

POLITECNICO DI TORINO

Department of Mechanical and Aerospace Engineering

Master's Degree in Biomedical Engineering

**Computational Modeling of the Heat Shock Protein  
HSP90 Alpha and A Search for Its Small Molecule  
Inhibitors**



**Thesis Supervisor**

Prof. Jack A. Tuszynski

**Co-supervisor**

Prof. Marco A. Deriu

Dr. Maral Aminpour

**Candidate**

Alexia Moracchiato

*March 2023*



## ABSTRACT

Heat shock protein 90 (Hsp90) is a chaperone protein that protects proteins from heat stress, aids in the correct folding of other proteins and assists in protein degradation. It also stabilizes several proteins required for tumor growth, which is why Hsp90 inhibitors are investigated as anti-cancer drugs. The inhibition of Hsp90 causes cell death, as it induces the ubiquitin-proteasome system to eliminate the incorrectly folded proteins within the tumor cells, whose proliferation is induced by the inhibition of Hsp90.

Another important role of Hsp90 in cancer is the stabilization of mutant proteins, it seems that Hsp90 can intervene to maintain the correct folding of the less stable proteins produced by DNA mutations, making the effect of these mutations phenotypically less relevant.

It is also worth mentioning the implication of this protein in neurodegenerative diseases such as Alzheimer's and Parkinson's diseases and, according to recent discoveries, also in multiple sclerosis and in spinal and bulbar muscular atrophy.

This project focuses on the creation of an equilibrated model of this protein under conditions typical of cancer cells and on the demonstration of the mode of action of some Hsp90 inhibitors. Investigating the chaperone using computational methodologies, specifically both molecular docking and molecular dynamics simulations, have proven to be valuable tools for exploring and fully understanding the binding between Hsp90 and these compounds. It was also important to find similar compounds in the ZINC database to see if other inhibitors could have better pharmacological profiles.

All of this was made possible by the abundance of data on X-ray resolved crystallographic structures, which aided the work by making available the structures of many known inhibitors. This also has influenced the choice of a starting model complex, on which the molecular docking protocol was later based.

Pharmacokinetic analysis carried out in this work, as well as the other strategies, are aimed at improving the efficacy of Hsp90 inhibitors to maximize the full potential of this pharmaceutical class.



## ACKNOWLEDGEMENTS

I would like to express my sincere gratitude to Prof. Jack A. Tuszynski and Dr. Maral Aminpour for their assistance at every stage of the research project.

I would like to thank Dr. Quian Wang, Dr. Giannoula Lakka Klement and Dr. Christos for their insightful comments and suggestions.

Without the pre-exam moral support from my peers, the laughter outside the classroom, the group works and the last-minute doubts I could never have gotten this far. Thank you for always being there, especially in moments of discouragement.

I wish to extend my special thanks to the “few but good”, my true friends, Federico and Euxhenio, for their unwavering support and belief in me. Thanks for listening to my outbursts, thanks for all the carefree moments.

To my grandmother Franca, who cannot be here today with me, but who I hope will look at me from up there and who is proud of me and of the woman I have become.

I sincerely thank my nephew Alessandro, who, despite his tender age, has shared with me the joys and hardships of these years. Whenever I needed him, despite the kilometers that separate us, he was always there.

I am deeply grateful to my family, Maria, Natale and Alien, which are the pillar of my life, the foundation of my days. This thesis is for them and to them I dedicate the joy that crossing the finish line of graduation ignites in my heart. With boundless gratitude.

Finally, I would like to dedicate this milestone to myself, for never giving up, which could be the beginning of a long and brilliant professional career.

*Ad maiora semper.*



## SUMMARY

<b>ABSTRACT.....</b>	<b>3</b>
<b>INTRODUCTION .....</b>	<b>9</b>
<b>OVERVIEW ON COMPUTATIONAL METHODS .....</b>	<b>24</b>
<b>Molecular Modeling &amp; Molecular Mechanics .....</b>	<b>24</b>
The Boltzmann Law .....	25
Simulating the environment .....	34
Energy minimization .....	37
<b>Molecular dynamics .....</b>	<b>39</b>
Free energy .....	44
<b>Molecular docking .....</b>	<b>47</b>
Docking simulation .....	48
Search algorithm .....	49
Flexibility of the ligand.....	50
Flexibility of the receptor .....	50
Score function .....	50
<b>ADMET properties and PK .....</b>	<b>51</b>
Absorption .....	52
Distribution .....	53
Metabolism .....	53
Excretion .....	54
Toxicity .....	54
QSAR model .....	54
<b>MOLECULAR DOCKING .....</b>	<b>57</b>
<b>Receptor .....</b>	<b>57</b>
<b>Ligands .....</b>	<b>58</b>
Rescoring .....	61
<b>Results .....</b>	<b>62</b>
<b>MOLECULAR DYNAMICS.....</b>	<b>64</b>
<b>ANALYSIS OF ADMET PROPERTIES.....</b>	<b>68</b>
<b>Material and methods.....</b>	<b>68</b>
ADMET Predictor Simulations Plus.....	68
SwissADME .....	69
pkCSM .....	74
<b>Results .....</b>	<b>77</b>
Absorption .....	78
Distribution .....	79
Metabolism .....	80
Excretion .....	81

Toxicity .....	82
<b>CONCLUSION .....</b>	<b>85</b>
<b>BIBLIOGRAPHY .....</b>	<b>87</b>



## INTRODUCTION

Native structure is the conformation that allows a protein to perform its correct function. The process that the protein goes through to reach its native conformation is called "folding". The protein complexes that assist proteins during folding are called "molecular chaperones".

Folding is a two-stage process: the first stage is fast and sees the formation of the so-called molten globule; the second stage, on the other hand, occurs more slowly and requires molecular chaperones to assist the protein in reaching its native structure, starting from the aggregation nucleus that was formed in the first stage.

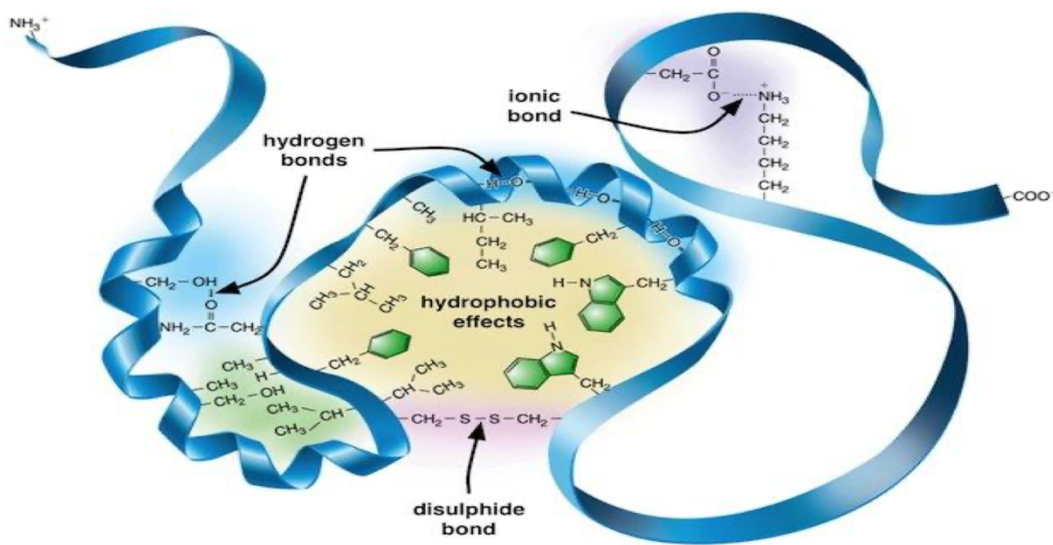


Figure 1: Protein folding.

