the molecules, and other secondary or derived descriptors. By implementing a QSAR model, they were able to obtain a multiple linear regression equation. Which allowed them to directly interpret the results, making it possible to describe 83% of the different biological activities predict others, despite the little information on the activities of the molecules used. Finally, they found that two of its most active molecules shared similarities in their structures, which allowed them to bind to a specific binding site which is still unknown.

If we take a look at the different structures that have been studied, we will notice that most large structures are the ones that play a role in the biological activity of STL. Double bonds and small functional groups such as hydroxyls are also relevant since their position within the larger structure (10-carbon rings) will also affect the recorded biological activity<sup>1</sup>.

The QSAR method has great advantages for researchers since it allows them to corroborate their results and draw robust conclusions. It is of great important to know what information we should put in our model because that is the basis from which all our possible outcomes originate. This is why it is important to emphasize that in all QSAR studies it is necessary that the test subjects have similar qualities, whether they are structural, physical-chemical, or biological. The model will also depend on the molecules we use (if they belong to the same base structure or if they vary) and what will be the purpose of the model (what relationship we seek to establish between structures). In other words, our model will depend on the selected descriptors. QSAR models provide an equation with specific parameters. These parameters or descriptors are those who best describe the system<sup>1</sup>. This equation is the one that will allow researchers to establish relationships and to predict possible biological activity values.

In this thesis, our computational studies combine quantum mechanical calculations at both the level of semiempirical tight binding and *ab initio* density-functional theory (DFT) with appropriate hybrid functionals that includes van der Waals dispersions to predicting the most energetically favorable conformers, their atomic and electronic structure, and physical chemical properties. These computed values are processed in quantitative structure-activity relationship models of antibacterial activity. The results obtained on the training set of the *Sesquiterpene lactones* molecules will allow us to propose and find more effective anti-bacterial *Sesquiterpene lactones*-based compounds.

## 1.1 Problem Statement

*Staphylococcus aureus* methicillin-resistant is a bacteria that has developed a great resistance to the medicines available to treat *staphylococcus*. This bacteria can be acquired in the most common health care places, generating a problem that involves every level of society. Some compounds isolated from plants show a wide spectrum of biological activities and plant-based antibiotics can be developed. Unfortunately, not all plant-based antibiotics present a considerable antibacterial activity against *staphylococcus aureus*. Moreover, it is difficult to select only the optimal antibiotic because the experimental methods to study the effectiveness of these antibiotics are expensive and time-consuming. Also, with experimental studies is not possible to fully determine which structure has the greater biological activity, generating a problem when structure-activity relationships are established. *In silico* studies allow us to explore all the possible structures of a molecule that combined with statistical methods help us to predict the possible relationships between structure and activity. This method could allows the whole research process to be much faster, cheaper and, in addition, it could potentially reduce the number of necessary experiments. In other words, we can optimize the process of designing and studying new effective antibacterial compounds.

## 1.2 General and Specific Objectives

### 1.2.1 General Objective

Develop a Quantitative Structure-Activity Relationship model based on a methodology that employs density-functional theory and tight binding calculations to predict the antibacterial activity of *Sesquiterpene Lactones* against Methicillin-Resistant *Staphylococcus Aureus*.

### 1.2.2 Specific Objectives

- Determine the DFT methodology which properly describes the electronic properties of STL.
- Estimate the possible biological activity of untested STL using the developed QSAR model.
- Identify the relationship between biological activity and electronic structure.





## Chapter 2

# Theoretical Background

### 2.1 Quantum mechanics

Newton's laws are a powerful tool used to describe and predict nature. For example, it is used to calculate the motion of a body in time, to see what forces are implied in a system. These equations have a huge range of applications since they can correctly be used with objects with the size of a simple fly or as big as a planet. This enormous range of applicability gives to this work the wrong idea that it was universally applicable. It was found that for very small objects such as atoms or molecules these laws do not explain correctly the phenomena. Therefore, a new theory was needed, the quantum mechanics theory<sup>14</sup>

#### 2.1.1 Principles of quantum mechanics

The main equation in quantum mechanics is the Schrödinger equation, that in 1D is expressed as<sup>15</sup>

$$i\hbar \frac{\partial \Psi(x, t)}{\partial t} = -\frac{\hbar^2}{2m} \frac{\partial^2 \Psi(x, t)}{\partial x^2} + V(x, t)\Psi(x, t), \quad (2.1)$$

where  $\Psi(x, t)$  is the wave function that depends on  $x$  and  $t$ ,  $m$  is the mass,  $\hbar$  is Planck's constant divided by  $2\pi$  and  $V(x, t)$  is the potential.

This equation aims to determine the wave function of a determined particle that will depend on the position and time. This equation is analogous to Newton's second law equation which says that  $F = ma$  and determines the position for all future times<sup>15</sup>.

$$m \frac{\partial^2 x}{\partial t^2} = -\frac{\partial V}{\partial x}. \quad (2.2)$$

To get any information from the Schrödinger equation it is imperative to solve it. Considering that  $V$  is independent of time ( $t$ ) we can solve Eq. 2.1 by the separation of variables method and obtain the wave function  $\Psi$ :

$$\Psi(x, t) = \psi(x) \varphi(t), \quad (2.3)$$

where  $\psi$  depends on  $x$ , and  $\varphi$  depends on  $t$ . Therefore, the Schrödinger equation can be rewritten as:

$$-\frac{\hbar^2}{2m} \frac{\partial^2 \psi}{\partial x^2} + V\psi = E\psi, \quad (2.4)$$

which is called the time-independent Schrödinger equation. Or using operator notation  $\hat{H}\psi = E\psi$ , where  $H$  is the so-called Hamiltonian and it is equal to  $-\frac{\hbar^2}{2m} \frac{\partial^2}{\partial x^2}$ . It is relatively easy to solve the time-independent Schrödinger equation for a one body system, but for a many-body system is not simple and the larger the system gets the more complicated the wave function becomes.

### 2.1.2 Electronic structure problem

Consider a non-relativistic time independent system, with  $K$  nuclei and  $N$  electrons. The Schrödinger equation will be written as:

$$\hat{H} \psi(R_1, R_2, \dots, R_K, r_1, r_2, \dots, r_N, \sigma_i) = E\psi(R_1, R_2, \dots, R_K, r_1, r_2, \dots, r_N, \sigma_i). \quad (2.5)$$

Now the wave function is function of all spatial coordinates of the nuclei  $\mathbf{R}_A$ , electrons  $\mathbf{r}_i$  and also of the spin  $\sigma$  (up or down) of a given electron  $i$ <sup>16</sup>.  $E$  is the energy of the eigenstate. The Hamiltonian is a sum of all the possible interactions between electrons and nuclei, and includes potential. It is expressed in atomic units (reduced Planck constant = elementary charge = Bohr radius = electron mass = 1). The units of energy are then Hartree, Eh (i.e., 1 Eh= 27.211396 eV). The Hamiltonian is then written as

$$\hat{H} = -\sum_{i=1}^N \frac{\nabla_i^2}{2} - \sum_{A=1}^K \frac{\nabla_A^2}{2M_A} + \sum_{i=1}^N \sum_{j>i}^N \frac{1}{|r_i - r_j|} + \sum_{A=1}^K \sum_{B>A}^K \frac{Z_A Z_B}{|R_A - R_B|} - \sum_{i=1}^N \sum_{A=1}^K \frac{Z_A}{|r_i - R_A|}, \quad (2.6)$$

here  $A$  and  $B$  runs over all  $K$  nuclei, while  $i$  and  $j$  runs over all  $N$  electrons. Also,  $M_A$  is the mass of atom  $A$ ,  $Z_A$  is the atomic number of nucleus  $A$ . The first two terms of the equation represent the kinetic energy of electrons and nuclei, respectively, and the third and fourth terms represent the Coulomb repulsion between electron and nuclei, respectively. The last term represent the attractive Coulomb interaction between electrons and nuclei. This many-body problem has no analytical solution for a system of more than three particles. A way to overcome this problem was to develop approximations that allow researchers to get accurate results.

### 2.1.3 Born-Oppenheimer approximation

The Born-Oppenheimer approximation assumption states that the wave function can be decoupled into two parts, the electronic part, and the nuclear part<sup>16</sup>. This assumption is based on the fact that the nuclei are much heavier than the electron and therefore the timescale of electron motion is a few orders of magnitude greater than the timescale for nuclei motion. In other words, due to the small mass of the electron, it is possible for them to almost instantaneously respond to any motion of the nuclei allowing electrons to rapidly get into their ground state configuration<sup>16,17</sup>. With this in mind, it is easy to imagine that for electrons the nucleus is apparently stationary and can be treated as classical particles. This allows us to decouple the Hamiltonian into the electronic and nuclear parts and solve the time-independent Schrödinger equation.

As a result of these considerations Eq. 2.6 can be manipulated, the second term of Eq. 2.6 can be neglected or  $M_A = \infty$  for practical purposes; the fourth term became a constant for a fixed configuration of nuclei. Finally, we get the electronic Hamiltonian in Hartree units expressed as

$$\hat{H}_e = - \sum_{i=1}^N \frac{\nabla_i^2}{2} + \sum_{i=1}^N \sum_{j>i}^N \frac{1}{|r_i - r_j|} - \sum_{i=1}^N \sum_{A=1}^K \frac{Z_A}{|r_i - R_A|}, \quad (2.7)$$

or using operator notation

$$\hat{H}_e = \hat{T} + \hat{W}_{ee} + \hat{V}_{ne}, \quad (2.8)$$

where  $T$  is the kinetic operator for electron,  $W$  is the operator for coulomb interaction between electrons and  $V$  is the operator for coulomb interaction between electron and nucleus<sup>18,19,20</sup>.

The Schrödinger equation using the electronic Hamiltonian becomes:

$$\hat{H}_e(r_1, r_2, \dots, r_N)\psi_e(x_1, x_2, \dots, x_N) = E \psi_e(x_1, x_2, \dots, x_N), \quad (2.9)$$

or

$$\hat{H}_e\psi_e(x_i) = E \psi_e(x_i), \quad (2.10)$$

here  $E_e$  is the associated energy and  $x_i = (r_i, \sigma_i)$  is a variable that results from the combination of the spatial coordinates ( $r_i \in R^3$ ) and the spin ( $\sigma = \uparrow$  or  $\downarrow$ ),  $X_i$  is named space-spin coordinates<sup>18</sup>. Since we have a fixed configuration of nuclei, we can suppress their spatial coordinates  $R_A$ <sup>16</sup>.

Now the total energy of the system will include the electronic part and also the nuclear repulsion part, leading to

$$E_{tot} = E_e + E_{nuc}. \quad (2.11)$$

Even though some considerations and modifications have been made to the Schrödinger equation it is still difficult to solve analytically, just keep in mind that the system depends on the spatial coordinates of  $N$  electrons, a total of  $3N$  variables. Such a system may not be easy to handle. Knowing this limitation, some approximation methods have been developed. In the next section we will see some of these methods and how density-functional theory (DFT) was developed.

## 2.2 Density Functional Theory

The density functional theory (DFT) is a theory that instead of using the many-electron wave function it uses the electron density of the system as the main variable to solve the electronic structure of solid, surface, defects, etc<sup>21</sup>. DFT shows a great versatility since, as mentioned before, DFT is implemented to solve the electronic structure of much larger systems which presents more atoms (hundreds or thousands), this versatility is because DFT only uses three spatial coordinates (3 variables) for the electronic density<sup>22</sup>. The importance of DFT rises from the fact that the ground state properties of a system strongly depend on the ground state density<sup>17,23</sup>. DFT is a method with almost 30 years which it has been improved and nowadays it is the preferred method in condensed matter physics. It has also been implemented in computational physics and chemistry.

For DFT the electron density is defined as:

$$\rho(r) = N \int \dots \int |\psi_e(r, r_2, \dots, r_N)|^2 d\sigma_1 dr_2 \dots dr_N, \quad (2.12)$$

and normalizing it:

$$N = \int \rho(r) dr, \quad (2.13)$$

here  $\rho$  is the probability of finding any of the  $N$  electrons at position  $r$ .  $\sigma$  is the spin (up or down) and  $\psi$  is the wave function for all the  $N - 1$  electrons which have arbitrary positions<sup>22</sup>.

### 2.2.1 Thomas-Fermi Theory

Thomas and Fermi were the first to express the electronic energy in terms of the electronic density in 1927<sup>24</sup>. In the Thomas-Fermi (TF) theory, they propose that starting from a uniform electron gas the kinetic energy of the

electrons is a functional of the electron density, but to calculate the total energy of the system two extra terms the nuclear-electron and the electron-electron interaction are treated classically<sup>25</sup> giving 2.14 as result.

$$E(\rho) = C_F \int \rho^{5/3}(r) dr - Z \int \frac{\rho(r)}{r} dr + \frac{1}{2} \int \int \frac{\rho(r_1)\rho(r_2)}{|r_1 - r_2|} dr_1 dr_2, \quad (2.14)$$

Where  $C_F = \frac{3}{10}(2\pi^2)^{2/3} = 2.871$ . The second term is the nucleus-electron interaction and the third term is the electron-electron interaction.

Thomas and Fermi treated electrons as independent particles that do not interact between them. TF method can be used to roughly determine e.g., the electrostatic potential and the charge density<sup>26</sup>, but this model still has some deficiencies and failures.

## 2.2.2 Hartree-Fock Approximation

The Hartree-Fock (HF) approximation is the first step towards fundamental results in approximations<sup>27,28</sup>. The main idea behind HF is that the ground state antisymmetric wave function of interacting  $N$ -electrons can be expressed by a single Slater determinant. This idea rises from considering an  $N$ -electron system, the spin functions, and the variational principle which states that any approximation of the wave functions has a energy above or equal to the ground state energy<sup>27,28</sup>. Therefore, the electronic wave function of the approximation must be antisymmetric and obey the Pauli principle. This antisymmetric wave function can be obtained from the Slater determinant; here the columns are single-electron wave functions and the rows are linear combinations of electron wave functions. These wave functions are also called spin-orbital and come from the product of spatial orbital and spin function. The Slater determinant looks like:

$$\psi(x_i)^{HF} = \frac{1}{\sqrt{N!}} \begin{vmatrix} \phi_1(x_1) & \dots & \phi_N(x_1) \\ \vdots & \ddots & \vdots \\ \phi_1(x_N) & \dots & \phi_N(x_N) \end{vmatrix}. \quad (2.15)$$

The antisymmetric property can be proved by considering two-electron system:

$$\psi(x_1, x_2) = \frac{1}{\sqrt{2!}} [\phi_1(x_1)\phi_2(x_2) - \phi_2(x_1)\phi_1(x_2)], \quad (2.16)$$

also, the electron wave functions are orthonormal,  $\langle \phi_i(x_i) | \phi_j(x_j) \rangle = \delta_{ij}$ . This leads to the Schrödinger equation and one-electronic Hamiltonian of the form:

$$\hat{F}_i \phi_i(x_i) = \epsilon_i \phi_i(x_i), \quad \hat{F}_i = -\frac{\nabla^2}{2} + V^{HF}(x_i) + V_i^{xc}(x_i), \quad (2.17)$$

where  $\epsilon_i$  and  $\phi_i$  corresponds to the eigenvalues and eigenvectors. The first term is the kinetic energy of N independent electron and the second term is the external potential,  $V_i^{xc}(x_i)$ , which is the Coulomb attraction on  $i^{th}$  electron due to all the nuclei. The third and fourth terms approximately account for the many-body electron-electron interactions. The Hartree potential is denoted by  $V^{HF}(x_i)$ .

The HF approximation has a special method to be solved, this method is called the self-consistent-field (SCF). What this method does is that with an initial guess of the spin orbitals, the Hartree potential can be calculated. Then with  $V^{HF}$  the eigenvalue can be calculated for new spin-orbital. This new spin-orbital is then used to repeat the procedure until self-consistency is reached, or in other words, the spin-orbital does not change any more<sup>27,28</sup>.

The main idea of HF theory is to instead of considering a many-electron system, the system is now a one-electron problem in which electron-electron repulsion is an average effect.

### 2.2.3 Hohenberg-Kohn Theorem

We can say that the basis of DFT is the connection between the electron density and unique external potential that allows expressing the Hamiltonian as functional of the density ( $\rho(r)$ ). This idea was developed by Hohenberg and Kohn<sup>29</sup>. Hohenberg-Kohn theorem (HK) theorem has two key statements, the first says:

“There exist a unique external potential  $V_{ext}(r)$  determined by the ground state electron density  $\rho(r)$ ”.

Proof:

Let us consider two external potentials  $V_{ext}(r)$  and  $V'_{ext}(r)$  that at the ground state gives the same electronic density  $\rho(r)$ . A Hamiltonian would depend on the potential meaning that there are two Hamiltonians  $H$  and  $H'$  which have the same ground state density. But since there exist two Hamiltonians two different wave functions also exist  $\psi$  and  $\psi'$ , which gives  $H\psi = E_0\psi$  and  $H'\psi' = E'_0\psi'$ . Then by applying the variational principle:

$$E_0 < \langle \psi' | \hat{H} | \psi' \rangle, \quad (2.18)$$

$$\langle \psi' | \hat{H}' | \psi' \rangle + \langle \psi' | \hat{H} - \hat{H}' | \psi' \rangle, \quad (2.19)$$

$$\langle E'_0 - \int \rho(r) [v(r) - v'(r)] dr, \quad (2.20)$$

and

$$E'_0 < \langle \psi | \hat{H}' | \psi \rangle, \quad (2.21)$$

$$\langle \psi | \hat{H} | \psi \rangle + \langle \psi | \hat{H}' - \hat{H} | \psi \rangle, \quad (2.22)$$

$$\langle E_0 - \int \rho(r) [v(r) - v'(r)] dr. \quad (2.23)$$

By adding equation 2.20 and equation 2.23 the following is obtained.

$$E_0 + E'_0 < E'_0 + E_0. \quad (2.24)$$

Which is a contradiction. This means that there cannot be two external potential  $V(r)$  that gives the same ground state electronic density  $\rho(r)$ . Thus  $\rho(r)$  determines  $V(r)$  uniquely.