Molecular chaperones are proteins and protein complexes responsible for assisting proteins during folding. These are distinguishable into sHsp (small Heat-shock proteins), which have a molecular weight between 10 and 40 kDa, and Hsp, which have a molecular weight ranging from 40 to about 200 kDa. Furthermore, some chaperones (such as Hsp70) are part of a first level quality control (QCI), which deals with the folding of nascent proteins; others (mostly families of chaperone proteins) play a role in second level quality control (QCII), which has the task of controlling the post-synthesis protein conformation [1-8]. The Heat Shock Proteins (Hsp), whose acronym is accompanied by the molecular weight, play several essential roles:

- The control of the folding of the polypeptide chains arising in the active conformation.

- The prevention of aggregation of multiple polypeptide chains in inactive conformation.
- The quality control of proteins that have lost their structure native tertiary.
- The control of the activity and stability of the proteins they have taken their tertiary structure.

The division based on the molecular weight allows to distinguish six families: Hsp100 (100-110 kDa), Hsp90 (83-90 kDa), Hsp70 (66-78 kDa), Hsp60, Hsp40 and small Hsp (15-30 kDa). They deputies to perform specific functions in different cellular compartments.

Hsp90s are involved in the conformational maturation of nascent peptics and in the renaturation of denatured proteins, and they also appear to be involved in signal transduction, given their interactions with steroid hormone receptors and various cell cycle kinases. The chaperone Hsp90 is responsible for mechanisms of maturation and renaturation of proteins in their biologically active form. <sup>[5-10]</sup>

Hsp90 is also the most abundant molecular chaperone in eukaryotic cells, constituting 1-2% of the total protein content.<sup>[11]</sup>

In the human proteome there are 4 isoforms of Hsp90:

- Hsp90 $\alpha$ , inducible and quantitatively more abundant isoform.
- Hsp90 $\beta$ , constitutive minor form.
- GRP94 or Glucose-Regulated Protein expressed in the lattice endoplasmic, member of 94 kDa.
- TRAP-1 or hsp75 (Tumor Necrosis Factor Receptor Associated Protein 1) which is located in the mitochondrial matrix [12].

Hsp90 forms a dimer at physiological temperatures. Each protomer consists of three domains: N-terminal domain of 25 kDa (NTD), middle-domain of 35 kDa (MD) and C-terminal domain of 12 kDa (CTD). Some members of the Hsp90 family, including cytosolic eukaryotic Hsp90s and Grp94, include a disordered region known as the charged linker that separates NTD and MD. Aside from the charged linker, cytosolic eukaryotic Hsp90s feature a MEEVD C-terminal extension. This domain is essential for Hsp90 dimerization. It also aids in the binding of client proteins including the tumor suppressor.

In the N-terminal region, on the other hand, they contain a binding site for ATP (also known as a GHKL type ATPase domain) with hydrolytic activity. The drugs analyzed in this project prevent the binding between ATP and Hsp90 and thus causes the accumulation of Hsp90 complexes and misfolded proteins in cancer cells. This causes the ubiquitin-proteasome system to destroy the abnormal

proteins and as a result the cancer cells die because the signaling pathway that controls cell growth is altered.

As a result, Hsp90 inhibition causes the ubiquitin-proteasome system to degrade critical oncogenic client proteins, inhibiting tumor development and activating apoptosis in cancer cells.

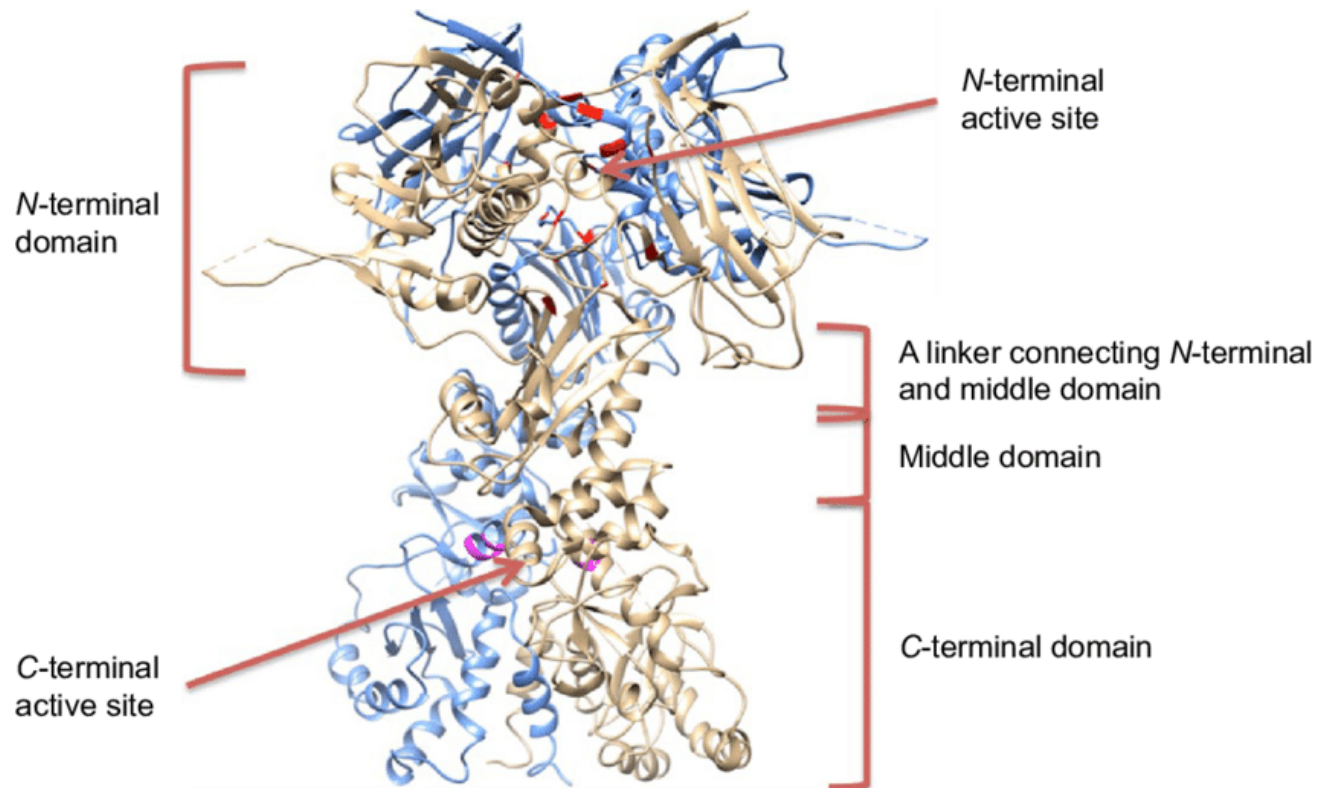


Figure 2: Domains of Hsp90.

Inhibition of Hsp90 results in simultaneous destabilization and degradation of client proteins which result in suppression of tumor growth.

Hsp90 inhibitors currently described in the literature are divided into different classes based on the different modes of inhibition:

- Blocking the ATP bond.
- Interruption of the cochaperon/Hsp90 interaction.
- Client protein antagonism/Hsp90.
- Interference with post-transductional modifications of Hsp90.

Numerous Hsp90 inhibitors work by binding to the N-terminal ATP pocket, however the chaperone can also be inactivated in other ways. Compounds that bind with its C-terminus or change its posttranslational state are two further strategies to inhibit Hsp90 action. In contrast to the N-terminus, the crystal structure of the C-terminal region has not been solved so far. This area is thought

to be involved in the binding of a second ATP molecule. According to research, the region is only accessible to ATP once the N-terminal ATP pocket is filled by ATP or an inhibitor such as geldanamycin. Although the role of this location in the function of Hsp90 is unknown, it is thought to modulate the ATPase activity of the N-terminal region [13]. Interactions with this area of the chaperone may potentially impede Hsp90 function and have anticancer consequences.