Now, if we integrate  $\rho(r)$  it yields  $N$ , it also determines the kinetic energy, and the total energy can be rewritten as:

$$E[\rho] = T[\rho] + E_{ne}[\rho] + E_{ee}[\rho] = \int \rho(r)V(r)dr + F_{HK}[\rho], \quad (2.25)$$

where

$$F_{HK}[\rho] = T[\rho] + V_{ee}[\rho], \quad (2.26)$$

or

$$F[\rho] = \langle \psi[\rho] | \hat{T} + \hat{V}_{ee} | \psi[\rho] \rangle. \quad (2.27)$$

As result we get the full Hamiltonian which is a universal functional of  $\rho(r)$  and therefore all the properties of the ground state can be calculated.

The second statement says:

“By applying the variational theorem, the ground state energy can be obtained”.

This means that the lowest energy of the system (the ground state energy) is obtained if and only if the minimized  $E[\rho]$  of the system has as input the real ground state density.

$$E_0 = \min_{\rho} \left\{ F[\rho] + \int V_{ne}(r)\rho(r)dr \right\}. \quad (2.28)$$

By knowing  $F_{HK}$  the Schrödinger equation for DFT could be solved exactly. But, the first HK theorem does not say anything about the form of this term, it just says that it exists<sup>20,25</sup>. The explicit form of the kinetic term and the electronic term is the major challenge of DFT.

### 2.2.4 The Kohn-Sham Equations

Since DFT theory was not a complete because there was not a full expression for the ground state energy and still had some failures, Kohn and Sham in certain way correct the equations by changing the considerations in the kinetic energy made by Thomas and Fermi and developed a method to approximate the ground state density and therefore its energy<sup>30</sup>. They consider a non-interacting system with the same density as the real one, and rewrite  $F_{HK}$ .

$$F_{HK}[\rho] = T_s[\rho] + E_{cl} + E_{xc}. \quad (2.29)$$

Where  $T_s$  is the kinetic energy of the non-interacting system and the electron-electron interaction is separated into the classical part or Hartree energy ( $E_{cl}$ ) and the exchange-correlation energy ( $E_{xc}$ ). All the unknown non-classical terms (the many-body quantum effects) are put into this  $E_{xc}$  term. So, the energy functional becomes:

$$E[\rho(r)] = \int \rho(r) V(r) dr + T_s[\rho(r)] + E_{cl}[\rho(r)] + E_{xc}[\rho(r)]. \quad (2.30)$$

With equation 2.30 the Euler-Lagrange equation can be calculated and therefore an expression for the effective potential  $V_{eff}$ <sup>25</sup>.

$$\delta E[\rho(r)] = \int \delta\rho(r) \left\{ V_{eff}(r) + \frac{\delta}{\delta\rho(r)} T_s[\rho(r)] - \varepsilon \right\} dr, \quad (2.31)$$

where

$$V_{eff}(r) = V(r) + V_{xc}(r) + \int \frac{\rho(r')}{|r-r'|} dr', \quad (2.32)$$

and

$$V_{xc}(r) = \frac{\delta}{\delta\rho(r)} E_{xc}[\rho(r)], \quad (2.33)$$

where  $V_{xc}$  is the exchange correlation potential and  $\varepsilon$  is the Lagrange multiplier to assure particle conservation<sup>25</sup>.

With these equations the central Kohn-Sham DFT equation can be written as

$$\left( -\frac{1}{2}\nabla^2 + V_{eff}(r) - \varepsilon_j \right) \varphi_j(r) = 0, \quad (2.34)$$



with

$$\rho(r) = \sum_{j=1}^N |\varphi_j(r)|^2, \quad (2.35)$$

and  $\varepsilon_j$  are the one-electron orbital energies.

Now, the ground state energy is given as

$$E = \sum_j \varepsilon_j + E_{xc}[\rho(r)] - \int V_{xc}(r) \rho(r) dr - \frac{1}{2} \int \frac{\rho(r)\rho(r')}{|r-r'|} dr'. \quad (2.36)$$

Note that  $V_{eff}$  and  $V_{xc}$  depend on the density of the system. Equations 2.32, 2.34 and 2.35 are the self-consistent Kohn-Sham equations.

There exists a procedure that in principle will solve the KS equations. It starts with guessing the charge density  $\rho(r)$ , calculate the effective potential ( $V_{eff}$ ) and obtain the KS orbitals<sup>16,23</sup>. But the issue here is that there is not a full expression for  $E_{xc}$ . So, the only solution is to approximate the exchange-correlation Energy ( $E_{xc}$ ) to obtain an approximation of the exchange-correlation potential ( $V_{xc}$ ). Once the first approximation was made, it is used to calculate a new form of the ground state energy (2.36) iteratively until the energy converges within some tolerance range. With the energy it is possible to determine the electron density. In the next section the most used approximations are discussed.

## 2.2.5 Exchange Correlation Functional

As mentioned in the Kohn-Sham section the exchange-correlation energy functional is the only term that remains unknown. However, there exist some approximations that may help to get a value for  $E_{xc}$ . These different approximations can be ranked or ordered according to their accuracy, precision, and complexity; this ranking is named Jacob's ladder (see Figure 2.1)<sup>31</sup>. Some models separate the  $E_{xc}$  into correlation part  $E_c$  and the exchange part  $E_x$ , in that way it is easy to deal with the  $E_{xc}$ .

The selection of an approximation would be determined based on the complexity of the system under study, this means that in some cases selecting the basics ones would be enough to get good results. Some of the most commonly used approximations are the local density approximation (LDA), generalized gradient approximation (GGA), meta-GGA, and Hybrid functional which are going to be discussed next.

## 2.2.6 Local Density Approximation

Local density approximation can be considered as the basic approximation within DFT; it is in the lower section of Jacob's ladder<sup>31</sup>. In LDA the objects under study are infinitesimal volumes of a uniform electron gas, on which a

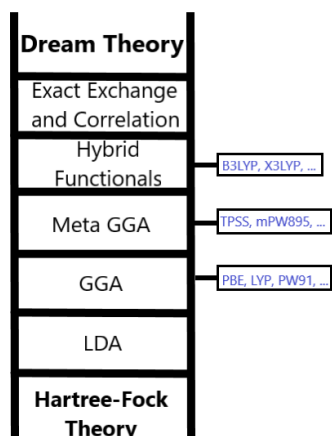


Figure 2.1: Schematic diagram of Jacob's Ladder of density functional approximations proposed by Perdew and Schmidt in 2001<sup>31</sup>

constant external potential act and the charge density varies slowly. In these infinitesimal volumes, the exchange-correlation energy is the same. In this approximation  $E_{xc}$  is defined as<sup>30</sup>.

$$E_{xc}^{LDA} = \int \rho(r) e_{xc}(\rho) dr, \quad (2.37)$$

where  $e_{xc}(\rho)$  is the exchange-correlation energy of an infinitesimal part of the gas for a density  $\rho$ . Here  $e_{xc}(\rho)$  is given in atomic units<sup>25</sup>. In practical applications of LDA,  $E_{xc}$  is split into the exchange part  $e_x$  and the correlation part  $e_c$ , therefore  $e_{xc}(\rho) = e_x(\rho) + e_c(\rho)$ .

$e_x$  is given as:

$$e_x(\rho) = \frac{0.458}{r_s}, \quad (2.38)$$

where  $r_s$  is the radius of a sphere that contains one electron<sup>25</sup> and is defined as:

$$r_s = \left( \frac{3}{4\pi\rho} \right)^{1/3}. \quad (2.39)$$

Unfortunately, the correlation part cannot be solved analytically and Quantum Monte Carlo calculations are needed. Such a complicated term needs to be parametrized<sup>18</sup>. There exist two preferred parametrizations that are

accurate enough to be used, these are the Vosko, Wilk, and Nusair (VWN) parametrization<sup>32</sup> and the Perdew and Wang (PW92)<sup>31</sup> parametrization.

### 2.2.7 Generalized Gradient Approximations

The LDA was a good start for DFT but it was not enough. It was realized that considering a uniform electron gas will not work for all the different structures where the density is not uniform and vary rapidly such as in molecules. To improve the results from such cases the solution was to consider a gradient of the electronic density. This approximation called Generalized Gradient Approximation (GGA) has a development that arise from LDA.

In order to improve approximations, the first and logical step was to add a gradient portion to LDA, this was done by adding an external potential that slowly varies, and then expand the  $E_{xc}$  with respect to the gradient density, this was named gradient-expansion approximation (GEA). But the low-order gradient corrections ( $\nabla\rho$  and  $\nabla^2\rho$ ) did not work and in some cases made even worse approximations than LDA<sup>33</sup>.

This failure helps to understand that the solution was not to use a power series expansion but to use more general functions. That is how GGA was developed and has a generic form:

$$E_{xc}^{GGA}[\rho] = \int f^{GGA}(\rho(r), \nabla\rho(r)) dr, \quad (2.40)$$

where  $f$  is some function.

GGA are also called “semi-local” approximations because of the density gradient. Notice that as new considerations are taking into account, the new approximations get placed higher in the Jacob’s ladder. There exists numerous GGA that differ in  $f$ , some of them includes spin-density dependent, and even some functions take into account the gradient of the spin density.

There exist several GGA but some of them are used more often such is the case of the Becke 88 (B88)<sup>34</sup>, Lee-Yang-Parr (LYP)<sup>35</sup>, the Perdew-Wang 91 (PW91)<sup>36</sup>, and the most used GGA functional is the Perdew-Burke-Ernzerhof (PBE)<sup>37</sup>. These techniques are still used today and work is being done to improve them<sup>38</sup>. The use of GGA reduce the LDA errors of atomization energies, by a factor of 3-5 for small molecules<sup>25</sup>.

### 2.2.8 Meta-Generalized Gradient Approximations

The meta-generalized-gradient approximation belongs to the third generation or third stair of the Jacob’s ladder and it takes into account the second derivative of the gradient or the Laplacian of the density, and/or the kinetic energy gradient  $\tau(r)$ .

$$E_{xc}^{mGGA}[\rho] = \int f^{mGGA}(\rho(r), \nabla\rho(r), \nabla^2\rho(r), \tau(r)) dr, \quad (2.41)$$

and,

$$\tau(r) = \frac{1}{2} \sum_{i=1}^N |\nabla \phi_i(r)|^2, \quad (2.42)$$

where  $\phi_i$  are the Kohn-Sham orbitals. For meta-GGA we can also take into account the spin and equation 2.41 becomes a bit more complicated.

## 2.2.9 Hybrid functional

In 1993 Becke<sup>39</sup> proposed an adiabatic-connection formalism which propose to combine a Hartree-Fock (HF) exchange energy  $E_{HF,x}$  and a GGA functional, here HF is an exact exchange functional<sup>40</sup>. One common formula for the adiabatic formalism is:

$$E_{xc} = \int_0^1 U_{xc}^\lambda d\lambda, \quad (2.43)$$

where  $\lambda$  is an interelectronic coupling strength parameter that switches on the  $\frac{1}{|r_i-r_j|}$  Coulomb repulsion between electrons and  $U_{xc}^\lambda$  is the potential energy of exchange with a coupling strength  $\lambda$ . This equation connects the non-interactive Kohn-Sham system  $\lambda = 0$  with a real interactive system  $\lambda = 1$ , all with the same density  $\rho(r)$ . Given that at this point a combination of functionals takes place, it is expected that it gives better results for the exchange-correlation functional than GGA or LDA alone.

There exists different hybrid functional that differs on the number of parameters they have. It is known that the most widely used hybrid functional is the Becke, 3-parameter, Lee–Yang–Par (B3LYP), this method use three parameters  $a_0 = 0.20$ ,  $a_1 = 0.72$ ,  $a_2 = 0.81$  and LYP is the correlation functional and B88 is the exchange functional<sup>41</sup>. B3LYP has the following form:

$$E_{xc} = E_{xc}^{LDA} + a_1(E_x^{HF} - E_x^{LDA}) + a_2 E_x^{GGA} + a_3 E_c^{GGA}. \quad (2.44)$$

The parameters obtained by Stephens in 1994<sup>41</sup> can also be calculated with fitting the experiment data. By varying equation 2.44 we can obtain other approximations, for example, instead of three parameters only one parameter and one GGA functional are considered.

## 2.3 Basis Sets

Mathematically talking, a basis set is a collection of vectors that spans and define a space on which a problem is solved, for example  $i, j, k$  defines the Cartesian space. A basis set in quantum computation can be seen as the one-particle function used to build molecular orbitals. Basis set are used to calculate the  $\psi$  of the electronic Schrödinger

equation. Generally, the unknown  $\psi$  is expanded in terms of known basis functions. The expansion cannot be done infinitely, it is necessary to have a finite set of functions, these basis functions can be any function or family of functions, but it needs to fulfill some requirements:

- For an isolated atom or molecule, they should decay when the distance between electrons and nuclei is large just like it happens in reality.
- The computational cost has to be low and the calculations need to be accurate enough.

Basis set is also called the linear combination of atomic orbitals theory (LCAO). It is important to mention that an unknown molecular orbital can only be completely represented by an infinite number of functions, but for practical calculations this is impossible<sup>42</sup>. There exists different type of basis functions such as:

### Slater type orbitals

Slater-type orbital (STO) are used to calculate electronics structures. The STO have the following form:

$$\chi_{\zeta,n,m,l}(r, \theta, \varphi) = NY_{l,m}(\theta, \varphi) r^{n-1} e^{-\zeta r}, \quad (2.45)$$

where  $N$  is a normalization constant,  $Y_{l,m}$  are the spherical harmonic functions and  $\zeta$ ,  $n$ ,  $m$  and  $l$  are the radius of the orbit and quantum numbers: principal, angular momentum and magnetic, respectively. These types of functions have an exponential decay as the distance to the nuclei increase. This dependence on distance is what allows STO to perfectly describe the hydrogen-like atoms<sup>42</sup>. Also due to the exponential dependence it has a rapid convergence if the number of functions increase, but contrary to this advantage, the calculations of three- and four-center two-electron integrals has no analytical solution and the computational time increase rapidly. This is why STO are essentially used for one-atom or two-atom systems.

### Gaussian type orbitals

Gaussian type orbital (GTO) present a better calculation time than STO. For this simple reason GTO are the preferred function in quantum chemistry<sup>42</sup>. GTO has the following form:

$$\chi_{\zeta,l_x,l_y,l_z}(x, y, z) = Nx^{l_x}y^{l_y}z^{l_z}e^{-\zeta r^2}. \quad (2.46)$$

In this case we do not have the quantum numbers, instead  $l_x$ ,  $l_y$ , and  $l_z$  are the new parameters. These parameters have the particularity that when they are added,  $L = l_x + l_y + l_z$ ,  $L$  is the angular momentum for atoms and determines the type of orbital e.g. s-orbitals ( $L = 0$ ), p-orbitals ( $L = 1$ ), d-orbitals ( $L = 2$ ) and f-orbitals ( $L = 3$ ). GTO are very efficient in terms of computational cost. However, GTO has two specific disadvantages, it does not work near the nucleus ( $r = 0$ ) where it has zero slope and very far from it where it falls off rapidly. This is because of the  $r^2$  exponent-dependence of GTO compared to the  $r$  exponential-dependence of STO makes GTO poorly describe the system with the same amount of functions. By adding more functions, the accuracy of GTO is increased and roughly

talking three times more GTO's than STO's are needed to achieve the same accuracy level of STO<sup>42</sup>. One might think that with three times more GTO's the efficiency will decrease, but the required integrals can be easily solved and therefore the computational efficiency does not exceed the time needed for STO.

Equation 2.46 is also called primitive Gaussian and as mentioned before several primitive Gaussian's are group into one big Gaussian known as contracted GTO's which is a linear combination of Gaussian's equation 2.47.

$$g(c) = \sum_i a_i g_i(p), \quad (2.47)$$

where  $c$  and  $p$  designate contracted and primitive.

It is important to determine the number of functions to be used, this will depend on the type of basis function (STO/GTO) and the element of the periodic table which is being studied. In theory, just the minimum or single-zeta (SZ) basis set is needed to describe an occupied orbital of a neutral atom. However, this statement is not applicable in practical problems. Normally it is needed to use more than one basis set (SZ), generally two basis set which is called double zeta (DZ) or triple zeta (TZ). The number of basis sets can grow and continue to higher numbers<sup>16,42</sup>. In some cases, it is important to also take into account higher angular momentum functions called polarization functions.

### Def2 basis set

Def2 is a family of Gaussian basis set that were included in the computational program called TURBOMOLE<sup>43</sup>. These family of basis set has almost the same level of accuracy for all the elements of the periodic table<sup>44</sup>. The characteristic of this family is that they can implement different functions for polarization and valance electrons. This is the case of the def2-TZVPP which is a Gaussian function with a triple zeta valance (TZV) basis function and a large polarization function (PP)<sup>42,44,45</sup>. These functions are optimized to be use at the HF level<sup>45,46</sup>. The def2/J is an auxiliary basis set used to fit the Coulomb interaction.

## 2.4 Tight Binding

The free electron model was good enough to deal with single atoms or systems that are not so complicated. But if the atoms start to get close to each other and form solids or crystals this model is no longer appropriate. The tight binding model aims to describe the electronic structure of crystals or big molecules. The general idea is that when atoms get close to each other their wave function overlap. With this in mind tow cases can happen; the wave functions of atom  $A$  and  $B$  are added or subtracted, leading to two energy levels. For a system with  $N$  atoms,  $N$  orbitals are formed for each orbital in the atom<sup>47</sup>. Because of the coulomb interaction between atoms in a crystal the energy levels are split into bands and the width of the bands is proportional to the overlapping of near neighbor atoms. Degenerate energy states will form different bands. Tight binding method is also called Linear Combination of Atomic Orbital (LCAO)<sup>47,48</sup>.

## 2.5 RIJCOSX Algorithm

The RIJCOSX approximation is an algorithm used to speed up the calculations of the Hartree-Fock and Hybrid Density Functional Theory<sup>49</sup>. It is basically a combination of two algorithms the split RI-J and the COSX that aims to deal with Coulomb and exchange parts of the Hartree approximation. The split RI-J<sup>50</sup> approximation uses Gaussian basis functions and it is used to describe the Coulomb interactions specially in the near-field part. The Chain-Of-Spheres exchange (COSX) is an algorithm that implements semi-numerical approximation for the exchange matrix<sup>27,49</sup>. The RIJCOSX algorithm has been proved and shows that the self-consistent calculations are highly accurate and efficient<sup>49</sup>.

## 2.6 Dispersion correction

In DFT some important interactions that are not properly treated because they are small in magnitude or occurs at far distance from the local vicinity of the atom. Some of these interactions are the van der Waals interaction and among them is the long-range London interaction. It is well known that these interactions play an important role on chemical and physical accuracy<sup>51</sup>. To make DFT and HF approximation more accurate it is necessary to take into account the van der Waals interaction. There exist various approximations but the ones used in this thesis is the Becke-Johnson<sup>52</sup> and the damped DFT proposed by Grimme<sup>51,53</sup>.

The Becke-Johnson model provides an excellent treatment to the van der Waal interactions because it considers bigger intermolecular dispersion coefficients. In concrete this method adds to the HF model a dynamical correlation energy terms and a dispersion term of the form<sup>52</sup>.

$$E_{total} = E_{HF} + E_C^{BR} + E_{disp}, \quad (2.48)$$

where  $E_C^{BR}$  is the dynamical correlation energy term<sup>54</sup>.

The damping DFT (DFT-D3) method simply adds an atom-pairwise specific dispersion coefficient to the result in Kohn-Sham DFT<sup>51,53</sup>.

$$E_{DFT-D3} = E_{KS} - E_{disp}, \quad (2.49)$$

where  $E_{KS}$  is the self-consistent energy obtained from KS model and  $E_{disp}$  is the dispersion correction for two- or three- body.

## 2.7 Ab Initio Method

*Ab Initio* methods aims to study the electronic structure of solids, surfaces, etc, with a good computational accuracy-time ratio. They use the same idea behind the Born-Oppenheimer adiabatic approximation<sup>55</sup> and the self-consistent

Hartree-Fock approximation<sup>56</sup>. Here the wave function can be rewrite as a linear combination of Slater determinants. *Ab Initio* method allows to generate accurate results but when dealing with larger system or condense-phase systems the computational time becomes enormous and the computational effort way to expensive. For these reason the methods are used as complementary, for example to calculate small regions of a system<sup>57</sup>.

## 2.8 Sesquiterpene Lactones

*Sesquiterpene Lactones* (STL) are a group of secondary metabolites isolated from plants belonging to the *Asteraceae* (*Compositae*) family<sup>1</sup>. These organic molecules have a wide variety of chemical structures and have shown a wide spectrum of biological activities, which is “the capacity of a specific molecular entity to achieve a defined biological effect” on a target<sup>58</sup>. The biological activity is measured in terms of the needed concentration to generate the specific effect<sup>59</sup> that can include antimicrobial, anti-fungal, anti-inflammatory, anticancer, among others<sup>2</sup>. It is also well known that some STL are toxic to human and animal parasites<sup>2,4</sup>, because of these two characteristic STL are of great interest in this thesis. These molecules have been used mainly in the pharmaceutical industry and in ancient medicine.

*Asteraceae* family is one of the most widely distributes and abundant family of plants, this had allowed them to evolve and develop different chemical structures which is one of the reasons why they present different biological activities. Currently, it has been possible to identify the structure of several of these molecules and in the Dictionary of Natural Products (DNP), there are at least 5000 different structures for STL<sup>3</sup>.

The chemical structure of STL is very particular, it is characterized by a ring of carbons (typically from 6 to 10 atoms)<sup>60</sup>. The  $\gamma$ -lactone ring containing an  $\alpha$ -methylene group is a common feature that most of the STL shares in their structures<sup>2</sup>, but it is not always observed. There exist many other features that can vary from one group of STL to other and this is because, as mentioned before, the *Asteraceae* family is one of the most widely distributed and abundant plant families. In simple word their structures are as varied as their biological diversity.

*Sesquiterpene lactones* can be classified by their structure, the most important groups are the Germacranes, Eudesmanes, Elemans, Eremophilanes, Guaianes, Xanthanes and Pseudoquaianes<sup>3</sup>. To get a general idea of how the main groups are structured, we can describe them in a general way and without going into detail about the possible functional groups, since if we do that, we would have a very large number of possibilities, not to say infinite. Germacranes are generally formed by a central ring of 10 carbons; the Eudesmanes and Eremphilanes are made up of a 6- carbon bicyclic ring, only one ethyl group changing position; Guaianes and pseudoquaianes have a 5/7 carbon bicyclic ring; for the Elemans the structure is made up of a ring of 6 and several methyls and ethyl groups; just like for the Xanthanes we have a single central ring of 7 carbons and a butyl group<sup>3</sup>. But it is to be expected that the naturally formed structures that can be found in plant extracts are more complex and with many more extra elements than the structures previously described. In this research, we will focus on studying those STL that have shown great antibacterial activity and for which there is a record of their toxicity.



## 2.9 Cytotoxicity

Cytotoxicity often used in biology, medicine or chemistry. It is defined as “the toxicity caused due to the action of chemotherapeutic agents on living cells”<sup>61</sup>. For practical purposes cytotoxicity is an “*in vitro* test to determine whether the medical device will cause any cell death due to leaching of toxic substances or from direct contact”<sup>62</sup>. Determining the cytotoxicity of a molecule is of importance since the possible applications of it might depend on how much of this agent. There exist different cytotoxic methods the common ones are when a cell destroys the cell membrane of a target cell, or when it prevents protein synthesis, or when it binds to receptors, etc<sup>63</sup>.

## 2.10 Methicillin-resistant *Staphylococcus aureus*

Methicillin is a semi-synthetic penicillin<sup>64</sup> (antibiotic) used to treat bacterial infections. *Staphylococcus aureus* is a bacteria member of the *Micrococcaceae* family that cause infections of skin, soft-tissue, respiratory among others<sup>65</sup>. *Staphylococcus Aureus* methicillin-resistant (MRSA) is a pathogen that emerged in the earliest 1960<sup>66</sup> that developed resistance to methicillin. The Minimum inhibitory concentration (MIC) is defined as the lowest concentration of antimicrobial at which microorganism visible stop growing<sup>67</sup>. The effectiveness of new antibiotics against bacteria usually is measured or reported in MIC in units of mass over volume [mass/vol].

## 2.11 Quantitative Structure-Activity Relationship

Quantitative Structure-Activity Relationship (QSAR) is an *in silico* statistical method uses to establish relationships between the molecular structure of compounds (organic and inorganic)<sup>68</sup> and the physical or biological activity. This method aims to construct a model to describe the above, describe relationships and to predict the possible physico-chemical and biological activity of a group of (untested, new) compounds<sup>69</sup> at the early stage of design (before synthesis). In results designer may save time and reduce cost of whole process as well as reduced number of animal testing (if needed)<sup>70</sup>. QSAR models have been used in different areas of research such as biology<sup>1,3,11–13,71–74</sup>, chemistry and even for nanoparticles in physics<sup>68,70,75,76</sup>.

The general process to build a QSAR model can be summarized in the following way<sup>77,78</sup>:

1. Determination and selection of the data set
2. Molecular modeling and generation of the structural data
3. Development and calculation of structural descriptors
4. Model development based on the selected descriptor
5. Analysis and interpretation of the model
6. Estimation of the validity and predictability of the model.

The starting point for a QSAR model is generally based on the available information about the molecule and the physicochemical or biological activity. It is important to mention that all the data points have to come from the same procedure, they have to be homogeneous. This means that they have to follow the same process in order to be used in the same model. Different data points obtained from different procedures can lower the quality of the model<sup>70</sup>.

The generation of the descriptors is a critical step in the development of QSAR model. The structure itself cannot be used for QSAR model because it does not explicitly contain the information related to the activity, it has to be extracted from it. Also, because most methods use uniform numerical vectors as inputs for all molecules, and molecular structures have diverse size, shape, orientation and forms, so they do not fit in the models<sup>69</sup>. Mathematically QSAR can be seen as<sup>78</sup>:

$$P_i = f(D_i), \quad (2.50)$$

where  $P_i$  is the specific activity or property of a compound  $D_i$ ,  $f(D_i)$  is a function that depends on  $D_i$ . Generally, a linear function is not sufficient to generate a good QSAR model, instead a multiple linear regression is preferred.

$$P = f(D_1, D_2, D_3, \dots, D_n). \quad (2.51)$$

For our purposes  $P$  would be taken as the biological activity and  $D_1, D_2, D_3, \dots, D_n$  are the structural properties (descriptors). The relation between  $P$  and  $D_i$  could be linear or not.

### 2.11.1 Descriptors

Actually, what is used to build a QSAR model are called descriptors. A descriptor is a parameter of the system which varies with a desired property of the system. It describes qualitatively the expected value of the desired property. In other words it is a mathematical representation of a molecule, obtained as an output of a well-specific algorithm which is applied to a specific and defined molecular representation<sup>79</sup>. Descriptor can be theoretical calculations or experimental physico-chemical properties of molecules.

Descriptors are obtained by applying different theories such as quantum mechanics, DFT, graph theory. As mentioned, descriptors encode specific information about the molecule, information such as toxicology, pharmaceutical, physical, electronic, etc<sup>79,80</sup>. Descriptors can be classified by the dimensionality of the structure from which they are derived (0D, 1D, 2D, 3D)<sup>79,81,82</sup>. The 0D –descriptors are obtained directly from the formula (e.g., number of atoms of an element); the 1D –descriptors are structural fragments<sup>76</sup>; 2D –descriptors are topological or geometrical properties of the molecule<sup>69,82</sup>; 3D-descriptors take into account distances, angles, size, volume and other geometrical parameters<sup>69,79,81</sup>.

Contrary to what might seem intuitively when performing a QSAR model, the number of descriptors is not the same as the number of compounds. In fact, there exists a “rule” which suggests that for every 6 to 10 data points just one descriptor is needed<sup>83</sup>. We follow this rule in our study to obtain accurate results. The next step is to select the more representative descriptors and develop a function that links the values of the descriptors and the analyzed

activity<sup>69</sup>.

### 2.11.2 Endpoint

Endpoint is defined as the measurement of any activity for chemical made under specific conditions (following the same experimental protocols)<sup>84</sup>. Endpoints are a crucial part of QSAR models since they are the observable expression of a specific phenomena that the model will predict. For QSAR model, it is ideal to have all endpoints obtained following the same procedure (homogeneous data set), but it is common to mix different data sets from different protocols. To decrease the impact of mix data set a "defined endpoint" can be use, it is referred to any physicochemical (boiling, melting point, etc), biological (antibacterial, anti-inflammatory activity, etc) or environmental effect related to chemical structures shows that can be measured and modelled<sup>84</sup>.

### 2.11.3 Statistical variables

To have a tangible idea of how good a QSAR model is, an external and internal validation is needed. There exist three specific statistical variables that can be calculated for each model based on the data on which the are build to that will gives information about the model quality. These variables are the determination coefficient in the training set  $R^2$ , the leave-oneout cross-validation determination coefficient  $Q_{CV}^2$  and the determination coefficient for the validation set  $Q_{Ext}^2$ <sup>85</sup>. These variables give information about the goodness-of-calibration, robustness and predictability capacity. They are defined as:

$$R^2 = 1 - \frac{\sum_{i=1}^n (y_i^{obs} - y_i^{pred})^2}{\sum_{i=1}^n (y_i^{obs} - \bar{y}^{obs})^2}, \quad (2.52)$$

$$Q_{CV}^2 = 1 - \frac{\sum_{i=1}^n (y_i^{obs} - y_i^{cvpred})^2}{\sum_{i=1}^n (y_i^{obs} - \bar{y}^{obs})^2}, \quad (2.53)$$

$$Q_{Ext}^2 = 1 - \frac{\sum_{j=1}^{nval} (y_j^{obs} - y_j^{pred})^2}{\sum_{j=1}^{nval} (y_j^{obs} - \bar{y}^{obs})^2}, \quad (2.54)$$

and the errors for equations 2.52, 2.53 and 2.54 are respectively defined as:

$$RMSE C = \sqrt{\frac{\sum_{i=1}^n (y_i^{obs} - y_i^{pred})^2}{n}}, \quad (2.55)$$

$$RMSE CV = \sqrt{\frac{\sum_{i=1}^n (y_i^{obs} - y_i^{cvpred})^2}{n}}, \quad (2.56)$$

$$RMSEP = \sqrt{\frac{\sum_{j=1}^{nval} (y_j^{obs} - y_j^{pred})^2}{nval}}, \quad (2.57)$$

here  $y_i^{obs}$  is the observable endpoint of compound  $i$ -th,  $y_i^{pred}$  is the predicted or calculated value of the endpoint for  $i$ -th compound,  $\bar{y}^{obs}$  is the mean values for all observable  $y$ ,  $y_i^{cvpred}$  is the predicted observable value for compound  $i$ -th using a model calibrated without using  $i$ -th compound (this  $i$ -th compound is not taken into account temporarily for the model calibration),  $n$  is the number of compounds in the training set, and  $nval$  is the number of compounds in the validation set. Ideally a good QSAR model will have values for  $R^2$ ,  $Q_{CV}^2$  and  $Q_{Ext}^2$  close to 1 and errors closer to 0 as possible.

# Chapter 3

## Methodology

### 3.1 Computational Methods

#### 3.1.1 Conformer-Rotamer Ensemble Sampling Tool

Conformer-Rotamer Ensemble Sampling Tool (CREST) is a semiempirical tight-binding computational method. It is specialized in obtaining proper thermodynamic conformers of STL (and in general of any organic or inorganic molecules) at a quantum chemical level. A conformer is an ensemble of low-energy structure, from a chemical point of view conformers are stereoisomers of a molecule with different spatial conformation. Conformers are a specific molecular conformation that carry information about the properties of the molecule at a specific temperature, and can be differentiated by the potential energy minimum<sup>86</sup>. A graphical description of how conformers can be differentiated is showed in Figure 3.1.

CREST is a software specialized on finding and determining the possible conformers of a molecule. Since this code implements semi-empirical methods the computational cost is low when working with molecules that has hundreds of atoms. CREST has implemented two different conformational search procedures MF-MD-GC (combination of mode following, molecular dynamics sampling, and genetic z-matrix crossing) and iMTD-GC (metadynamics combiend with an CG step). The difference between them is that MF-MD-GC implements a molecular dynamics sampling, and iMTD-GC implements an extensive metadynamics sampling<sup>86</sup>.

The general procedure of CREST starts with quantum-chemical method that optimize the structure and then a Conformer Rotamer Ensemble (CRE) is calculated within a certain energy window, the smaller the energy window the greater accuracy is obtained. These calculations use meta-dynamics simulation and the root-mean-square deviation (RMSD) combined with energetic for structure comparison. For the code, conformers are those structures that either differ on the potential energy surface (PES) or RMSD and rotational constant ( $B_e$ ). In the meta-dynamics part, a history-dependent biasing potential is applied, and previous structures are used as collective variables. By

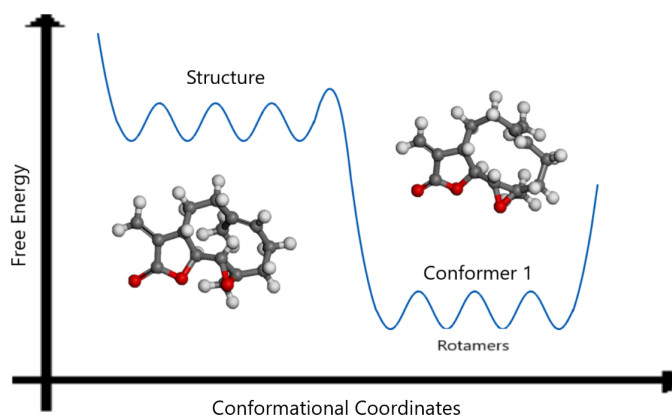


Figure 3.1: Graphical description of conformers and the difference in energy between possible conformations. Here the oscillations represents how the energy fluctuates and allows the molecule to adapt an specific conformation within that specific range of energy.

using this type of potential that only grows, is possible to explore huge regions of the PES and prevent calculations in regions that were already explored<sup>86</sup>.