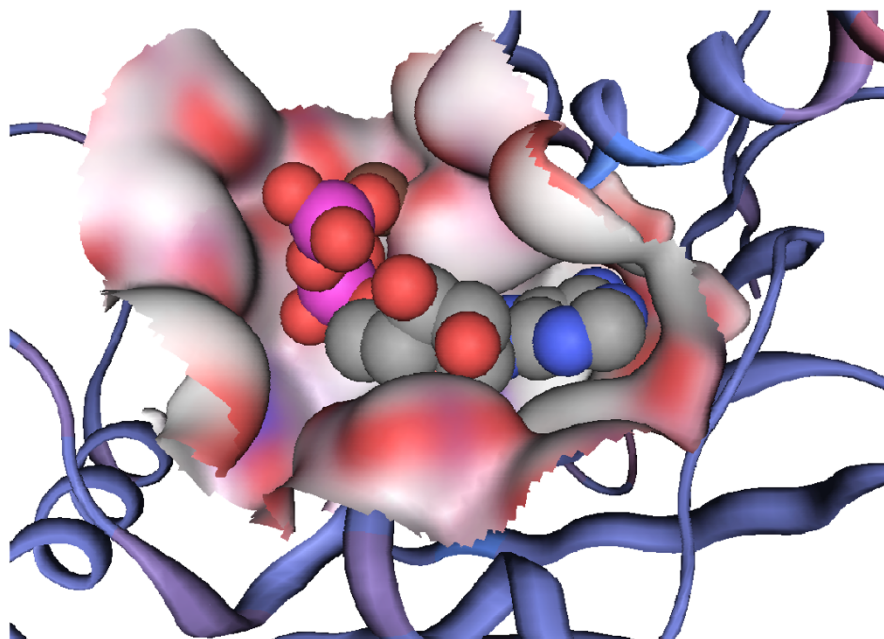


Figure 3: Hsp90 binding Geldanamycin.

The first Hsp90 inhibitor found was Geldanamycin, an ansamycin derived from bacteria with a 19-membered macrocycle, a benzoquinone ring, and lactam activity. It has antibacterial action against a wide range of Gram-positive and Gram-negative bacteria. It functions as an antiviral, antineoplastic, antibacterial, cysteine protease inhibitor, and a Hsp90 inhibitor. It is a carbamate ester, an ansamycin, an organic heterobicyclic molecule, and a 1,4-benzoquinone. Despite having antibacterial and anticancer activities, research was discontinued due to toxicity and low solubility. Toxicity is caused by a reaction of biological nucleophiles at the 19-position of geldanamycin's quinone ring.

Affinity chromatography and crystallographic analyses indicated that geldanamycin binds in a U-shaped conformation inside a deep pocket at the Hsp90 ATP-binding site, with the ansa ring and the benzoquinone folded roughly parallel to one other and the lactam moiety in a cis configuration. This understanding has enabled the application of structure-based design tactics that include structural elements that favor the cis-amide bond configuration.

Even if geldanamycin was the first Hsp90 inhibitor to enter clinical trials, it was not advanced because of its unacceptable hepatotoxicity. As a result, substituting different substituents for the 17-methoxy group resulted in less lethal analogs such as Tanespimycin (17-allylaminogeldanamycin, 17-AAG) [9].

Tanespimycin is a geldanamycin with a 19-membered macrocycle in which the methoxy substituent linked to the benzoquinone moiety has been replaced with an allylamino group. It is an effective inhibitor of the heat shock protein 90 (Hsp90). It is a less toxic derivative of geldanamycin that promotes apoptosis and has antitumor properties. It functions as an antineoplastic agent, a Hsp90 inhibitor, and an inducer of apoptosis. It is an ansamycin, a carbamate ester, an organic heterobicyclic molecule, and a member of the 1,4-benzoquinones. It is derived from the antibiotic geldanamycin.

It has been tried in over 30 clinical studies (phase I/II) in both solid and hematologic malignancies, both as a single agent and in combination with chemotherapy or targeted medicines [10]. Tanespimycin's early phase I studies were unsatisfactory, with only little efficacy shown in several tumor types [11]. Single-agent Tanespimycin's low effectiveness has been ascribed, at least in part, to inefficient inhibition of target client proteins, most likely due to insufficient drug dosage or frequency of administration, unpredictable pharmacokinetics, inappropriate formulation, and other factors and dose-limiting toxicities, including hepatotoxicity. However, promising activity was seen in a phase II study in HER2<sup>+</sup> breast cancer [14]. Tanespimycin is also susceptible to multidrug resistance mechanisms such as p-glycoprotein-mediated efflux, as well as polymorphic-reductive metabolism of the benzoquinone by the enzymes NQO1/DT-diaphorase or CYP3A4 [15]. Although quinone metabolism boosts the drug's HSP90 inhibitory efficacy, it is likely to contribute to the reported liver damage and may represent a primary and acquired resistance mechanism [16].

Tanespimycin is the first Hsp90 inhibitor to be tested in a clinical trial.

The invention of the water-soluble analog Alvespimycin solved the solubility problem (17-dimethylaminoethylaminogeldanamycin, 17-DMAG) [17].

Alvespimycin is a geldanamycin with a 19-membered macrocycle in which the methoxy group linked to the benzoquinone moiety has been replaced with a 2-(N,N-dimethylamino)ethylamino group. It functions as a Hsp90 inhibitor. It is a tertiary amino molecule, an ansamycin, a member of the 1,4-benzoquinones, and a carbamate ester. It is derived from the antibiotic geldanamycin.

It has been used in trials studying the treatment of solid tumor in various cancer as an antitumor agent. Tanespimycin is also susceptible to multidrug resistance mechanisms such as p-glycoprotein-mediated efflux, as well as polymorphic-reductive metabolism of the benzoquinone by the enzymes NQO1/DT-diaphorase or CYP3A4 [18]. Although quinone metabolism boosts the drug's



HSP90 inhibitory efficacy, it is likely to contribute to the reported liver damage and may represent a primary and acquired resistance mechanism.

Rifabutin is a rifamycin antibiotic that is structurally and functionally similar to rifampin and rifapentine and is primarily used to prevent *Mycobacterium avium* complex (MAC) illness in individuals with advanced HIV infection. Rifabutin is linked to temporary and asymptomatic increases in serum aminotransferase and is a potential cause of clinically evident, acute liver damage. It contains an ansamycin moiety [19-20]. It works by preventing or slowing the growth of some bacteria by preventing them from synthesizing RNA. Other research has revealed that rifabutin is beneficial against cryptosporidiosis, another parasite infection of the intestine [21-22].