It is difficult to determine the “true” conformation of any molecule without experiments to confirm. But even with experiments the “true” conformation is hard to identify since it will depend on the experimental conditions, meaning that it changes. That is why it is important to use theoretical calculations. With theory the "experimental conditions" can be fixed and the real conformation at that specific conditions will appear. CREST also allow us to protonate molecules, in our case the STL structures. Protonation is the process were we add a hydrogen atom [ $H^+$ ] to the structure<sup>87</sup>, so we generate a new structure. In general we use protonation procedure to destroying a double bond in the original structure and add the [ $H^+$ ] atom.

### 3.1.2 ORCA

ORCA is an *ab initio* quantum-chemistry program package developed in 1999 by the group of Frank Neese, which implements many electronic structure methods such as semiempirical, DFT and others<sup>88</sup>. In this thesis ORCA is used to compute accurately the electronic properties of the STL. For this purpose, the level calculation used hybrid functional B3LYP within the RIJCOSX approximation with def2-TZVPP basis set, def2/J auxiliary basis set and D3 for atom-pairwise dispersion correction

## 3.2 Computational procedure

This thesis aims to establish a reasonable relationship between the structure of *Sesquiterpene lactones* and its biological activity (against the methicillin resistance staphylococcus aureus). To achieve this goal, a general overview of the thesis procedure and methodology is outlined in Figure 3.2. Each step will be developed with more detail in the next section.

A general overview of the benchmark around *Sesquiterpene lactones* is necessary to generate a clear idea of the state-of-the-art of the topics involving these organic molecules. First, the general *in silico* studies around STL was essential to elucidate that STL has been studied for a long time, but essentially the study of their structure and the establishment of relations between biological activity and structure started in 1971 by Kupchan<sup>11</sup>. Afterward, a small review of different approaches including docking, DFT, and previous QSAR models, was carried out. Having studied the previous work done over STL, we were able to determine that not many works involve the antibacterial activity of STL neither *Staphylococcus aureus* resistant to methicillin and *in silico* studies. It was also evident that there is not much information available about the different biological activities of STL, this is due to that experimental procedures are needed to determine this kind of information. Experiments are time and resource consuming which cause a lack of information about STL biological activity. From this point, we determine that the study will involve STL that presents effective antibacterial activity against MRSA and that an effective *in silico* study was needed to generate a model able to predict the antibacterial activity.

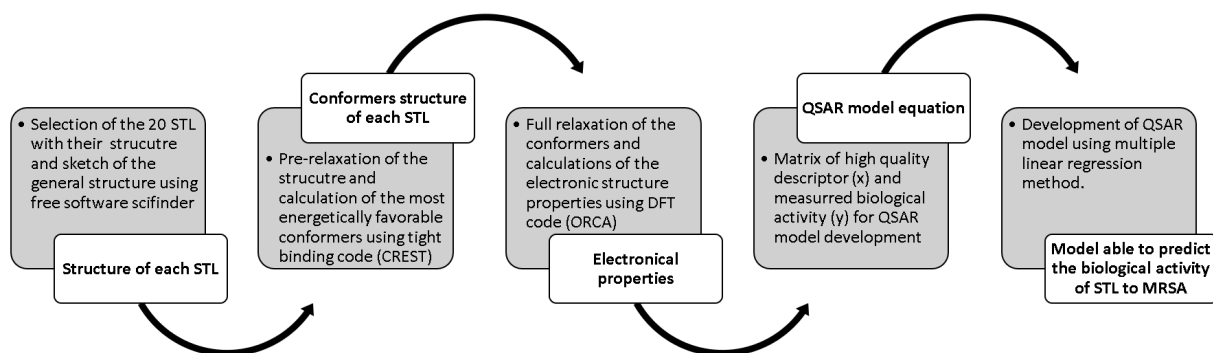


Figure 3.2: Schematic diagram of the process followed to obtain a QSAR model able to predict the the biological activity of *Sesquiterpene lactones* to MRSA

The next step needs an extensive bibliography review about reported STL with effective antibacterial activity against MRSA. Due to the nature of the QSAR model is important to have experimental information. After a full revision of the STL benchmark, 21 STL were selected based on the available structure reported with the respective antibacterial activity. It is of importance to have the 2D or 3D information about the molecule structure and their

biological activity since they are related, and using different activities from different STL will give false results. The structure has to be already determined, reported, and available in the literature.

With the structural information, we built a 3D model of each selected STL using Scifinder software (see Figure 3.3)<sup>89,90</sup>. Once the molecules were sketched, hydrogen atoms were used to fill the free bonds that the molecules have. Finally, a .mol file was generated with the structure of each STL. An important feature in the generation of these files and the construction of the molecule itself, is that it is important to conserve the same stereochemistry of the molecule. As mentioned before the form or distribution of the atoms in the space will determine the future calculations over the molecule.

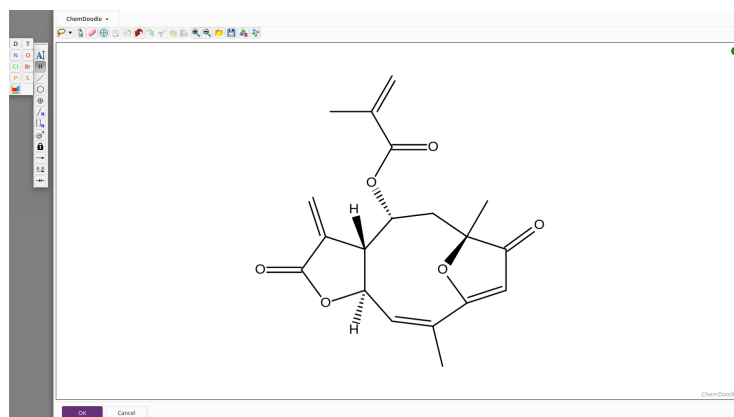


Figure 3.3: Scifinder user interface to graph molecules. The sketched molecules is Calaxin structure, it belongs to the *Sesquiterpene lactones* group used in this thesis.

Before starting with the calculations with Conformer-Rotamer Ensemble Sampling Tool (CREST) we transform the format of the .mol files to .pdb files. This is made because these types of files are the preferred ones when conformational calculations are performed. This action is made by CREST before the process begins. A small overview of this code was given before, but for a more detailed information see Ref.<sup>86</sup>.

All the possible conformers for each molecule were calculated using CREST. This code generates different files that contains different information, for our purposes we only use the file that contains the statistical information about the conformers (.out file) and the file that contains the structure of each conformer (.xyz file). For a single STL many different conformations can exist, but just the energetically stable are reported (which still are various possible structures in some cases). To illustrate the idea of how conformers can differ from the original structure, in Figure 3.4 a STL after a pre-relaxation with CREST and their respective conformer is showed.

We apply a criteria selection based on the statistical wight of each conformer, we need to achieve a 50% of the total



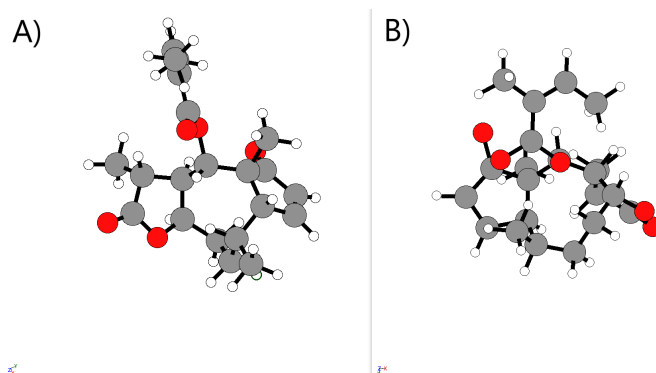


Figure 3.4: *Sesquiterpene lactones 6-O-angeloylplenolin*. A) Structure of the calculated conformer; B) Reported structure after a pre-relaxation using CREST. In these figures, the gray, red and white circles represent carbon, oxygen and hydrogen atoms, respectively.

possibilities of conformers with a few structures. This means that for a determined STL if one conformer structure has more than 50%, just that conformer will be selected for that particular molecule. If the sum of statistical weights of two or three conformers achieve 50%, then the two or three structures can be selected as the most representative conformations of that STL. The conformer(s) files for each molecule were saved in `.xyz` format.

For the calculation of electronic properties of each structure a small but effective code was built in ORCA based on specific requirements. First, it is important that the molecular structure of each conformer pass through a relaxation process, to then be used to calculate the properties. With these full relaxation ORCA can start the appropriated process to determine the properties.

This code implements *ab-initio* DFT with a level calculation that use hybrid functional B3LYP within the RIJCOSX approximation with def2-TZVPP basis set, def2/J auxiliary basis set and a combination of the atom-pairwise dispersion but with a damping function developed by Becke and Johnson. An overview of the main input file is sketched in Figure 3.5.

ORCA generates a file were all the information is contained, this is a `.out` file and a `.xyz` file with the structure of the relaxed conformer. In the `.out` file the information about all the iterations that ORCA made is detailed. From this file we need to extract the relevant information that we will use for the QSAR model. The selected information is the total energy ( $E_T$ ), the electronic energy ( $E_e$ ), the core-core repulsion energy ( $E_{c-c}$ ), the energy of the highest occupied molecular orbital (*HOMO*), the energy of the lowest unoccupied molecular orbital (*LUMO*), the *HOMO-LUMO* energy gap ( $E_g$ ). The band gap energy is the difference in energy between the lowest point of the conduction band and the highest point of the valence band; when temperature increase electrons can travel from the valance band to the conduction band creating electron-hole pairs which are directly linked to electrical conductivity. Other properties that were selected are the dipole moment ( $p$ ), the quadrupole moment ( $Q$ ) and the polarizability

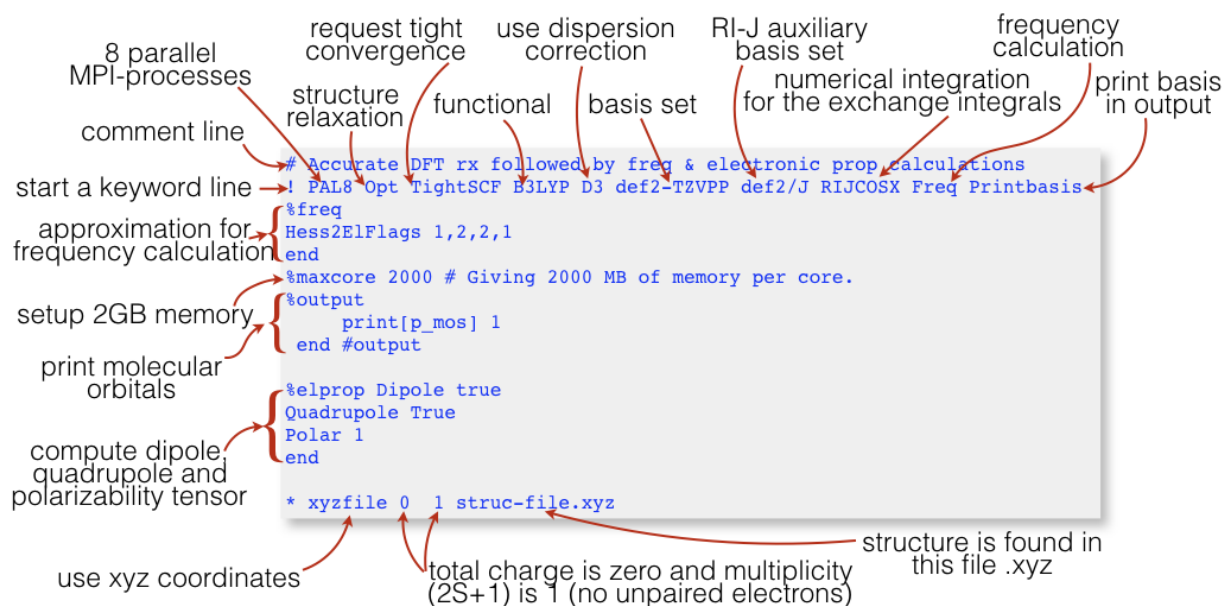


Figure 3.5: A production input file used in ORCA to compute the electronic properties of STL molecules in this work.

( $\alpha$ ). This information is related to most of the electronic characteristics of the structure. Also, another important feature of the structure is the radius of gyration which is defined as the root-mean-square average of the distance of all scattering elements from the center of mass of the molecule<sup>91</sup>.

### 3.3 Quantitative structure-activity relationship model development

The QSAR model used in this thesis is programmed in MATLAB<sup>92</sup> language and it is a multiple linear regression (MLR) fitting process prepared for logarithmic values of MRSA. The general equation for an MLR model is given by:

$$\mathbf{y} = b_0 + b_1x_1 + b_2x_2 + b_3x_3 + \dots + b_nx_n, \quad (3.1)$$

here the molecular descriptors are used as independent variables ( $x_i$ ) and the model parameters ( $b_1, b_2, b_3, \dots, b_n$ ). To have a good model is necessary to sort molecules by their log(MRSA) (endpoint) from lowest to highest, and split them into two sets: calibration or training set, and validation set. The calibration set is normally the 70~75% of the total set of molecules. The validation set will be the other 25~30% of the molecules. It is important to use molecules with known MRSA values for both sets. The model parameters ( $b_i$ ) are determined in the calibration phase, using the training set of molecules. Generally a QSAR model has to be internally and externally validated. For internal validation two statistical variables are be calculated, the  $R^2$  and  $Q_{CV}^2$  which gives information about

the goodness-of-calibration and robustness. External validation is developed based on the validation set and can be measured by  $Q_{Ext}^2$  which gives information about the predictability capacity. For our purposes the training and validation set is divided into 1:2 (see table 3.1).

### 3.3.1 Endpoint

The endpoint was measured based on MRSA concentration expressed in  $\mu\text{g/mL}$ . The final model was developed based on logarithmic value expressed as the  $\log(\text{MRSA})$ .

Table 3.1: Structure label, value of the descriptors electronic energy  $E_e$  and HOMO energy, observed  $\log(\text{MRSA})$ , set (validation V and calibration C).

Structure	$E_e$	HOMO	$\log(\text{MRSA})$	Set
7C1	-2039.5609	-0.2244	0.2900	C
7C2	-2038.5950	-0.2253	0.2900	C
6C1	-2088.2698	-0.2488	1.1931	V
1C1	-2226.8852	-0.2503	1.3979	C
2C1	-2141.5073	-0.2440	1.6902	C
21C1	-3998.0506	-0.2580	1.6990	V
9C1	-2479.0666	-0.2587	2.0000	C
19C1	-4455.5364	-0.2562	2.3010	C
3C1	-2178.0091	-0.2429	2.3909	V
4C1	-1992.0913	-0.2356	2.3979	C
17C1	-3723.6660	-0.2781	2.3979	C
17C2	-3732.3715	-0.2788	2.3979	V
5C1	-3473.8755	-0.2514	2.4771	C
14C1	-3414.9182	-0.2572	2.4771	C

10C1	-3582.9435	-0.2559	2.6021	V
10C2	-3570.3207	-0.2550	2.6021	C
18C1	-4034.5419	-0.2530	2.6021	C
20C1	-4071.2748	-0.2584	2.6021	V
20C2	-4066.3747	-0.2611	2.6021	C
16C1	-3884.1986	-0.2632	2.6990	C
16C2	-3881.4599	-0.2633	2.6990	V
15C1	-3600.2516	-0.2564	2.7782	C
8C1	-2820.0698	-0.2644	2.9542	C
12C1	-2413.8204	-0.2540	3.1761	V
13C1	-2406.9504	-0.2566	3.1761	C
13C2	-2395.5465	-0.2564	3.1761	C
11C1	-4824.9528	-0.2668	5.0000	V
11C2	-4824.4102	-0.2686	5.0000	C

---

### 3.3.2 Descriptors

The different numerical vectors obtained from ORCA and the radius of gyration are now called descriptors, and are going to be used in QSAR model. Due to the amount of data obtained, it is important to select only 1 or 2 descriptors. Since it is not possible to determine *a priori* which combination of descriptors will be the most accurate for the calibration set we develop several combinations of them and generate a QSAR model equation for each combination. All the different models use the antibacterial activity of each molecules and the selected descriptors. Finally, the selected model will depend on the relationship between descriptors and toxicity.

### 3.3.3 Prediction

Once all  $b_i$  parameters are determined, the model is able to predict the endpoint of new molecules based only on descriptors. In other words QSAR model is then used calculate the  $\log(\text{MRSA})$  and therefore the MRSA concentration of each STL which were not used in any set of the model, these are STL with no experimental value of MRSA. Modified structures were also included into the prediction set, these molecules are protonated molecules obtained from previous STL structures.

For the addition of two H atoms to the structure, three structures were selected from the calibration and validation set (structure 2, 6 and 21). The addition process is implemented to see if some significant changes are showed in the biological activity when two hydrogen atoms are added to the original structure of STL. With the relaxed structure of the conformers obtained from ORCA we can start the addition process using CREST. This process is implemented automatically by the software, and generates a .xyz file with the manipulated structure. The files obtained from the most stable manipulated molecules follows the same procedure as the original STL. Once the electronic properties are calculated, the HOMO energy and the electronic energy descriptors are used to predict their possible biological activity.

### 3.3.4 Molecular orbitals and Structure activity relationship

Using the output files obtained from ORCA, a 3D graphic of the HOMO of each STL is visualized using the Gabedit<sup>93</sup> software. Here all the information about the orbital (filled and empty) is used to create an approximated image of the orbitals and where are they located is generated. Finally, this information is saved in a .png file.



## Chapter 4

# Results & Discussion

### 4.1 Structures modeling and electronic properties calculation

A set of 21 *Sesquiterpene Lactones* were selected based on the antibacterial activity (MRSA) and the reported structure (see Table 4.1). STL can be classified into the groups mentioned in chapter 2 that differ in the skeleton or basic structure. With these molecules and based on the graphical representation of their structure, a 3D model of each one was constructed as mentioned in chapter 2. The 2D representation of these molecules can be found in Table 4.5.

Table 4.1: Selected *Sesquiterpene lactones* with their MRSA concentration in  $\mu\text{g/mL}$  and the  $\log(\text{MRSA})$  value

Number	Name	MRSA	$\log(\text{MRSA})$
1	Parthenolide <sup>94</sup>	25.00	1.39
2	Dehydroleucodine <sup>95</sup>	49.00	1.69
3	Leucodine <sup>95</sup>	246.00	2.39
4	Costunolide <sup>96</sup>	100.00-400.00	2.39
5	Arnicolide C <sup>97</sup>	300.00	2.47
6	Xanthatin <sup>98</sup>	7.80-15.60	1.19

*Continued on next page*

Table 4.1 – Continued from previous page

Number	Name	MRSA	log(MRSA)
7	Ketolactone <sup>97,99</sup>	1.95	0.29
8	Artemisinin <sup>94,100</sup>	900.00	2.95
9	Parthenin <sup>94,101</sup>	100.00	2.00
10	Calaxin <sup>102</sup>	400.00	2.60
11	Acanthospermal B <sup>103</sup>	100x10 <sup>3</sup>	5.00
12	Armexifolin <sup>104</sup>	1500.00	3.17
13	Armexifolin (alternative) <sup>104</sup>	1500.00	3.17
14	6-O-methylacrylylplenolin <sup>94</sup>	300.00	2.47
15	6-O-angeloylplenolin <sup>94</sup>	600.00	2.77
16	Vernolide <sup>105</sup>	500.00	2.69
17	Vernodaline <sup>105</sup>	250.00	2.39
18	8B-(epoxyangeloyloxy)-14-hydroxy-tithifolin <sup>106</sup>	400.00	2.60
19	8B-(epoxyangeloyloxy)-14-acetoxy-eupatolide <sup>106</sup>	200.00	2.30
20	Niveusin-C <sup>106</sup>	400.00	2.60
21	Budlein-A <sup>106</sup>	50.00	1.69

Considering Table 4.1, CREST calculations were carried out for each structure. These calculations were made with the default parameters of CREST, and the specific structure for each of the possible conformers were calculated. A total of 28 structures were finally obtained and selected as conformers for the 21 original *Sesquiterpene Lactones* (STL) (the first number in the labeling refers to each STL, "C" is a shortening for Conformer, and the second number refers to the first or second conformer). A different file is generated for each structure by CREST, then these files



were used to compute the electronic properties of the conformers using ORCA. After a full relaxation, the electronic structure was computed and the total energy, the electronic energy, the core–core repulsion energy, the energy of the highest occupied molecular orbital (HOMO), the energy of the lowest unoccupied molecular orbital (LUMO), the *HOMO-LUMO* energy gap or band gap energy ( $E_g$ ), the dipole moment ( $p$ ), the quadrupole moment ( $Q$ ) and the polarizability  $\alpha$ . All calculated values are tabulated in Table 4.2. Finally, since ORCA generates a new .xyz file for each of the fully-relaxed conformers, we also calculated the radius of gyration ( $R_g$ ) and the total mass of each conformer.

Table 4.2: DFT-B3LYP computed properties for selected conformers of studied molecules: total energy  $E_T$ , dipole moment  $p$ , quadrupole moment  $Q$ , polarizability  $\alpha$ , electronic energy  $E_e$ , core–core repulsion energy  $E_{c-c}$ , HOMO, LUMO, and HOMO-LUMO energy gap  $E_g$ , total mass  $M_t$ , and radius of Gyration  $R_g$ . In this table,  $p$ ,  $Q$  and  $\alpha$  are in a.u., energies in Eh, total mass in u and  $R_g$  in Å.

Mol	$E_T$	$p$	$Q$	$\alpha$	$E_e$	$E_{c-c}$	HOMO	LUMO	$E_g$	$M_t$	$R_g$
1C1	-809.149	2.641	-84.428	176.933	-2226.885	1417.818	-0.250	-0.057	0.194	248.322	8.502
2C1	-806.784	1.545	-83.991	180.642	-2141.507	1334.797	-0.244	-0.069	0.175	244.290	8.860
3C1	-808.014	1.473	-85.507	179.631	-2178.009	1370.072	-0.243	-0.066	0.176	246.306	8.948
4C1	-733.942	2.231	-80.747	178.478	-1992.091	1258.230	-0.236	-0.052	0.183	232.323	8.590
5C1	-1115.691	1.717	-111.641	222.651	-3473.875	2358.299	-0.251	-0.061	0.191	334.412	10.156
6C1	-807.976	1.013	-88.878	194.548	-2088.270	1280.367	-0.249	-0.080	0.168	246.306	13.333
7C1	-804.394	1.759	-82.466	216.220	-2039.561	1235.233	-0.224	-0.111	0.113	240.258	9.790
7C2	-804.393	1.739	-82.523	216.273	-2038.595	1234.268	-0.225	-0.111	0.114	240.258	9.808
8C1	-960.795	2.271	-91.429	181.086	-2820.070	1859.368	-0.264	-0.002	0.263	282.336	7.771
9C1	-883.203	1.208	-86.377	174.527	-2479.067	1595.951	-0.259	-0.069	0.189	262.305	7.342
10C1	-1187.252	1.830	-111.393	230.787	-3582.943	2395.796	-0.256	-0.066	0.190	344.363	10.140
10C2	-1187.252	1.683	-111.932	229.132	-3570.321	2383.172	-0.255	-0.065	0.190	344.363	10.393
11C1	-1456.847	2.445	-137.867	273.109	-4824.953	3368.243	-0.267	-0.089	0.178	420.458	11.762
11C2	-1456.846	2.497	-137.775	272.713	-4824.410	3367.701	-0.269	-0.089	0.180	420.458	11.713
12C1	-883.217	2.415	-88.628	182.558	-2413.820	1530.686	-0.254	-0.066	0.188	262.305	9.145
13C1	-883.219	0.885	-90.904	182.392	-2406.950	1523.813	-0.257	-0.069	0.188	262.305	9.138
13C2	-883.217	2.150	-90.647	183.797	-2395.546	1512.411	-0.256	-0.067	0.189	262.305	9.476

---

14C1	-1114.473	2.177	-109.908	223.212	-3414.918	2300.557	-0.257	-0.071	0.186	332.396	10.328
15C1	-1153.780	2.353	-114.455	238.680	-3600.252	2446.589	-0.256	-0.071	0.185	346.423	11.331
16C1	-1263.628	1.825	-114.863	231.601	-3884.199	2620.685	-0.263	-0.059	0.204	362.378	11.187
16C2	-1263.629	1.736	-114.952	230.542	-3881.460	2617.945	-0.263	-0.061	0.202	362.378	11.256
17C1	-1262.461	2.475	-123.516	234.521	-3723.666	2461.311	-0.278	-0.077	0.201	360.362	13.517
17C2	-1262.461	2.506	-122.789	234.105	-3732.372	2470.017	-0.279	-0.077	0.201	360.362	13.338
18C1	-1304.140	2.117	-118.398	249.511	-4034.542	2730.514	-0.253	-0.048	0.204	378.421	13.367
19C1	-1381.594	2.130	-126.791	273.245	-4455.536	3074.071	-0.256	-0.052	0.204	404.459	13.483
20C1	-1304.208	0.874	-121.333	252.143	-4071.275	2767.185	-0.258	-0.054	0.204	378.421	13.446
20C2	-1304.208	1.196	-121.829	251.365	-4066.375	2762.284	-0.261	-0.053	0.208	378.421	13.321
21C1	-1301.779	1.423	-120.302	250.534	-3998.051	2696.384	-0.258	-0.066	0.192	374.389	12.036

---

To the best of our knowledge, not many studies combine DFT with Tight Binding, here we apply these methods that ensure the calculated descriptors are of high quality and correctly describe the characteristics of the structure.

## 4.2 Quantitative structure-activity relationship model

The matrix of calculated descriptors and measured endpoints was then used to start the construction of the QSAR models. At this point, the descriptors were combined in all the possible ways (without repeating the pair), and used alone. Just the most significant models were selected. In table 4.3 the combinations and the resulting correlation coefficient of each developed model is presented.

The selected models were developed based on MLR method using one and two descriptors. These descriptors represents the electronic energy  $E_e$  and the energy of the highest occupied molecular orbital (HOMO) both in Hartree units. The equations that describes each model are:

Model 2

$$\log(MRSA) = -9.9184 - 0.0003 E_e - 45.3675 \text{ HOMO.} \quad (4.1)$$

Model 8

$$\log(MRSA) = -11.9959 - 56.7318 \text{ HOMO.} \quad (4.2)$$

For internal and external validation different statistical variables were calculated. These variables gives information about goodness-of-calibration ( $R^2$ ), robustness ( $Q_{CV}^2$ ), and predictability ( $Q_{EP}^2$ ) of the model. The obtained

Table 4.3: Different combination of descriptors with their goodness-of-calibration  $R^2$ 

Model	Descriptor 1	Descriptor 2	$R^2$
1	Core-core repulsion energy [Eh]	HOMO [Eh]	0.5787
2	Electronic energy [Eh]	HOMO [Eh]	0.5732
3	Quadrupole moment [a.u.]	HOMO [Eh]	0.5686
4	HOMO [Eh]	Total mass	0.5644
5	Polarizability [a.u.]	HOMO [Eh]	0.552
6	dipole moment [a.u.]	HOMO [Eh]	0.5457
7	HOMO [Eh]	Radius of Giration	0.5456
8	HOMO [Eh]		0.5441
9	Quadrupole moment [a.u.]	HOMO-LUMO energy gap [Eh]	0.5285
10	Core-core repulsion energy [Eh]	HOMO-LUMO energy gap [Eh]	0.5171

results for both models are presented in Table 4.4 with their respective errors.

The computed values displayed in Table 4.4 are not good from a scientific point of view because they are not close enough to an acceptable value of prediction. However, even though the obtained values for  $R^2$  are far from the ideal value of 1, not all information is enclosed in the correlation coefficient, in fact most information can be obtained analysing the relationships that can exist between the used descriptors and the biological activity. This is because we do not only look for a good  $R^2$ , which gives information about how far our point are from the function, but also the physical meaning behind the equations and variables. For this reason Models 2 and 8 were selected. In Figure 4.1, the graphics obtained using the selected models are displayed where the predicted and experimental values of  $\log(MRSA)$  are compared, in this figure, the line indicates exact prediction.

The descriptors used here were the Electronic energy and the HOMO energy, both of them are characteristic values of a specific conformation and are strictly related to the electronic structure. It is possible that a toxic mechanism that involves the electric structure takes place, and influence the biological activity.

It is clear that biological activity is tied to structural components such as functional groups, double bonds, or other type of structures<sup>3,71</sup>. But what is not straightforward to notice is that the electronic properties of a molecule are also involved in the toxic mechanism and biological activity<sup>71,107</sup>. In both models a common descriptor is implemented,

Table 4.4: Statistical variables obtained from Equations 4.1 and 4.2

Model	$R^2$	$Q_{CV}^2$	$Q_{Ext}^2$	RMSEC	RMSECV	RMSEP
2	0.5732	0.3046	0.1992	0.6570	0.8386	0.9132
8	0.5441	0.3626	0.1231	0.6790	0.8028	0.9577

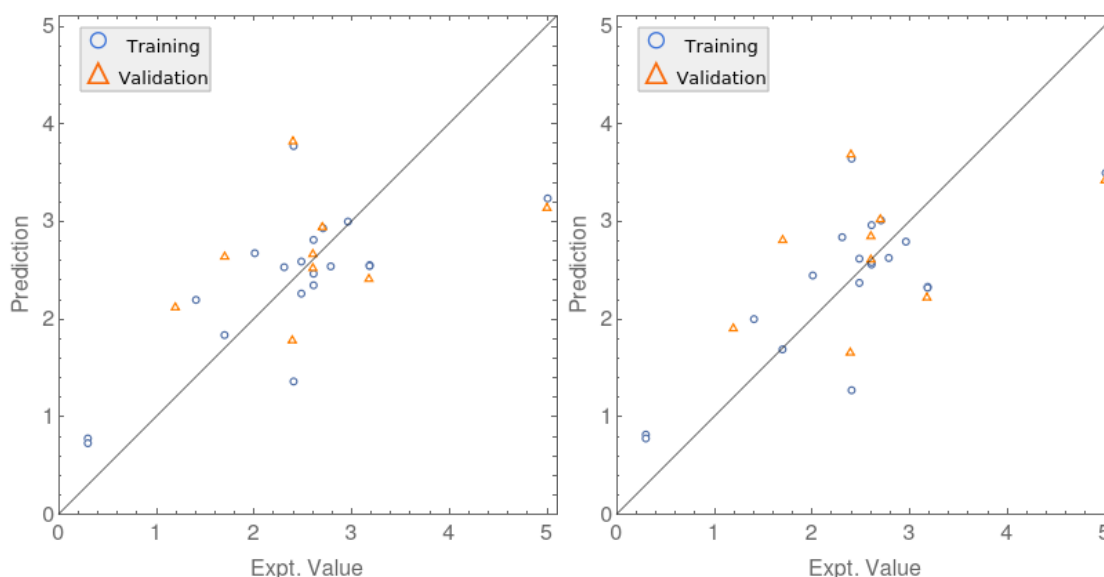


Figure 4.1: QSAR models graphics displaying the predicted value of log(MRSA) for calibration and validation test

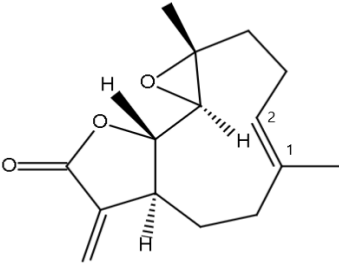
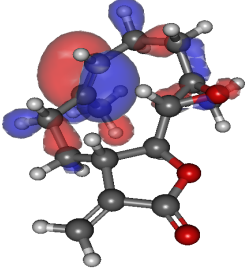
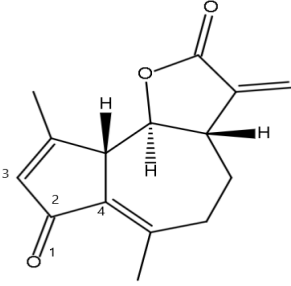
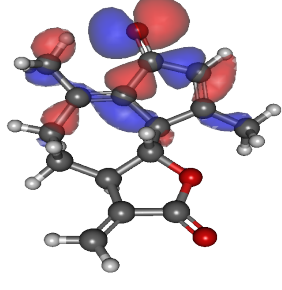
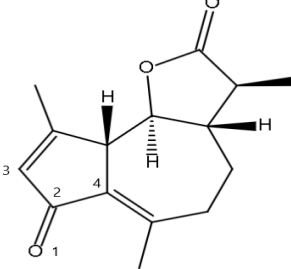
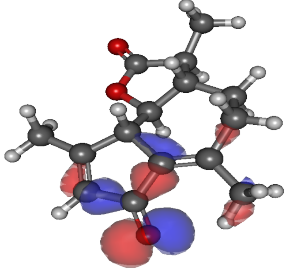
the Highest Occupied Molecular Orbital energy. This descriptor usually appears in QSAR studies<sup>108-112</sup>, and it is related to different biological activities. What these results suggest, is the existence of a relationship between HOMO energy and the biological activity, specifically antibacterial activity against MRSA, through some toxicity mechanism of actions. Since HOMO energy is related to how easily an electron can be detached from the last orbital, the greater the HOMO energy value the easier is to detach one electron from the molecule, this electron could migrate inside the bacteria producing free radicals that can damage and eventually contribute to kill the bacteria, leading to an increase in antibacterial activity.

Several studies indicate that the relationship between HOMO, band gap energy and cytotoxicity exist and it depends also to where the molecule is going to be attached<sup>113</sup>. The electronegativity of a molecule and even the atoms can affect the potential biological activity that STL can have<sup>1</sup>. Taking into account that not all molecules follows the same experimental protocol, and their antibacterial activity is reported by different research groups in different times, it is interesting that just with the HOMO energy more than half of MRSA activity of molecules could

potentially be predicted. In fact, in Figure 4.1 most molecules are concentrated near the center, except those whose values differ greatly from the mean. These results from QSAR model suggest the important role that electronic structure plays in molecules properties.