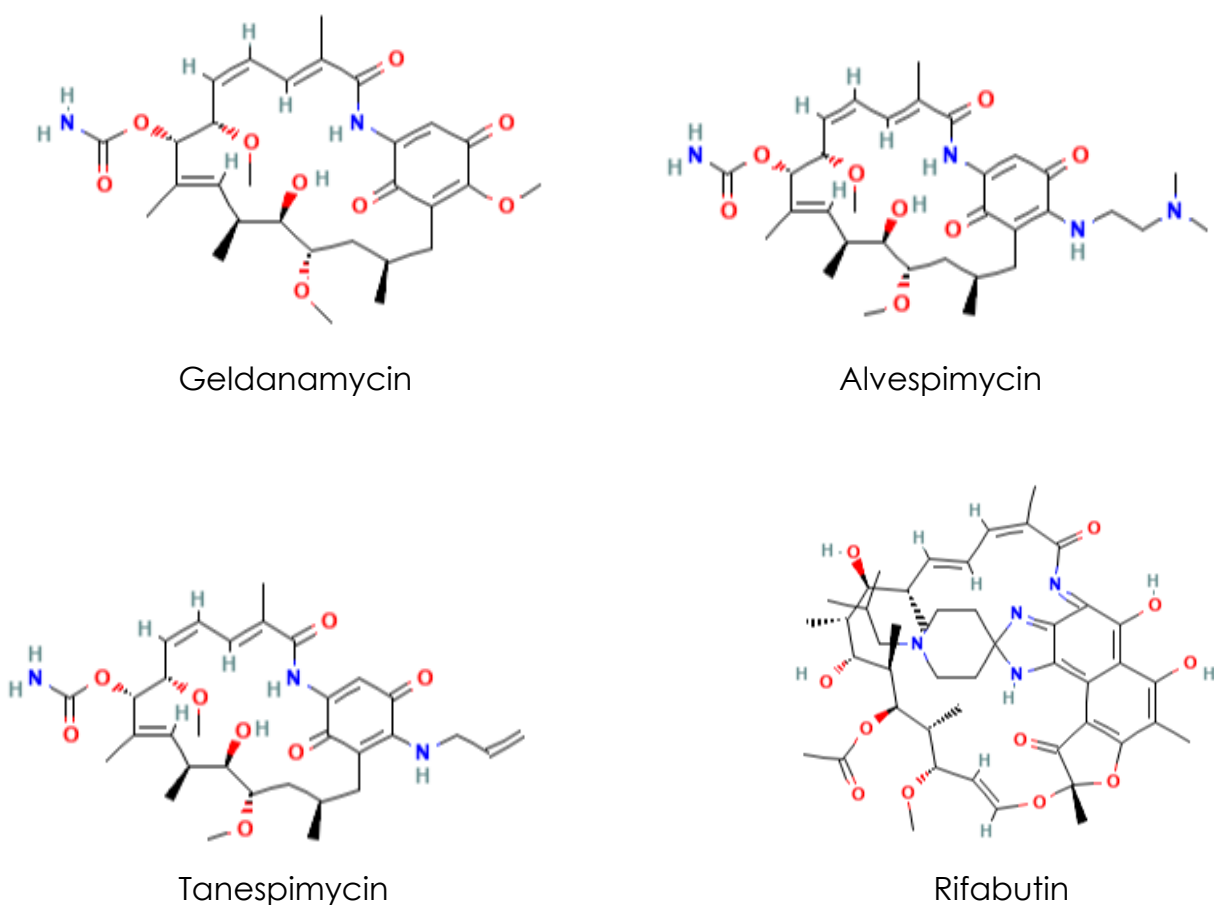


Figure 4: Chemical structures of the reference molecules [62].

The expression of Hsp90 in tumor cells [23] and mutagenesis experiments have shown that the survival of eukaryotic organisms is dependent on Hsp90, however, cancer cells have shown particular sensitivity to small molecules that inhibit its activity.

The overexpression or accumulation of denatured proteins are responsible for diseases such as cancer, multiple sclerosis, spinal and bulbar muscular atrophy, Alzheimer and Parkinson [16]; for this the inhibitors of Hsp90 appear to be promising chemotherapy and that target has become one of the most studied in the world.

The terms “cancer” and “tumor” refer to a pathological condition characterized by the uncontrolled proliferation of cells that have the ability to infiltrate the normal organs and tissues of the body, altering their structure and functioning. Cancer also has the ability to localize itself at a distance from the primary disease, and, in this case, we speak of secondary or metastatic disease. Tumors are divided into solid tumors, characterized by a compact mass of tissue, and blood tumors (lymphomas and leukemias). It is possible to classify tumors in various different ways, depending on the organ in which they develop, the type of cells that are formed, the stage at which the disease is at diagnosis, aggression and the possibility of development of metastasis. Cancer is caused by DNA mutations within cells. Cellular DNA contains information on how cells need to grow and multiply. Errors in these instructions cause the cell to become cancerous.

Genetic mutations can cause a healthy cell to do the following:

- it multiplies abnormally, thus creating more diseased cells.
- doesn't fight abnormal cell growth: normal cells contain genes called tumor suppressors, which recognize abnormal cell growth and act to stop it. When there is an error in these genes, that function can be weakened or even disrupted. This allows the mutated cells to continue growing and dividing.
- make mistakes in DNA repair: Genes are able to identify and repair errors that may be present within the DNA itself. A mutation can mean that some errors are not detected, allowing the accumulation of different mutations and, ultimately, the appearance of cancer.

Genetic mutations can be present from birth, or, in other cases, they can be caused by viruses, chronic inflammation or by the same hormones produced by the body. However, they can also be caused by factors external to the body, such as ultraviolet (UV) rays, carcinogenic chemicals or radiation.

Scientists think that the simultaneous presence of more than one mutation is necessary to give rise to most cancers.

Some hematological cancers can result from a single mutation, but most of those that form in internal organs, such as the lungs and colon, start from many different mutations. It is not yet clear how many mutations need to be added together to give rise to a tumor, although it is thought that this varies according to the type of tumor.

The genetic mutations present at birth add up with those acquired over time and give rise to cancer. This means that the mutations present at birth are necessary but not sufficient for the development of cancer. One or more other mutations will be needed to cause the tumor.

The genetic mutation only makes it more likely that a person will develop cancer when exposed to other risk factors. Mutations initiate the process, while risk factors play a role in the development of the disease.

Cancer can therefore be defined as a multifactorial disease, both genetic mutations and environmental factors play a fundamental role in it.

One in five persons will acquire cancer over their lives. The authors of the report entitled "Global Cancer Statistics 2020", produced in collaboration by the American Cancer Society (ACS) and the International Agency for Research on Cancer (IARC), say so.

Many proteins responsible for malignant progression in cancer cells are Hsp90-dependent client proteins. [24] Indeed, more than 40 oncogenic substrates of Hsp90 have been identified so far, in addition Hsp90 is overexpressed in malignant cell lines, and its expression is correlated with the proliferation of this type of cells. [25] Interestingly, these proteins score they are involved in each of the key processes that lead to malignant neoplastic proliferation; and that the inhibition of Hsp90 allows a unified mechanism for the simultaneous degradation of multiple oncogenic targets [26]. As a result, Hsp90 inhibitors have emerged as a promising class of drugs for the treatment of numerous types of cancer. There are currently more than 20 ongoing clinical trials targeting Hsp90. These inhibitors have a high differential selectivity between malignant and untransformed cells, furthermore the inhibition occurs at concentrations well tolerated by patients [27].

Multiple sclerosis is an inflammatory disease of the central nervous system characterized by the loss of myelin (the substance that lines the nerve fibers of the white matter) in multiple areas (hence the name "multiple"). The demyelination process can cause damage or loss of myelin and the formation of lesions (plaques) that can evolve from an initial inflammatory phase to a chronic phase in which they take on scar-like characteristics from which the term "sclerosis" derives.

Multiple sclerosis occurs at any age of life, but the most affected subjects are those between 20 and 40 years of age; women are affected twice as many as men. The prognosis is very variable: the most common form is characterized by



phases in which the disease manifests itself interspersed with phases of remission of different duration. In the early stages of the disease, the regression of the signs is almost complete, but with the passage of time the symptoms persist longer and longer, giving rise to progressive disability. The underlying causes of multiple sclerosis are still unknown. It is currently believed that multiple sclerosis is an autoimmune disease: at the basis of the loss of myelin there is in fact an alteration in the response of the immune system that would attack the myelin as if it were an external agent to be fought. There are several factors that play a role of some importance in the onset of multiple sclerosis, which are still being studied:

- the environment (countries with a temperate climate are at greatest risk).
- ethnicity (Caucasian origin determines a greater predisposition).
- exposure to infectious agents (viruses, bacteria), especially in the first years of life.
- genetic predisposition.

MS is widespread all over the world, but the distribution of the disease is not uniform; it is more widespread in areas far from the equator, in particular in the United States, Northern Europe, New Zealand, Canada and Australia. The prevalence is 100-190 per 100,000 inhabitants in Northern Europe, the United States, Canada and New Zealand, while it drops to 2-25 per 100,000 inhabitants in Asia, Africa and South America. At the regional level, there are fewer gender disparities in Europe and America than in Asia, Africa and Oceania.

According to the "Atlas of MS" cases worldwide increased from 2.1 million in 2008 to 2.3 million in 2013. As regards the American continent, the United States and Canada are the countries with the highest prevalence rates (respectively 135 and 291 cases per 100 thousand inhabitants). The spread of the disease is considerably lower in the countries of Central and South America. More recent studies estimate that around one million people are affected by MS in the US, about double the previous estimates.