It is mentioned that there exist different types of descriptors, HOMO energy can be considered as a basic descriptor. Our preliminary QSAR model can be enhance by using more complex descriptors such as topological ones, standardizing the used data (obtained following the same protocol), and selecting specific family of structures STL. Many improvements can be done by refining the base data.

Table 4.5: Summary of results displaying the experimental value of  $\log(MRSA)$  within parenthesis and the predicted values according to model 2 (M2) and model 8 (M8), respectively. It is also displayed the 2D structure and the the B3LYP computed HOMO.

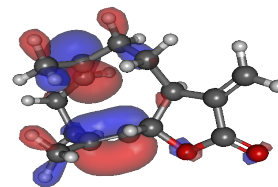
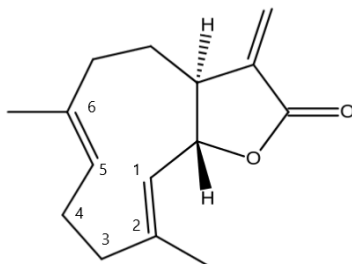
Name	$\log(MRSA)$	Structure	3D HOMO
Parthenolide <sup>94</sup>	(1.398) M2: 2.008 M8: 2.205		
Dehydroleucodine <sup>95</sup>	(1.690) M2: 1.698 M8: 1.845		
Leucodine <sup>95</sup>	(2.391) M2: 1.658 M8: 1.783		

Costunolide<sup>96</sup>

(2.398)

M2: 1.280

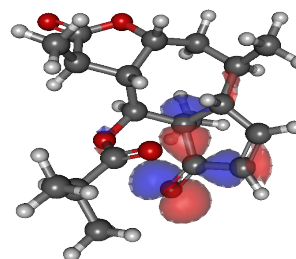
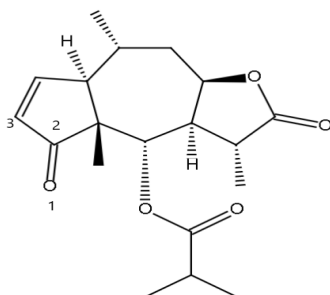
M8: 1.370

Arnicolide C<sup>97</sup>

(2.477)

M2: 2.379

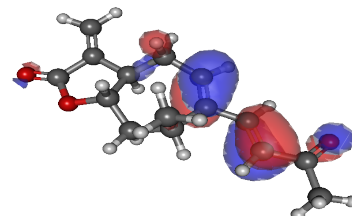
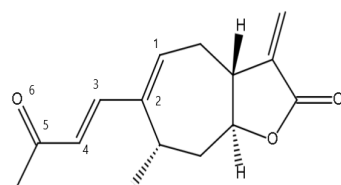
M8: 2.269

Xanthatin<sup>98</sup>

(1.193)

M2: 1.905

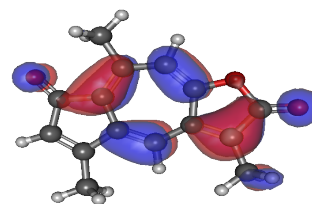
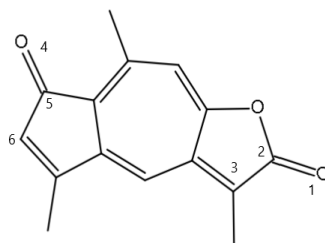
M8: 2.120

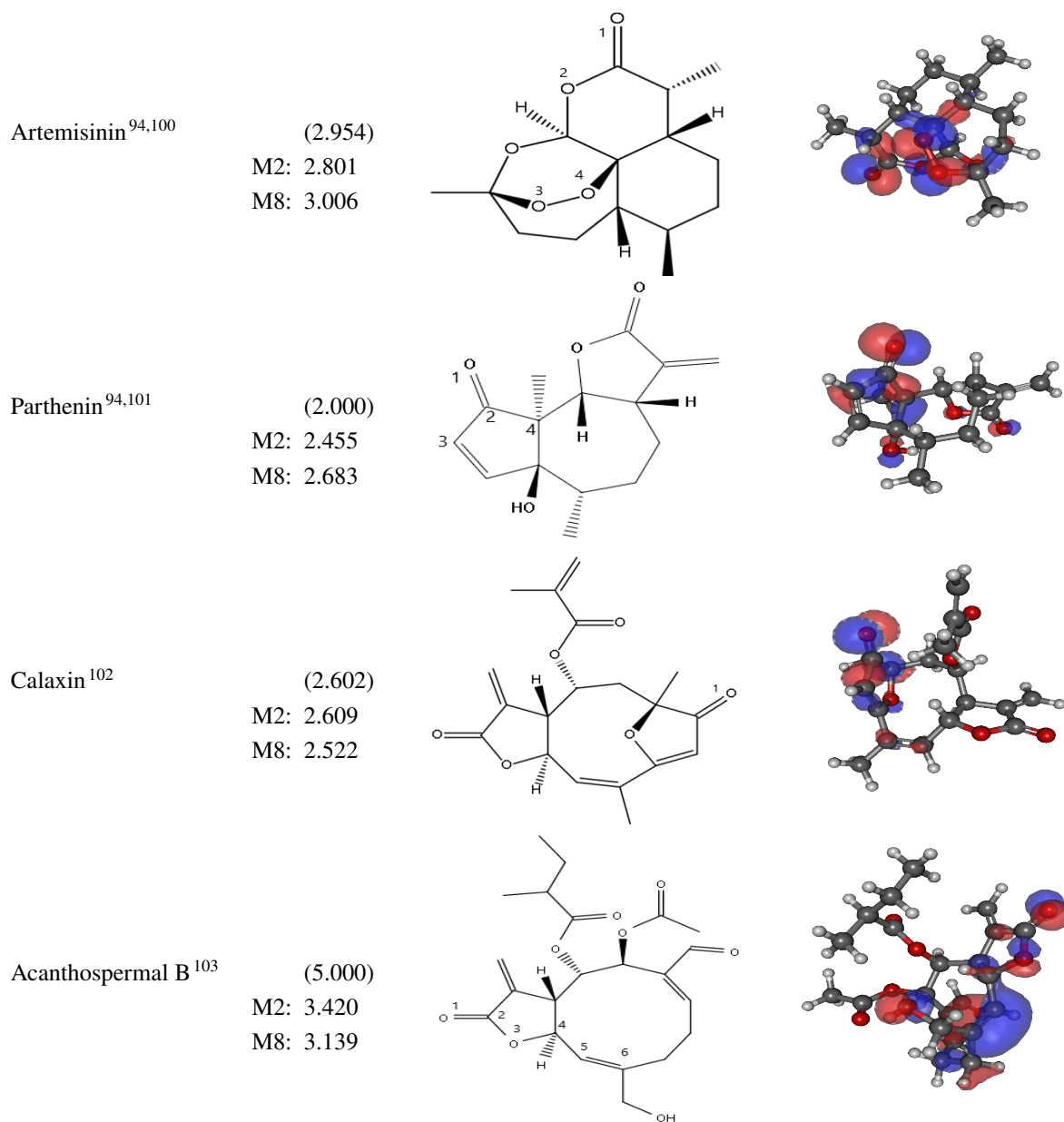
Ketolactone<sup>97,99</sup>

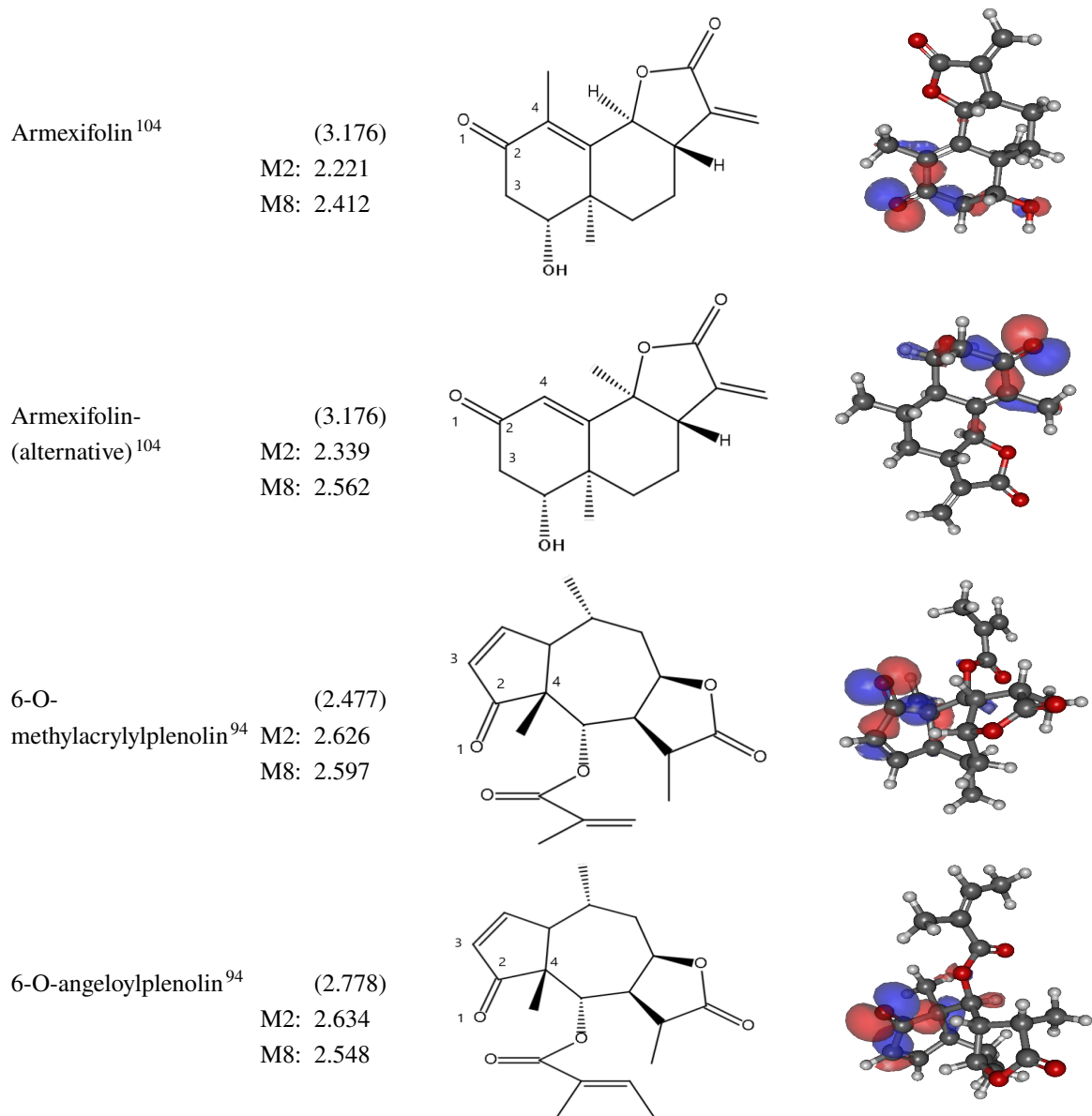
(0.290)

M2: 0.787

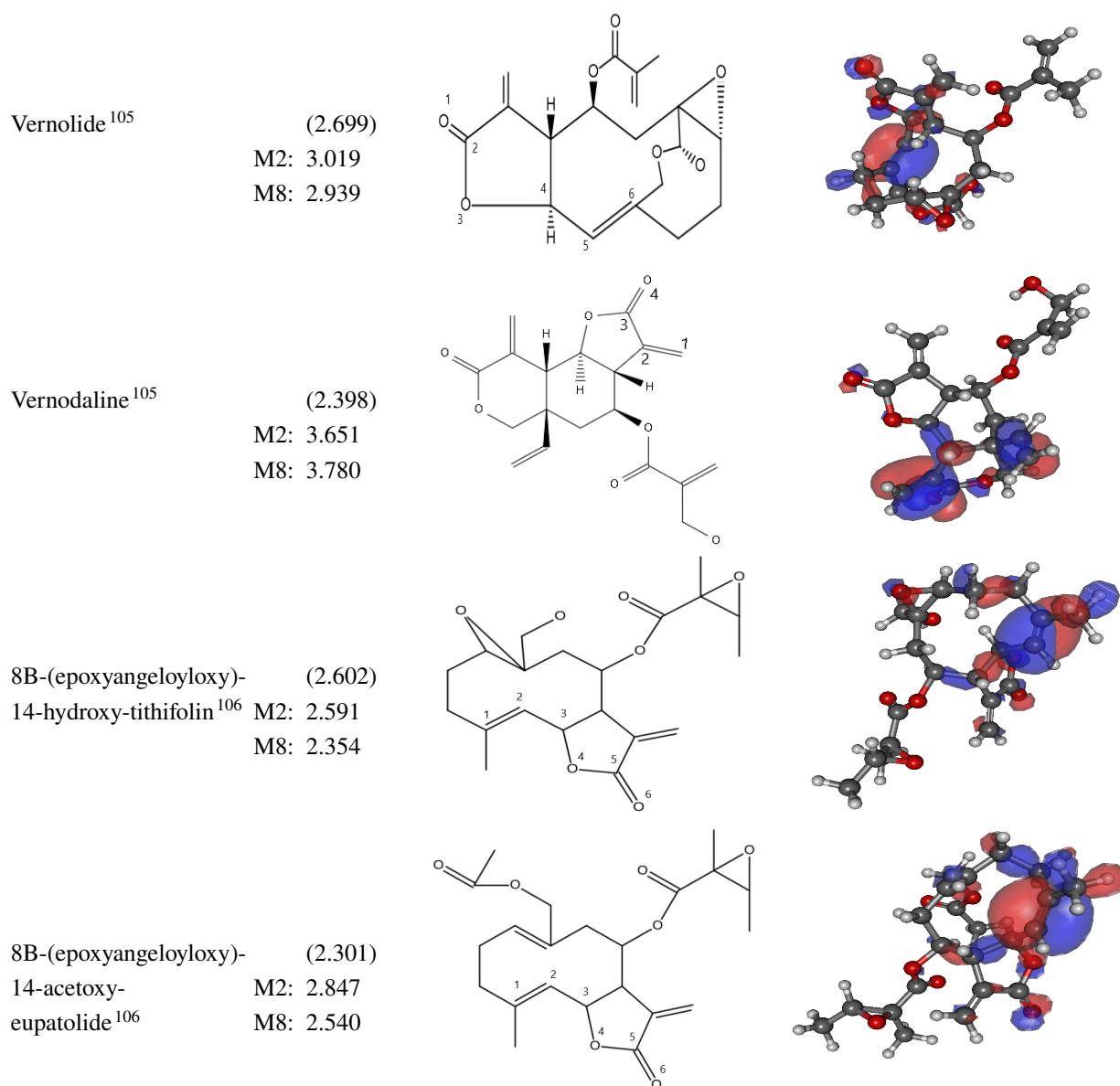
M8: 0.737

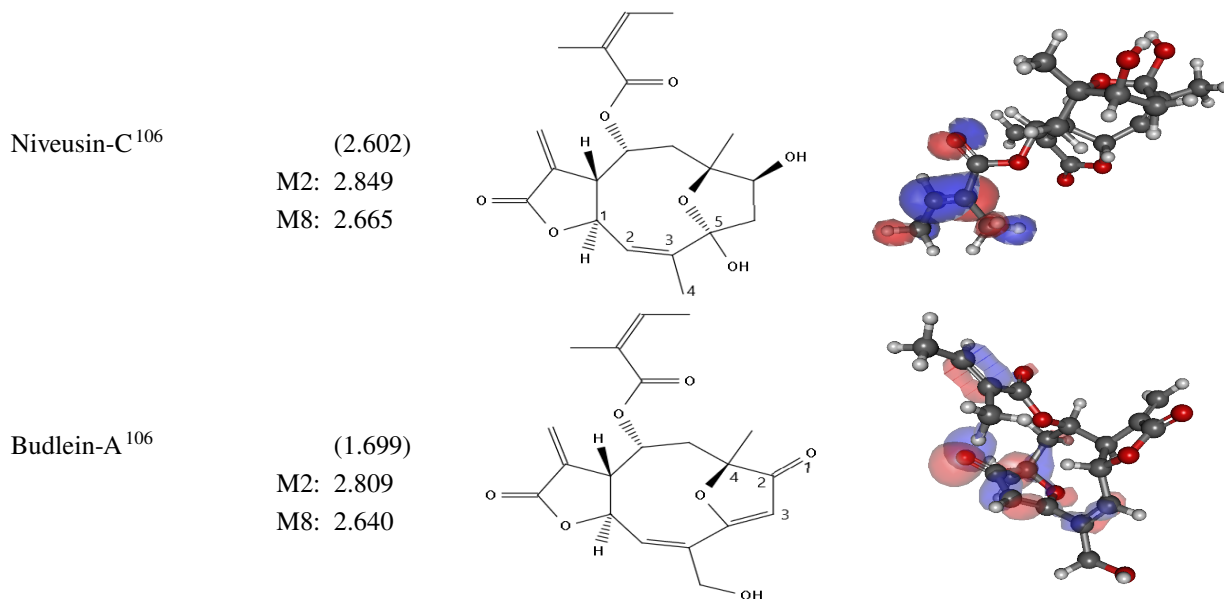












Apart from the relationship between descriptors (HOMO and electronic energy) and biological activity, QSAR model can establish relations with the specific parts of the structure. For this, the molecular orbitals of each structure were sketched using Gabedit<sup>93</sup> software generating a 3D image. These images give information about where the electrons are more likely to be placed, in other words where the electron density is concentrated. These places where more electrons are located are the active places where the toxicity mechanism can take place. This idea implies that if the structure is modified in such a way that the HOMO energy (molecular orbitals) change, the biological activity can be enhanced or reduced. Since the biological activity is related to structural components and by identifying the common places where the HOMO energy is mostly concentrated in each structure, active places for biological activity could be determined.