Most African countries do not have data on the spread of MS. However, the disease is more widespread in South Africa (5 cases per 100 thousand inhabitants), Egypt (25), Morocco (20), Tunisia (20) and Algeria (20).

It is important to note that the risk of SM can differ substantially between different ethnic groups within the same geographical region. For example, in South Africa, MS is less common in blacks than in whites, in Australia in natives than in those born overseas, in New Zealand in Maori than in Europeans. For this reason, ethnic origin (ethnicity) must be considered in future prevalence and incidence studies.

The expression of Hsp90 (as antigen) on the surface of OPCs of MS patients has recently been reported. In addition, IgG specifically directed to Hsp90 were found in the spinal fluid, thus creating complexes that lead to the death of OPCs. These data also indicate that the inhibition of hsp90 could lead to an increase in the expression of Hsp90 with a potential decrease in the number of deaths caused by the loss of OPCs [28].

Bulbar spinal muscular atrophy (also called Kennedy's disease) is a progressive neuromuscular disease. It is characterized by muscle weakness that mainly affects the muscles of the lower limbs and those responsible for wording and chewing. It is often associated with endocrinological disorders such as infertility and insensitivity to androgens. In addition to muscle weakness, the disease manifests itself (in adulthood and in males) with involuntary muscle contractions, cramps or tremors.

Bulbar spinal muscular atrophy is caused by alterations in the gene coding for the androgen receptor, located on the X chromosome. These alterations consist of an excessive number of repetitions of a sequence of three nucleotides of DNA, normally present in the gene. The disease is transmitted in an X-linked manner: only males (who have only one X chromosome) show symptoms, while females are healthy carriers, as the toxicity of the mutated androgen receptor occurs only when it is activated by the testosterone (a typical male hormone). The prevalence of BSMA is 1/30,000 male births. The incidence is 1/526,315 males/year.

It is known that AR is a client protein of Hsp90, and that the complex with mutated AR is more stable than the complex with wild type AR. Treatment with 17-AAG has demonstrated the ability to initiate specific degradation of the mutated AR, suggesting another potential therapeutic use [29].

Alzheimer's disease is a neurological illness that is chronic and progressive. It is the most common cause of dementia in the elderly population of developed countries: currently it is estimated that about 5% of the population over 65 and about 20% of those over 85 are affected, although in several cases it can also occur. An early onset around the age of 50. This disease, which takes its name from the German neurologist Alois Alzheimer who first described its characteristics in the early 1900s, is characterized by a progressive degenerative process that destroys brain cells, causing irreversible deterioration of cognitive functions (memory, reasoning and language), to the point of compromising autonomy and the ability to carry out normal daily activities. 1% of Alzheimer's cases is caused by the presence of an altered gene that determines its transmission from one generation to another in the same family. The remaining 99% of cases occurs in a "sporadic" way, i.e., in people who are not clearly familiar with the disease. The core cause of Alzheimer's appears to be a change in the metabolism of a protein, the precursor protein of

beta amyloid (called APP) which, for reasons still unknown, at some point in some people's lives begins to come metabolized in an altered way leading to the formation of a neurotoxic substance – beta amyloid – which slowly accumulates in the brain leading to progressive neuronal death.

It is estimated that, worldwide, in 2015 there were 46.8 million people affected by a form of dementia (in Italy over 1 million and 200 thousand). This figure is set to almost double every 20 years, reaching 74.7 million people in 2030 and 131.5 million in 2050.

The abnormal aggregation of neurofibrils can be decreased by overexpression of Hsp70, Hsp27 and Hsp40 which, as already mentioned, can be triggered by inhibiting Hsp90 [30].

Parkinson's disease is a chronic, slowly progressive neurodegenerative disease that involves various motor, vegetative, behavioral and cognitive functions, with consequences on the quality of life of those who suffer from it.

Parkinson's is the most common of the "movement disorders". It occurs when dopamine production in the brain drops consistently due to the degeneration of neurons in an area called the "substantia nigra" (cell loss is over 60% at the onset of symptoms). From the medulla to the brain, accumulations of a protein called "alpha-synuclein" also begin to appear, which according to some may be responsible for the spread of the disease throughout the brain. The length of the preclinical phase (the time between the initiation of neuronal degeneration and the development of motor symptoms) is unknown, however some studies place it at about 5 years.

The causes of this disease are not yet known, but it seems that several elements contribute to its development. First of all, genetic factors: mutations in some genes are associated with Parkinson's and about 20% of patients have a positive family history for the disease. Exposure to toxic substances such as pesticides, hydrocarbon-solvents and heavy metals (iron, zinc, copper) is also relevant.

These symptoms occur asymmetrically: one side of the body is more affected than the other.

Parkinson's disease is a neurological disease that affects 5 million people worldwide today and occurs on average around the age of 60. This number is expected to increase.

The aggregation of  $\alpha$ -synuclein and the resulting neurotoxicity can be attenuated by the overexpression of Hsp70 stimulated by treatment with GDA. In studies carried out on *Drosophila* neurons, the treatment has shown an effective neuroprotective action [31].

Modern technology and ongoing breakthroughs in the field of molecular biology have also fundamentally altered how a new medicine is designed and identified among a large number of potential chemicals.

The primary motivation for researchers to investigate novel medications is what is known as clinical necessity, or the requirement to identify a chemical capable of treating or preventing a certain condition.

Modern pharmacological research is founded on a thorough understanding of the disease to be treated, the cellular and molecular systems that govern it, and the “target” against which the new medicine must be aimed.

Choosing the best target (a molecule or a real biological mechanism) is difficult because it depends on the disease: a virus or a bacterium for infectious diseases like hepatitis or flu, a lack of a hormone in metabolic diseases like diabetes, or a mechanism that leads to cell degeneration like Alzheimer's.

Once the molecule or mechanism responsible for the disease has been identified, they do not necessarily have the characteristics to become pharmacological targets. In fact, even while it is obvious that targeting a particular molecule has therapeutic benefits, meaning it affects the disease, the target is only optimal if it can attach to other small molecules (future medications) that change the activity.

With the use of methods like X-rays, crystallography, and spectroscopy, these targets may now be thoroughly described and examined, even in their three-dimensional structure, to prepare for the next phase in the creation of the new drug: the discovery of the “compound guide” (lead compound).

The new drug's precursor is the main chemical. A chemical that can bind to the selected pharmacological target and change the activity of that target.

Not all molecules that attach to a target may become the lead chemical; it is crucial that this substance, which can be either natural or synthetic (made in a lab), is highly selective, meaning that it exclusively acts on the target in order to reduce side effects. Additionally, it must be non-toxic, have a high bioavailability, and be feasible for pharmaceutical corporations to manufacture on a big scale.

Random screening and so-called rational drug discovery, also known as drug design, are the two major methods used by researchers to find the lead chemical.

Many natural or synthetic compounds are evaluated in the random screening process, and some of them show the desired qualities. Today, tens or hundreds of thousands of compounds may be generated and evaluated in a matter of months owing to technological advancements, making a procedure that would have previously taken years of effort incredibly quick. On the other hand, in rational drug development, the lead chemical is developed ad hoc by the researcher based on the properties of the target with which he wishes to engage, therefore it is unavoidably a synthetic molecule. If you wish to use this second route, you need to know a lot about the target and its physical and

biological properties, because only by studying the target molecule in depth will you be able to design a complimentary one and model it to achieve the best outcomes.

The computer and accompanying software are frequently required in these drug identification and design procedures.