The plot of HOMO for each of the 21 structures are displayed in Table (4.5). Molecules were classified according to their skeletons (see Figure 1.1), as result 8 molecules (Parthenolide, Calaxin, Acanthospermal B, Vernolide, 8B-(epoxyangeloyloxy)-14-hydroxy-tithifolin, 8B-(epoxyangeloyloxy)-14-acetoxy-eupatolide, Niveusin C and Budlein A ) belong to Germacranolides group (G1), Costunolide belongs to the Heliangolides group (G2), 3 molecules (Dehydroleucodine, Leucodine, Ketolactone) are part of the Guaianolides group (G3), 4 STL (Arnicolide C, Parthenin, 6-O-methylacrylylplenolin and 6-O-angeloylplenolin) belong to the Pseudoguaianolides group (G4), and 3 molecules (Vernodaline and both Armexifolin) are part of Eudesmanolides group (G5). Xanthantin and Artemisinin were not classified into any of these groups because they have different skeletons. It is clear that among all molecules, there exist two predominant parts on which molecular orbitals are concentrated, double bond to an oxygen atom (p-orbitals) and double bond between carbon atoms ( $\pi$ -orbitals) inside or outside the principal ring of STL.

In G1 Parthenolide, which has a high antibacterial activity mainly has the HOMO located at a double bond between carbon 1 and 2, Niveusin-C and Budlein-A are interesting because both structures are similar but Budlein-A has much higher antibacterial activity and has MO concentrated in the p-orbitals in oxygen 1, but Niveusin-C has the HOMO concentrated in a  $\pi$ -orbital (carbon 2 and 3) inside the principal ring. The other 5 molecules have also have a  $\pi$ -orbital, but these molecules presents less biological activity that could be attributed to the "tail" that is part of their structure. In G2, Costunolide has a medium antibacterial activity and the HOMO distribution is concentrated at two  $\pi$ -orbital (carbon 1-2 and 5-6) in the principal rings. From G3 group Dehydroleucodine and Leucodine have almost the same HOMO distribution in p-orbitals (oxygen 1), but Ketolactone which is a rigid molecule has several active sites in it ( $\pi$  and p-orbitals) suggesting that for this reason this molecule is the most active one. G4 molecules have similar HONO distribution around p-orbitals (O=R), but only Parthenin has no "tail", which in fact seems to affect the antibacterial activity of the other molecules in this group. For G5, Armexifolin and its alternative structure have a concentration of HOMO in oxygen 1 with double bond to a 6 carbon ring but their activity is way to low. Vernodaline presents better biological activity that could be attributed to the MO distribution which is composed by a contribution of  $\pi$ -orbital in carbon 1-2 and p-orbitals in oxygen 4, despite it has a tail. Artemisinin has a unique HOMO distribution, they are concentrated in a single bond between carbon 3 and 4, and also at a p-orbital in oxygen 1 which has a double bond. For Xanthantin the MO distribution is mainly concentrated in two  $\pi$ -orbitals (C=C) and a double bond with oxygen 6 (p-orbital), this molecules is the second most active and it could be due to the presence of various active points in the molecule.

The presence of these particular features (O=R and C=C) in the structure seems to affect in certain way the possible antibacterial activity that STL can have. Also the quantity and position of the bonds could be a determinant when performing antibacterial test.

#### 4.2.1 Minimum inhibitory concentration prediction

Different structures were used to test the model apart from the first 21 STL, such as Arglablin<sup>114</sup>, Dehydrocostus lactone<sup>94</sup>, and Helenalin<sup>94</sup>, but for these molecules there is not a MRSA value reported in the review literature. The protonation of the molecules (Dehydroleucodine, Xanthatin, and Budlein-A) were performed successfully giving as a result three structures that were added to the training set. For this prediction set molecules, MRSA activity has not been reported yet.

The electronic properties of these training set molecules were calculated, and equations 4.1 and 4.2 were used to calculate their possible  $\log(MRSA)$ . The results of the calculations are shown in table 4.6 with their respective HOMO and band gap energy.

Table 4.6: Training set molecules used to predicted their MRSA concentration value using equations 4.1 and 4.2. Here DWN and UP refers to the position of extra Hydrogen atom (down and up respectively), and the numbers are used to identify which conformer it is.

Name	$E_e$	HOMO	Model 2	Model 8
Arglabin	-2224.9962	-0.2467	1.8420	1.9978
Dehydrocostus L	-1845.1131	-0.2490	1.8485	2.1277
Helenalin	-2460.7882	-0.2564	2.3448	2.5510
2C1-DWN1	-2176.4804	-0.2464	1.8169	1.9819
6C1-DWN1	-2126.7680	-0.2310	1.1040	1.1064
21C1-UP1	-4056.3024	-0.2415	2.0765	1.7045

## Chapter 5

# Conclusions & Outlook

The relationship between molecular structure and biological activity of *Sesquiterpene lactones* have been studied over the pass 50 years giving positive results for computational models. We are conscious that this is not the first *in silico* studied involving STL and QSAR methods, but so far with the reviewed literature we can say that we are the first that combines semi-empirical tight binding and DFT to obtain appropriate molecular structures (conformers) and electronic properties to then be used in QSAR models. With this type of combinations we can assure that the calculated structures and results are accurate.

We were able to calculate and select the specific conformers of each STL. The selection of conformers were carried out based on statistical and energetically favorable criteria. With this, the calculated properties are specific for each structure which is stable within a pertinent range of temperature<sup>86</sup>. We were able to built a code that implements different functions of ORCA. Knowing that these functions includes corrections for van der Waal forces the results are precise and correctly represents the electronic properties of the real molecules. The results tabulated in Table 4.2 were used as descriptors to develop the multiple linear regression models which establish a relation between the electronic structure properties and the biological activity.

Several QSAR models were developed, giving equations 4.1 and 4.2 as best models which presents correlation coefficient 0,57 and 0.54. These models were used to estimated the possible antibacterial activity of STL against MRSA. Even though the  $R^2$  value for this first model is not close enough to 1, it allows us to first determine that the structures and properties are well calculated. This is because there exist a relationship between the descriptors (HOMO energy value and Electronic energy) and the biological activity of molecules, the relationship between descriptors and cytotoxicity has been reported in previous works<sup>68</sup>. These results suggest a relation between Highest Occupied Molecular Orbital energy and the mechanism of toxic actions: the greater the HOMO energy the easier to ionize the molecule by detaching one electron from the molecule that could migrate inside the bacteria producing free radicals that can damage and eventually contribute to kill the bacteria. Also, the presence and the quantity of structural features such as double bonds with oxygen atoms and double bonds between carbon atoms in the main skeleton of STL apparently intervene in the biological activity performance of molecules.

The development of more effective descriptors to better describe the toxicity of *Sesquiterpene lactones* is needed;

we are aware that more extensive exploration of parameters and molecules are needed to find the best correlation with the observed toxicity, this work is underway.

We can recommend that once the specific group of molecules to be studied has been defined, *insilico* studies can be implemented in an early stage before the experimental synthesis of each compound. Knowing in advance the possible performance and effectiveness of the components can save time and resources in not studying those compounds that computationally do not present the specific biological activity. Later the experimental results would be compared with the *insilico* ones to corroborate the information and improve the quality of the models.

# Bibliography

- [1] Scotti, M. T.; Fernandes, M. B.; Ferreira, M. J.; Emerenciano, V. P. Quantitative structure–activity relationship of sesquiterpene lactones with cytotoxic activity. *Bioorganic & Medicinal Chemistry* **2007**, *15*, 2927–2934.
- [2] Picman, A. K. Biological activities of sesquiterpene lactones. *Biochemical systematics and Ecology* **1986**, *14*, 255–281.
- [3] Schmidt, T. J. *Studies in natural products chemistry*; Elsevier, 2006; Vol. 33; pp 309–392.
- [4] Rodriguez, E.; Towers, G.; Mitchell, J. Biological activities of sesquiterpene lactones. *Phytochemistry* **1976**, *15*, 1573–1580.
- [5] Morris, G. M.; Lim-Wilby, M. *Molecular modeling of proteins*; Springer, 2008; pp 365–382.
- [6] da Silva, A. C. B.; da Silva, D. R.; de Macêdo Ferreira, S. A.; Agripino, G. G.; Albuquerque, A. R.; do Rêgo, T. G. In Silico Approach for the Identification of Potential Targets and Specific Antimicrobials for *Streptococcus mutans*. *Advances in Bioscience and Biotechnology* **2014**, *2014*.
- [7] Aliyu, A. B.; Koorbanally, N. A.; Moodley, B.; Singh, P.; Chenia, H. Y. Quorum sensing inhibitory potential and molecular docking studies of sesquiterpene lactones from *Vernonia blumeoides*. *Phytochemistry* **2016**, *126*, 23–33.
- [8] Berman, H. M.; Westbrook, J.; Feng, Z.; Gilliland, G.; Bhat, T. N.; Weissig, H.; Shindyalov, I. N.; Bourne, P. E. The Protein Data Bank. *Nucleic Acids Research* **2000**, *28*, 235–242.
- [9] Hegazy, M.-E. F.; Ibrahim, A. Y.; Mohamed, T. A.; Shahat, A. A.; El Halawany, A. M.; Abdel-Azim, N. S.; Alsaid, M. S.; Paré, P. W. Sesquiterpene lactones from *Cynara cornigera*: acetyl cholinesterase inhibition and in silico ligand docking. *Planta medica* **2016**, *82*, 138–146.
- [10] Luo, H.-J.; Wang, J.-Z.; Deng, W.-Q.; Zou, K. DFT Calculations and Docking Study on Sesquiterpene Lactones: Inhibition of Aromatase. *Procedia Environmental Sciences* **2011**, *8*, 446–450.
- [11] Kupchan, S. M.; Eakin, M.; Thomas, A. Tumor inhibitors. 69. Structure-cytotoxicity relations among the sesquiterpene lactones. *Journal of Medicinal Chemistry* **1971**, *14*, 1147–1152.

- [12] Reyes, C. P.; Muñoz-Martínez, F.; Torrecillas, I. R.; Mendoza, C. R.; Gamarro, F.; Bazzocchi, I. L.; Núñez, M. J.; Pardo, L.; Castanys, S.; Campillo, M.; Jiménez, I. A. Biological evaluation, structure-activity relationships, and three-dimensional quantitative structure-activity relationship studies of dihydro- $\beta$ -agarofuran sesquiterpenes as modulators of P-glycoprotein-dependent Multidrug Resistance. *Journal of medicinal chemistry* **2007**, *50*, 4808–4817.
- [13] Schomburg, C.; Schuehly, W.; Da Costa, F. B.; Klempnauer, K.-H.; Schmidt, T. J. Natural sesquiterpene lactones as inhibitors of Myb-dependent gene expression: Structure–activity relationships. *European journal of medicinal chemistry* **2013**, *63*, 313–320.
- [14] Haynes, P. Linear-scaling methods in ab initio quantum-mechanical calculations. Ph.D. thesis, University of Cambridge, 1998.
- [15] Griffiths, D. J. Introduction to quantum mechanics. *2nd, Pearson, Chapter2. The time-independent schrodinger equation* **2005**, 70–73.
- [16] Santra, B. Density-functional theory exchange-correlation functionals for hydrogen bonds in water. Ph.D. thesis, Technische Universität Berlin Berlin, 2010.
- [17] Cheung, D. L. G. Structures and properties of liquid crystals and related molecules from computer simulation. Ph.D. thesis, Durham University, 2002.
- [18] Toulouse, J. Introduction to density-functional theory. **2015**,
- [19] Abreu, J. Density Functional Theory: Systematic benchmarking of exchange and correlation functionals for the G2 dataset. M.Sc. thesis, 2015.
- [20] Dreizler, R. M.; Gross, E. K. *Density functional theory: an approach to the quantum many-body problem*; Springer Science & Business Media, 2012.
- [21] Kohn, W.; Becke, A. D.; Parr, R. G. Density functional theory of electronic structure. *The Journal of Physical Chemistry* **1996**, *100*, 12974–12980.
- [22] Parr, R. G. *Horizons of Quantum Chemistry*; Springer, 1980; pp 5–15.
- [23] Hossain, A. Introduction to density functional theory. **2004**,
- [24] Thomas, L. H. The calculation of atomic fields. **1927**, *23*, 542–548.
- [25] Kohn, W. Nobel Lecture: Electronic structure of matter—wave functions and density functionals. *Reviews of Modern Physics* **1999**, *71*, 1253.
- [26] Jones, R. O.; Gunnarsson, O. The density functional formalism, its applications and prospects. *Reviews of Modern Physics* **1989**, *61*, 689.



- [27] Zaleśny, R.; Papadopoulos, M. G.; Mezey, P. G.; Leszczynski, J. *Linear-Scaling Techniques in Computational Chemistry and Physics: Methods and Applications*; Springer Science & Business Media, 2011; Vol. 13.
- [28] Szabo, A.; Ostlund, N. S. *Modern quantum chemistry: introduction to advanced electronic structure theory*; Courier Corporation, 2012.
- [29] Hohenberg, P.; Kohn, W. Inhomogeneous electron gas. *Physical review* **1964**, *136*, B864.
- [30] Kohn, W.; Sham, L. J. Self-consistent equations including exchange and correlation effects. *Physical review* **1965**, *140*, A1133.
- [31] Perdew, J. P.; Schmidt, K. Jacob's ladder of density functional approximations for the exchange-correlation energy. **2001**, *577*, 1–20.
- [32] Vosko, S. H.; Wilk, L.; Nusair, M. Accurate spin-dependent electron liquid correlation energies for local spin density calculations: a critical analysis. *Canadian Journal of physics* **1980**, *58*, 1200–1211.
- [33] Capelle, K. A bird's-eye view of density-functional theory. *Brazilian journal of physics* **2006**, *36*, 1318–1343.
- [34] Becke, A. D. Density-functional exchange-energy approximation with correct asymptotic behavior. *Physical review A* **1988**, *38*, 3098.
- [35] Miehlich, B.; Savin, A.; Stoll, H.; Preuss, H. Results obtained with the correlation energy density functionals of Becke and Lee, Yang and Parr. *Chemical Physics Letters* **1989**, *157*, 200–206.
- [36] Perdew, J. P.; Ziesche, P.; Eschrig, H. Electronic structure of solids' 91. 1991.
- [37] Perdew, J. P.; Burke, K.; Ernzerhof, M. Generalized gradient approximation made simple. *Physical review letters* **1996**, *77*, 3865.
- [38] Wu, Z.; Cohen, R. E. More accurate generalized gradient approximation for solids. *Physical Review B* **2006**, *73*, 235116.
- [39] Becke, A. D. A new mixing of Hartree–Fock and local density-functional theories. *The Journal of chemical physics* **1993**, *98*, 1372–1377.
- [40] Perdew, J. P.; Ernzerhof, M.; Burke, K. Rationale for mixing exact exchange with density functional approximations. *The Journal of chemical physics* **1996**, *105*, 9982–9985.
- [41] Stephens, P. J.; Devlin, F. J.; Chabalowski, C. F.; Frisch, M. J. Ab initio calculation of vibrational absorption and circular dichroism spectra using density functional force fields. *The Journal of physical chemistry* **1994**, *98*, 11623–11627.
- [42] Jensen, F. *Introduction to computational chemistry*; John Wiley & sons, 2017.

- [43] Ahlrichs, R.; Bär, M.; Häser, M.; Horn, H.; Kölmel, C. Electronic structure calculations on workstation computers: The program system turbomole. *Chemical Physics Letters* **1989**, *162*, 165–169.
- [44] 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. *Physical Chemistry Chemical Physics* **2005**, *7*, 3297–3305.
- [45] Jensen, F. Atomic orbital basis sets. *Wiley Interdisciplinary Reviews: Computational Molecular Science* **2013**, *3*, 273–295.
- [46] Hill, J. G. Gaussian basis sets for molecular applications. *International Journal of Quantum Chemistry* **2013**, *113*, 21–34.
- [47] Kittel, C.; McEuen, P. *Introduction to solid state physics*; Wiley New York, 1976; Vol. 8.
- [48] Harrison, W. A. *Electronic structure and the properties of solids: the physics of the chemical bond*; Courier Corporation, 2012.
- [49] Neese, F.; Wennmohs, F.; Hansen, A.; Becker, U. Efficient, approximate and parallel Hartree–Fock and hybrid DFT calculations. A ‘chain-of-spheres’ algorithm for the Hartree–Fock exchange. *Chemical Physics* **2009**, *356*, 98–109.
- [50] Neese, F. An improvement of the resolution of the identity approximation for the formation of the Coulomb matrix. *Journal of computational chemistry* **2003**, *24*, 1740–1747.
- [51] Grimme, S.; Ehrlich, S.; Goerigk, L. Effect of the damping function in dispersion corrected density functional theory. *Journal of computational chemistry* **2011**, *32*, 1456–1465.
- [52] Johnson, E. R.; Becke, A. D. A post-Hartree-Fock model of intermolecular interactions: Inclusion of higher-order corrections. *The Journal of chemical physics* **2006**, *124*, 174104.
- [53] Grimme, S.; Antony, J.; Ehrlich, S.; Krieg, H. A consistent and accurate ab initio parametrization of density functional dispersion correction (DFT-D) for the 94 elements H–Pu. *The Journal of chemical physics* **2010**, *132*, 154104.
- [54] Becke, A. D. A multicenter numerical integration scheme for polyatomic molecules. *The Journal of chemical physics* **1988**, *88*, 2547–2553.
- [55] Ohno, K.; Esfarjani, K.; Kawazoe, Y. *Computational materials science: from ab initio to Monte Carlo methods*; Springer, 2018.
- [56] Johansson, M. P.; Kaila, V. R.; Sundholm, D. *Biomolecular Simulations*; Springer, 2013; pp 3–27.
- [57] Friesner, R. A. Ab initio quantum chemistry: Methodology and applications. *Proceedings of the National Academy of Sciences* **2005**, *102*, 6648–6653.