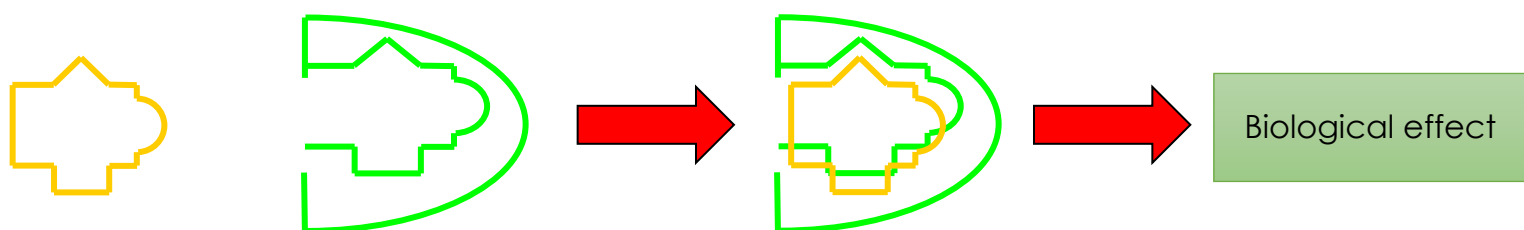


Figure 5: Receptor-ligand interaction scheme.

When the receptor structure is understood, the rational design of bioactive compounds can take place using a direct approach, followed by the use of molecular docking tools. Otherwise, if the receptor structure is unknown, an indirect approach might be used, due to the usage of QSAR algorithms (quantitative structure-activity relationships).

The latter refers to the process of relating chemical-physical qualities to biological activity or chemical reactivity.

A key assumption is that molecules with identical chemical-physical characteristics will have similar activity.

- Identification of the qualities (descriptors) appropriate for the task at hand (e.g., molecular weight, area, volume, dipolar moment, flexibility, ability to form hydrogen bonds, etc.).
- Descriptors are calculated for a collection of chemicals whose biological activity has been empirically determined.
- Calculation of the correlative equation (e.g., By PLS).
- Use the equation to forecast the activity of substances with unknown experimental activity.

Molecular docking, on the other hand, is a computational approach for studying the interactions between a generic ligand and a target biomacromolecule. It is therefore considered necessary to know the three-dimensional structure of both the ligand and the target biomacromolecule.

The purpose of both techniques is to create a model known as a pharmacophore from which new compounds will be built. It is the collection of drug molecule substructures required for receptor engagement.

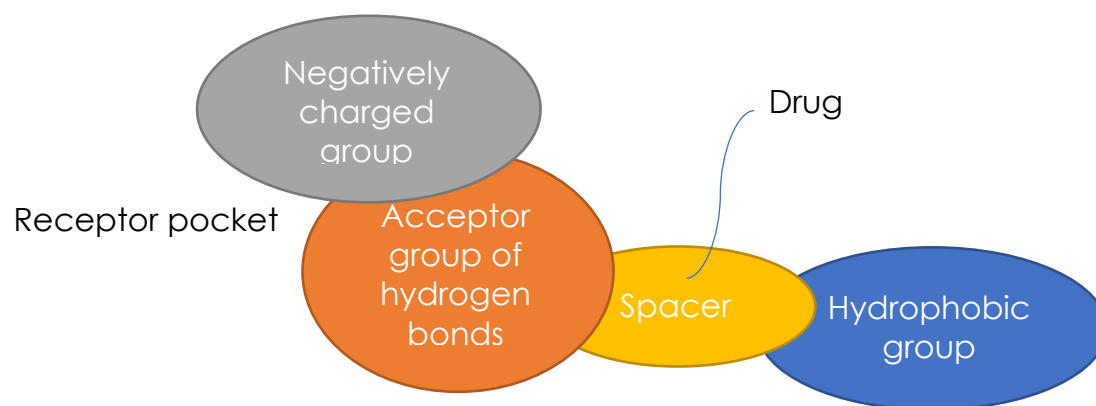


Figure 6: Scheme of a pharmacophore.



## OVERVIEW ON COMPUTATIONAL METHODS

### *Molecular Modeling & Molecular Mechanics*

When faced with a pathology, the challenge is interdisciplinary and involves medicine, biology, and computer science. Bioengineering is the only figure with this transdisciplinary trait. System biology is the study of the interactions and behavior of biological entities such as molecules, cells, and organisms. It is an interdisciplinary discipline that analyzes biological processes in terms of systems and combines models across different inquiry sizes. Structural biology is a subset of it that focuses on the potential structures of biological macromolecules and attempts to link them to their function, i.e., how structural changes in a biological macromolecule impact its function.

These disciplines are vital because we can only gain a thorough knowledge of a biological function if we can integrate all relevant information at different sizes in order to reproduce and comprehend dynamic interactions. Systems biology employs two distinct approaches:

- Reductionist method, from the study of these we deduce the properties of the system.
- The integrative method, as opposed to the reductionist approach, entails examining the qualities that result from the interaction of the components of a system. This method is utilized in complex systems analysis.

A complex system is one whose qualities cannot be determined only by analyzing its components, or one that exhibits emergent features not possessed by the individual elements. An emergent property is a property of a complex system that exists only via the interaction of its constituent pieces.

To examine the interaction between system components, first define them; nevertheless, depending on the objective and biological system under consideration, the scale moves from nanometer to meter. We progress from the microsecond to the gigasecond in terms of the temporal dynamics required to examine a process or phenomena. As a result, a single model for analyzing a biological system is insufficient; a multiscale approach is required that allows the biological system to be analyzed at several sizes, to transition from one model to another, and to comprehend how changes at one level are reflected at different levels. As a result, there are subatomic, electronic, and atomistic models, as well as coarse-grained and continuum approaches. However, because altering the size may cause the physics of events to change, multiscale modeling is frequently used in conjunction with multiphysics modeling. The biological organization is built on a structural hierarchy:



- Atomic
- Molecular
- Macromolecular
- Tissue
- Organs and organisms

Each level is distinguished by events that can be transmitted to other levels due to conformational changes. Modeling of the environment, that is, the environment with which the biological system interacts, is becoming significant, particularly to explore the ways in which it effects metabolic pathways.

### *The Boltzmann Law*

The equation that determines the macroscopic scale of the entropy  $S$  as a function of the multiplicity  $W$  of a system's degrees of freedom is at the core of statistical mechanics. According to this concept, entropy has a maximum that corresponds to the maximum of  $W$ .

$$S = k \ln(W)$$

As a result, entropy becomes a feature of a probability function as well as a physical notion that characterizes the degree of disorder in a system, giving it a broader meaning than statistical thermodynamics.

$$\frac{S}{k} = - \sum_{i=1}^t p_i \ln p_i$$

In the case of constraints, the Boltzmann distribution follows an exponential rule that maximizes entropy. The quantity  $q$  is known as the partition function; it is the total of all Boltzmann factors and counts the number of states that are truly accessible to the system; hence, it offers predictions on the system's attributes, even if it is generally difficult to know and must be sampled. The average energy may thus be represented in terms of the distribution function using the entropy-maximizing probability distribution:

$$\varepsilon = \sum_{i=1}^t \varepsilon_i p_i^* = \frac{1}{q} \sum_{i=1}^t \varepsilon_i e^{-\beta \varepsilon_i}$$

The Boltzmann distribution is a central result of statistical mechanics; it allows us to model and predict the properties of materials based on the structure of the 25

atoms that make them up, and it provides information on how states are arranged according to their energy and probability of being found. Some of the qualities predicted in this manner can be discovered empirically by proving that the model employed is a good reflection of reality.

According to the Boltzmann distribution, more particles will have low energies and fewer particles will have high energies. This is because particles have no preferences for energy levels, yet with this distribution, the system may be configured in a variety of ways. If each particle only holds a small portion of the total energy, the remaining energy can be distributed in a variety of ways by other particles.

Fluids, according to the kinetic theory, are made up of particles with Newtonian behavior, mass  $m$ , velocity  $v$ , and kinetic energy  $\varepsilon$ . Boltzmann's equation says that the likelihood that a particle in a container with constant volume and temperature has the velocity  $v_x$  is termed the Maxwell-Boltzmann distribution, and it allows for extremely precise prediction of the velocity distribution as a function of temperature. It is a distribution in which the average speed is zero because for every particle that moves at a specific speed in a certain direction, another particle moves at the same speed in the opposite direction on average.

The partition function, which determines how particles partition between accessible states, connects macroscopic and thermodynamic aspects to the microscopic model. Because the first conceivable energy state is usually zero, the first term of the sum representing the partition function is unitary. As previously stated, the partition function counts the number of states that are

truly accessible to the system and is the sum of the Boltzmann factors  $e^{-\frac{E_j}{kT}}$ .

At low or high energy levels, all states are accessible and equally populated. When the energy or temperature are increased or decreased, the particles tend to occupy the lower states until they occupy just the ground state, which is the only one that is accessible.

As a result, the amount  $\frac{E_j}{kT}$  decides whether or not a state  $j$  is genuinely accessible:

- At temperature  $T$ , states with energy larger than  $kT$  are comparatively inaccessible and unpopulated.
- States with energy less than  $kT$  are accessible and heavily populated.

The overall number of available states  $\Omega$  does not change and is determined by the system, but the number of states that are truly accessible fluctuates.

The number of microstates in a macroscopic system can reach  $10^{30}$ , and it is frequently impossible to identify one microstate from another because they change so fast. As a result, focusing on macrostates is beneficial. The density of

states  $W$  is the number of microstates in a particular macrostate ( $E$ ). In general, we focus on an energy-level-focused partition function, which has the following form:

$$Q = \sum_{j=1}^t W(E_j) e^{-\frac{E_j}{kT}}$$

Where  $W$  is the multiplicity of the microstate associated with the energy  $E_j$  and  $t$  are the energy levels.

We may also examine a system that is made up of a number of separate subsystems. In general, crystal particles are identifiable, but gas particles are indistinguishable since they do not hold permanent places. The partition function of a system composed of  $N$  independent subsystems is the sum of their partition functions. Independent and distinguishable particles will have  $Q = q^N$ , whereas those independent and indistinguishable will have  $Q = \frac{q^N}{N!}$ , where  $q$  is the partition function of each particle in the system.

The phrase for indistinguishable particle systems, in particular, arises from the overcounting caused by indistinguishability. This is a good estimate, but it is not totally accurate; in reality, there is no overcounting when all particles occupy the same energy level.

However, the likelihood of this scenario occurring is so minimal that the error made is insignificant. Furthermore, because the number of accessible states is typically significantly higher than the number of particles, the correction factor is appropriate.

All of the system's macroscopic features can be determined, except for the partition function, which cannot be calculated but can only be estimated. So far, we've looked at systems with constant temperature, volume, and particle count. In statistical mechanics, these systems are known as canonical ensembles. This is a form of statistical ensemble, which refers to both all the microstates of a system that share the same thermodynamic or macroscopic state (the entire set of configurations) and the usage of constraints. The Isothermal-Isobaric Ensemble (fixed  $T$ ,  $p$ ,  $N$ ), the Microcanonical Ensemble (fixed  $U$ ,  $V$ ,  $N$ ), and the Grand Canonical Ensemble (fixed  $T$ ,  $\mu$ ,  $V$ ) are further statistical ensembles. Because there can be no energy fluctuations in the microcanonical whole, unlike the canonical one, each state accessible to the system must have exactly the same energy, and the probability distribution that leads to a condition of equilibrium optimizes entropy.

A classical technique may be used to represent all atoms except hydrogen without making mistakes since quantum effects can be ignored. As a result, atoms are shown as hard things that may be found. However, hydrogens must be considered for hydrogen bonding. The phrase "molecular mechanics" refers

to the use of Newtonian mechanics to model molecular systems and the use of classical mechanics to find equilibrium molecule structures. In molecular mechanics, the potential energy of all systems is computed using force fields, which are the collection of parameters required to calculate the potential energy of a system and characterize the system's constituents. The molecular mechanics technique entails defining the atomic type, which comprises information on the state of hybridization, mass, and charge of an atom and entirely characterizes it: all information on electrons is lost since an atom is treated as a hard sphere. The geometry of the system is determined using molecular mechanics. Once the particles of the system are established, the potential energy, which is a function of the locations of the N atoms that make up the system, may be described as the total of two contributions, one owing to bonding interactions and one due to non-bonding interactions. Bonding interactions:

- Bond: Interaction between two covalently bonded atoms modeled as a harmonic interaction, or as if the covalent bond were a spring that connects the two atoms.

$$V(l) = \sum_{bonds} \frac{1}{2} k_l (l - l_0)^2$$

where  $k_l$  is the force constant and  $l_0$  is the reference bond length (length assumed when all the other terms of the force field are null), obtained experimentally or from quantum mechanical analyzes. The variable  $l$  is the bond length at equilibrium when all terms are considered. In reality the atoms are not stationary, but the bond length varies around an average value due to vibrations, this leads to slight experimental errors. This harmonic term is good around the energy minimum, it is not good for studying the breaking and formation of covalent bonds (it is very far from the energy minimum). In these cases, other models are used that use approximations of the Schrödinger equation. Another model is the morse, but in general the harmonic model is used. As the bond length decreases, the stiffness of the spring that represents it and the energy of the bond increases, and this means that the vibrations possible due to thermal agitation are of a lower amplitude. Basically, we are giving a penalty to the displacement from the minimum constituted by an increase in potential energy. But the stretching of a bond is not the only factor that can affect the energy of the system.

- Angle: represents the interaction of a three-particle system, two of which are bound in which the angle can often change uncontrollably by increasing or decreasing, for example, following collisions between

molecules. The angle term is then added, which gives a penalty due to the fact that the angle of the system varies in an uncontrolled manner.

$$V(l) = \sum_{\text{angles}} \frac{1}{2} k_{\theta} (\theta - \vartheta_0)^2$$

Again,  $k_{\theta}$  is the force constant,  $\theta_0$  is the reference bond angle (when all other terms are at zero) is the variable  $\theta$  is the bond angle when all other terms are considered. There are several estimates for the angle, but the most common is the harmonic, which is always good near the energy minimum or for tiny fluctuations. The waveform is identical to the first example. Thermal vibrations, like the preceding scenario, cause changes in the bond angle, resulting in experimental mistakes during detection.

- Dihedral angle: rotation in a system of four linked atoms of the fourth atom with respect to the plane identified by the first three. It takes into account the steric hindrance between atoms. Since it is possible that there are symmetries and that therefore there are different positions constituting the same minimum, the potential energy linked to the dihedral angle has a sinusoidal trend, i.e., different dihedral positions are possible.

$$V(l) = \sum_{\text{dihedrals}} k_{\varphi} [1 + \cos(n\varphi - \delta)]$$

$k_{\varphi}$  is the energy cost related to deformation, it is higher for the amide bond than for the C-C bond.  $n$  is the multiplicity of energy minima in a  $360^\circ$  rotation.

- In the case of planar molecules, mobility with regard to the plane of the molecule is feasible, but restricted. This "out-of-plane movement" is known as Improper Dihedral.

The binding terms are not the only ones; in fact, they are the least significant because non-binding interactions are responsible for essential structures such as protein folding. Two particles interact non-bindingly and arrange themselves so that they are at the distance that corresponds to the energy minimum. Each atom interacts with all the atoms around it, but the interactions that are represented are those that do not involve bonding. These interactions are typically described as functions that are inversely proportional to the distance between two atoms.