- [58] Jackson, C. M.; Esnouf, M. P.; Winzor, D. J.; Duewer, D. L. Defining and measuring biological activity: applying the principles of metrology. *Accreditation and quality assurance* **2007**, *12*, 283–294.
- [59] Pelikan, E. W. *Glossary of Terms and Symbols Used in Pharmacology*; Boston University School of Medicine, Pharmacology & Experimental Therapeutics, 1995.
- [60] Hohmann, M. S.; Longhi-Balbinot, D. T.; Guazelli, C. F.; Navarro, S. A.; Zarpelon, A. C.; Casagrande, R.; Arakawa, N. S.; Verri Jr, W. A. *Studies in natural products chemistry*; Elsevier, 2016; Vol. 49; pp 243–264.
- [61] Mukherjee, P. K. *Quality Control and Evaluation of Herbal Drugs: Evaluating Natural Products and Traditional Medicine*; Elsevier, 2019.
- [62] Ramakrishna, S.; Tian, L.; Wang, C.; Liao, S.; Teo, W. E. *Medical devices: regulations, standards and practices*; Woodhead Publishing, 2015.
- [63] Aslantürk, Ö. S. *In vitro cytotoxicity and cell viability assays: principles, advantages, and disadvantages*; InTech, 2018; Vol. 2.
- [64] Vardanyan, R.; Hruba, V. *Synthesis of essential drugs*; Elsevier, 2006.
- [65] Lowy, F. D. Staphylococcus aureus infections. *New England journal of medicine* **1998**, *339*, 520–532.
- [66] Mulligan, M. E.; Murray-Leisure, K. A.; Ribner, B. S.; Standiford, H. C.; John, J. F.; Korvick, J. A.; Kauffman, C. A.; Victor, L. Y. Methicillin-resistant Staphylococcus aureus: a consensus review of the microbiology, pathogenesis, and epidemiology with implications for prevention and management. *The American journal of medicine* **1993**, *94*, 313–328.
- [67] Andrews, J. M. Determination of minimum inhibitory concentrations. *Journal of antimicrobial Chemotherapy* **2001**, *48*, 5–16.
- [68] Gajewicz, A.; Schaeublin, N.; Rasulev, B.; Hussain, S.; Leszczynska, D.; Puzyn, T.; Leszczynski, J. Towards understanding mechanisms governing cytotoxicity of metal oxides nanoparticles: Hints from nano-QSAR studies. *Nanotoxicology* **2015**, *9*, 313–325.
- [69] Dudek, A. Z.; Arodz, T.; Gálvez, J. Computational methods in developing quantitative structure-activity relationships (QSAR): a review. *Combinatorial chemistry & high throughput screening* **2006**, *9*, 213–228.
- [70] Puzyn, T.; Leszczynski, J.; Cronin, M. T. *Recent advances in QSAR studies: methods and applications*; Springer Science & Business Media, 2010; Vol. 8.
- [71] Siedle, B.; García-Piñeres, A. J.; Murillo, R.; Schulte-Mönting, J.; Castro, V.; Rüngeler, P.; Klaas, C. A.; Da Costa, F. B.; Kisiel, W.; Merfort, I. Quantitative structure- activity relationship of sesquiterpene lactones as inhibitors of the transcription Factor NF- $\kappa$ B. *Journal of medicinal chemistry* **2004**, *47*, 6042–6054.

- [72] Schmidt, T. J.; Nour, A. M.; Khalid, S. A.; Kaiser, M.; Brun, R. Quantitative structure–antiprotozoal activity relationships of sesquiterpene lactones. *Molecules* **2009**, *14*, 2062–2076.
- [73] Merfort, I. Perspectives on sesquiterpene lactones in inflammation and cancer. *Current drug targets* **2011**, *12*, 1560–1573.
- [74] Lindenmeyer, M. T.; Hrenn, A.; Kern, C.; Castro, V.; Murillo, R.; Müller, S.; Laufer, S.; Schulte-Mönting, J.; Siedle, B.; Merfort, I. Sesquiterpene lactones as inhibitors of IL-8 expression in HeLa cells. *Bioorganic & Medicinal Chemistry* **2006**, *14*, 2487–2497.
- [75] Kar, S.; Gajewicz, A.; Puzyn, T.; Roy, K.; Leszczynski, J. Periodic table-based descriptors to encode cytotoxicity profile of metal oxide nanoparticles: A mechanistic QSTR approach. *Ecotoxicology and Environmental Safety* **2014**, *107*, 162–169.
- [76] Sizochenko, N.; Rasulev, B.; Gajewicz, A.; Kuz'min, V.; Puzyn, T.; Leszczynski, J. From basic physics to mechanisms of toxicity: The “liquid drop” approach applied to develop predictive classification models for toxicity of metal oxide nanoparticles. *Nanoscale* **2014**, *6*, 13986–13993.
- [77] Sizočenko, N. Optimal Selection of Descriptors for Structure-activity Modeling of Nanoparticles Based on Causality Analysis. Ph.D. thesis, 2016.
- [78] Berhanu, W. M.; Pillai, G. G.; Oliferenko, A. A.; Katritzky, A. R. Quantitative structure–activity/property relationships: the ubiquitous links between cause and effect. *ChemPlusChem* **2012**, *77*, 507–517.
- [79] Todeschini, R.; Consonni, V. *Handbook of molecular descriptors*; John Wiley & Sons, 2008; Vol. 11.
- [80] Cherkasov, A. *et al.* QSAR modeling: where have you been? Where are you going to? *Journal of medicinal chemistry* **2014**, *57*, 4977–5010.
- [81] Doucet, J. P.; Panaye, A. *Three dimensional QSAR: applications in pharmacology and toxicology*; CRC Press, 2010.
- [82] Todeschini, R.; Lasagni, M.; Marengo, E. New molecular descriptors for 2D and 3D structures. Theory. *Journal of chemometrics* **1994**, *8*, 263–272.
- [83] Schultz, T. W.; Netzeva, T. I.; Cronin, M. T. Selection of data sets for QSARs: analyses of Tetrahymena toxicity from aromatic compounds. *SAR and QSAR in Environmental Research* **2003**, *14*, 59–81.
- [84] OECD, O. Guidance document on the validation of (quantitative) structure-activity relationship [(Q) SAR] models. *Organisation for Economic Co-operation and Development: Paris, France* **2007**,
- [85] Puzyn, T.; Leszczynska, D.; Leszczynski, J. Toward the development of “Nano-QSARs”: Advances and challenges. *Small* **2009**, *5*, 2494–2509.

- [86] Pracht, P.; Bohle, F.; Grimme, S. Automated exploration of the low-energy chemical space with fast quantum chemical methods. *Physical Chemistry Chemical Physics* **2020**, *22*, 7169–7192.
- [87] Pracht, P.; Bauer, C. A.; Grimme, S. Automated and efficient quantum chemical determination and energetic ranking of molecular protonation sites. *Journal of Computational Chemistry* **2017**, *38*, 2618–2631.
- [88] Neese, F. The ORCA program system. *Wiley Interdisciplinary Reviews: Computational Molecular Science* **2012**, *2*, 73–78.
- [89] Schwall, K.; Zielenbach, K. SciFinder a new generation of research tool. *Chemical innovation* **2000**, *30*, 45–50.
- [90] Gabrielson, S. W. SciFinder. *Journal of the Medical Library Association: JMLA* **2018**, *106*, 588.
- [91] Rhodes, G. *Crystallography made crystal clear: a guide for users of macromolecular models*; Elsevier, 2010.
- [92] Grant, M. C.; Boyd, S. P. *Recent advances in learning and control*; Springer, 2008; pp 95–110.
- [93] Allouche, A.-R. Gabedit—A graphical user interface for computational chemistry softwares. *Journal of computational chemistry* **2011**, *32*, 174–182.
- [94] Chaturvedi, D. Sesquiterpene lactones: structural diversity and their biological activities, In-Opportunity, Challenges and Scope of Natural Products in Medicinal Chemistry. ISBN: 978-81-308-0448-4, *Research Signpost, Trivandrum* **2011**, 313–334.
- [95] Ordóñez, P. E.; Quave, C. L.; Reynolds, W. F.; Varughese, K. I.; Berry, B.; Breen, P. J.; Malagón, O.; Smeltzer, M. S.; Compadre, C. M. Sesquiterpene lactones from *Gynoxys verrucosa* and their anti-MRSA activity. *Journal of ethnopharmacology* **2011**, *137*, 1055–1059.
- [96] Sun, C.-M.; Syu, W.-J.; Don, M.-J.; Lu, J.-J.; Lee, G.-H. Cytotoxic Sesquiterpene Lactones from the Root of *Saussurea lappa*. *Journal of natural products* **2003**, *66*, 1175–1180.
- [97] Gibbons, S. Anti-staphylococcal plant natural products. *Natural product reports* **2004**, *21*, 263–277.
- [98] SATO, Y.; OKETANI, H.; YAMADA, T.; SINGYOUCHI, K.-I.; OHTSUBO, T.; KIHARA, M.; SHIBATA, H.; HIGUTI, T. A Xanthanolide with Potent Antibacterial Activity against Methicillin-resistant *Staphylococcus aureus*. *Journal of pharmacy and pharmacology* **1997**, *49*, 1042–1044.
- [99] Kawazoe, K.; Tsubouchi, Y.; Abdullah, N.; Takaishi, Y.; Shibata, H.; Higuti, T.; Hori, H.; Ogawa, M. Sesquiterpenoids from *Artemisia g. vilscens* and an Anti-MRSA Compound. *Journal of natural products* **2003**, *66*, 538–539.
- [100] Appalasaamy, S.; Lo, K. Y.; Ch'ng, S. J.; Nornadia, K.; Othman, A. S.; Chan, L.-K. Antimicrobial activity of artemisinin and precursor derived from in vitro plantlets of *Artemisia annua* L. *BioMed research international* **2014**, *2014*.

- [101] Siddhardha, B.; Ramakrishna, G.; Basaveswara, R. In Vitro Antibacterial Efficacy of a Sesquiterpene Lactone, Parthenin From *Parthenium Hysterophorus* L (Compositae) Against Enteric Bacterial Pathogens. *International journal of pharmaceutical, chemical and biological sciences* **2012**, *2*, 206–209.
- [102] Bohlmann, F.; Fritz, U.; King, R. M.; Robinson, H. Fourteen heliangolides from *Calea* species. *Phytochemistry* **1981**, *20*, 743–749.
- [103] Arena, M. E.; Cartagena, E.; Gobatto, N.; Baigori, M.; Valdez, J. C.; Bardon, A. In vivo and in vitro antibacterial activity of acanthospermal B, a sesquiterpene lactone isolated from *Acanthospermum hispidum*. *Phytotherapy Research* **2011**, *25*, 597–602.
- [104] Gören, N.; Woerdenbag, H. J.; Bozok-Johansson, C. Cytotoxic and antibacterial activities of sesquiterpene lactones isolated from *Tanacetum praeteritum* subsp. *praeteritum*. *Planta medica* **1996**, *62*, 419–422.
- [105] Rabe, T.; Mullholland, D.; Van Staden, J. Isolation and identification of antibacterial compounds from *Vernonia colorata* leaves. *Journal of Ethnopharmacology* **2002**, *80*, 91–94.
- [106] Villarreal, M. L.; Alvarez, L.; Alonso, D.; Navarro, V.; Garcia, P.; Delgado, G. Cytotoxic and antimicrobial screening of selected terpenoids from Asteraceae species. *Journal of ethnopharmacology* **1994**, *42*, 25–29.
- [107] Schmidt, T. J.; Heilmann, J. Quantitative Structure-Cytotoxicity Relationships of Sesquiterpene Lactones derived from partial charge (Q)-based fractional Accessible Surface Area Descriptors (Q<sub>fr</sub>ASAs). *Quantitative Structure-Activity Relationships* **2002**, *21*, 276–287.
- [108] Khalafi-Nezhad, A.; Rad, M. S.; Mohabatkar, H.; Asrari, Z.; Hemmateenejad, B. Design, synthesis, antibacterial and QSAR studies of benzimidazole and imidazole chloroalkoxyalkyl derivatives. *Bioorganic & medicinal chemistry* **2005**, *13*, 1931–1938.
- [109] Cardoso, F. J. B.; de Figueiredo, A. F.; da Silva Lobato, M.; de Miranda, R. M.; de Almeida, R. C. O.; Pinheiro, J. C. A study on antimalarial artemisinin derivatives using MEP maps and multivariate QSAR. *Journal of Molecular Modeling* **2008**, *14*, 39–48.
- [110] Eroglu, E.; Türkmen, H. A DFT-based quantum theoretic QSAR study of aromatic and heterocyclic sulfonamides as carbonic anhydrase inhibitors against isozyme, CA-II. *Journal of Molecular Graphics and Modelling* **2007**, *26*, 701–708.
- [111] Sharma, M.; Sahu, N.; Kohali, D.; Chaturvedi, S.; Sharma, S. QSAR, SYNTHESIS AND BIOLOGICAL ACTIVITY STUDIES OF SOME THIAZOLIDINONES DERIVATIVES. *Digest Journal of Nanomaterials & Biostructures (DJNB)* **2009**, *4*.
- [112] Rasulev, B.; Saidkhodzhaev, A.; Nazrullaev, S.; Akhmedkhodzhaeva, K.; Khushbaktova, Z.; Leszczynski, J. Molecular modelling and QSAR analysis of the estrogenic activity of terpenoids isolated from *Ferula* plants. *SAR and QSAR in Environmental Research* **2007**, *18*, 663–673.

- 
- [113] Kurihara, T.; Mine, H.; Satoh, Y.; Wakabayashi, H.; MOTOHASHI, N.; Sakagami, H. Relationship between electronic structure and cytotoxic activity of tropolones. *in vivo* **2006**, *20*, 391–395.
- [114] Zhangabylov, N.; Dederer, L. Y.; Gorbacheva, L.; Vasil'eva, S.; Terekhov, A.; Adekenov, S. Sesquiterpene lactone arglabin influences DNA synthesis in P388 leukemia cells in vivo. *Pharmaceutical Chemistry Journal* **2004**, *38*, 651–653.





# Abbreviations

**B3LYP** Becke, 3-parameter, Lee–Yang–Par 20, 30, 33

**B88** Becke 88 19, 20

**c-2** Carbon number two 5

**c-3** Carbon number three 5

**c-8** Carbon number eight 5

**CoMSIA** Comparative molecular similarity indices analysis 5

**COSX** Chain-Of-Spheres exchange 23

**CRE** Conformer Rotamer Ensemble 29

**CREST** Conformer-Rotamer Ensemble Sampling Tool iii, 29, 30, 32, 33

**DFT** density functional theory iii, 4, 6, 7, 12, 14, 16, 19, 23, 26, 30, 31, 33

**DFT-D3** damping DFT 23

**DNP** Dictionary of Natural Products 1, 24

**DZ** double zeta 22

**GEA** gradient-expansion approximation 19

**GGA** generalized gradient approximation 17, 19, 20

**GTO** Gaussian type orbital 21, 22

**HF** Hartree-Fock 13, 20

**HK** Hohenberg-Kohn theorem 14, 16

**HOMO** highest occupied molecular orbital iii, 4

**LCAO** linear combination of atomic orbitals theory 21, 22

**LDA** local density approximation 17–20

**LUMO** lowest unoccupied molecular orbital 4

**LYP** Lee-Yang-Parr 19, 20

**MLR** multiple linear regression 34

- MRSA** *Staphylococcus Aureus* methicillin-resistant iii, 25, 31, 34, 35, 37
- PBE** Perdew-Burke-Ernzerhof 19
- PDB** Protein Data Bank 4
- PES** potential energy surface 29, 30
- Pgp** P-glycoprotein 5
- PP** polarization function 22
- PW91** Perdew-Wang 91 19
- PW92** Perdew and Wang 19
- QSAR** Quantitative Structure-Activity Relationship iii, 4–7, 25–27, 31, 33, 34, 36
- RMSD** root-mean-square deviation 29
- SAR** Structure-Activity Relationship 5
- SCF** self-consistent-field 14
- STL** *Sesquiterpene Lactones* iii, 1–7, 24, 29–33, 37, 40
- STO** Slater-type orbital 21, 22
- SZ** single-zeta 22
- TF** Thomas-Fermi 12, 13
- TZ** triple zeta 22
- TZV** triple zeta valance 22
- VWN** Vosko, Wilk, and Nusair 19