- Van der Waals interactions: interactions between uncharged atoms at roughly Armstrong distances. At distances greater than a nanometer, the interactions are no longer visible, and they turn repulsive when the atoms are too near and overlap the electronic clouds. Because dispersion interactions are caused by the production of transient dipoles, they are electrostatic subatomic interactions. According to the model, as they overlap approaching, the energy must swiftly move to infinity, while it must slowly decline to zero after the nanoscale. VdW interactions are represented by the Lennard-Jones 12-6 model.

$$V(r) = 4\epsilon\left[\left(\frac{\sigma}{r}\right)^{12} - \left(\frac{\sigma}{r}\right)^6\right]$$

- Hydrogen bonds, which can be modeled as particular Lennard-Jones interactions 10-12.
- Electrostatic interactions are constantly present, even at distances greater than a nanometer, and so have a high computational cost. The defining of the charge is crucial because the atoms are given with partial charges that are optimized based on the other interactions and the force field while respecting the general criteria of total charge of the molecule. Coulomb's law is used to model electrostatic interactions. However, when the number of particles rises, many of these interactions must be deleted in order to lower the computational cost.

$$V(r_1, r_2, \dots, r_N) = \sum_{bonds} \frac{1}{2} k_l (l - l_0)^2 + \sum_{angles} \frac{1}{2} k_\theta (\theta - \theta_0)^2 + \sum_{dihedrals} k_\phi [1 + \cos(n\phi - \delta)] \\ + \sum_{improp\ dihedrals} \frac{1}{2} k_\zeta (\zeta)^2 \\ + \sum_{i=1}^N \sum_{j=i+1}^N \left[ \frac{q_i q_j}{(4\pi\epsilon_0\epsilon_r r_{ij})} + \frac{A(i,j)}{r_{ij}^{12}} - \frac{C(i,j)}{r_{ij}^6} \right]$$

Potential energy is just one of the concepts used to characterize the system; it is far from the sole one. For example, the system's kinetic energy must be used. The boundary conditions are critical in simulations because they have a large impact on the overall characteristics of the system, especially if the particles that interact with the border are numerous in comparison to the total.

One method for removing boundary conditions is to use periodic boundary conditions, which repeat the box that encloses the system in three dimensions of

space. This removes the system's stiff edges, and a moving particle flows into the adjoining box. The number of particles in the original box and duplicates remains constant. The system might therefore be duplicated endlessly. Only some forms, such as the cubic, tetrahedral, and hexagonal prisms, may be repeated endlessly. This approach, for example, allows you to construct a cell membrane without it bending and collapsing owing to water flow at the ends. This method has the advantage of avoiding edge effects and simulating an endless environment. However, it is an artifact, and each particle does not have to see itself (Minimum Image Convention), it becomes expensive if the system is large: in some cases, if not too expensive, it is better to simulate in a single very large box with the molecular system being analyzed in the center, away from the edges. However, because electrostatic interactions are long-range, the computing cost skyrockets even if we ignore the molecules in other boxes. To eliminate this effect, one technique is to use the plain cutoff, which involves making the particle blind from a specific distance, often 1 nm. At the cutoff radius, which must always be less than half the side of the box to meet the minimal image convention constraint, the potential energy curve quickly goes to zero (a good value is normally 1.2 nm). However, by doing so, I severely limit the interactions that a particle may have. Furthermore, at each step, I would have to compute all of the particles that are inside the cutoff radius and then calculate the interactions, which would be time-consuming. As a result, the neighbor list, which is a list of particles that are close to the considered particle, has been introduced (an example is the Verlet NL). It is updated on a regular basis (the precise number of steps is determined by the operator), and the electrostatic interactions with the specified particles are taken into account even if they have moved outside of the cutoff radius in the stages before the update. Instability is an issue with this method: there are significant variations in potential energy near the cutoff radius due to numerical inaccuracies caused by the potential energy function's rapid descent to zero. These fluctuations can cause forces that cause particles to splash and, as a result, a system crash; this is especially true for electrostatic interactions, which are more powerful and so have numerical errors bigger than those of VdW. One method for reducing the number of atoms in the neighbor list and saving computational expense is to identify charge groups and assign the charge of the group to only one atom in the list. Within the force field, these groupings have already formed. To avoid instability, the potential energy function can be altered more gradually at interatomic distances around the nanometer as an alternative to the cutoff. For example, some have pushed the function up to make it go to zero, but this changes the value of the minimum and causes instability: keep in mind that the area of the energy minimum is the most important during the analysis since the particles prefer to fluctuate about there. A third option is to establish a possible multiplier function in the cutoff zone to avoid making crucial errors and slow down the variation in this manner, they must specify two cutoff rays that delimit

the zone where this alteration occurs. You make a mistake with the switch just at the cutoff area, where you have a more gradual decrease to zero with less instability.

In fact, different strategies are utilized to lower the computational cost associated with long-range interactions, but they must be considered:

- Ewald's sum, which is excellent when the boundary conditions are periodic. Because each charged particle interacts with every other charged particle in the box and periodic boxes, the potential owing to these interactions may be expressed as follows:

$$V = \frac{1}{2} \sum_{|n|=1} \sum_{i=1}^N \sum_{j=1}^N \frac{q_i q_j}{4\pi\epsilon_0 |r_{ij} + n|}$$

The resulting series gradually converges. The secret is to divide the convergence into two different series that converge considerably faster, using a distance function that allows you to control the rapid changes in  $\frac{1}{r}$  for small  $r$  and the gradual decay for high  $r$ .

To avoid modifying the system, suppose we screen each charge outside the cutoff radius with a Gaussian distribution with mean in the position of the charge to be screened, equal intensity but opposite sign; at the same time, we add an analogous distribution but of the same sign as the original potential. After the charges are protected, the interaction shifts from long to short range, and it becomes almost meaningless beyond the nanoscale. The second Gaussian distribution may be examined in Fourier space, where it quickly converges. Due to the overlapping of the Gaussian tails, correction factors are required. The potential may thus be described as a sum in the real space of shielded charges and the Fourier space of the second Gaussian distribution, which converges extremely rapidly. This approach is the foundation of the most widely utilized, known as PME. Before using Ewald's sum, ensure that the system's overall charge is zero.

- Reaction field in which a sphere with a radius equal to the cutoff radius is created around the molecule. The short-range interactions, that is, those with the atoms inside the sphere, are precisely estimated. To take into account long-range interactions, the outside of the sphere is represented by a dielectric ( $\epsilon_s$ ), and the external electric field due to the dipoles ( $\mu$ ) inside the sphere is calculated (excluding the atom for which the potential is being calculated); the potential energy converges faster this way.



- Expansion of the multipole, in which space is split into uniform cubic cells and the precise interactions are determined. Furthermore, the multipole moments (charge, dipole, and quadrupole) for each cell are determined in order to calculate the long-range interactions as an atom-cell interaction using the multipole expansion.

There is torque additivity in electrostatic force, which means that I can represent the force on a charge owing to two others as the sum of two components. The same approach may be used for several charges, but it has a significant processing cost; consequently, to calculate the force acting on the charge, one must first calculate the potential energy (in reality it does not work like this due to the screen effect created when there are many atoms interacting with each other). The energy of a multibody system is the sum of the interactions between all conceivable pairs of particles, or the sum of the potentials. When analyzing a charge system in space, electric moments such as the dipole moment and the moment of charge with respect to a point can be studied. If there are  $n$  charges, the moment of first order ( $p_e$ ) determined with respect to a location is the sum of the  $i$ -th charge multiplied by the  $i$ -th distance from that point and may be expressed as a vector. It is also feasible to define the second order moment ( $q_e$ ), which has the dimension of a charge for a distance squared and may be expressed as a symmetric matrix. Because there is no standardization in the definition of electrical moments, it is required to identify the origin of the reference as well as the description of the instant itself. Sometimes, instead of the second order moment, the quadrupole moment ( $\Theta_e$ ) is used, which is expressed in the form of a matrix and has the feature of having diagonal components that sum up to zero, i.e., the matrix that defines it always has zero trace. Electric moments are scale or gauge invariant only if the preceding moment is zero, i.e., they are independent of the reference system: if the dipole moment and distribution charge are both zero, the quadrupole is gauge invariant (octahedral distribution of offices); a system of charges arranged in line and overall neutral does not require the definition of the origin when analyzing the dipole moment. The quadrupole moment quantifies the charge distribution's departures from spherical symmetry: if the distribution possesses spherical symmetry, the quadrupole moment is zero. Yes, they can define other moments of greater order in order to gather knowledge about the system and characterize it more accurately. We are interested in the distribution's lowest non-zero electric moment, which is the dipole moment for many molecules. Each moment may be represented by a charge distribution, for example, a quadrupole moment can be schematized as eight charges: this substantially simplifies the chemical system contained within a cell. But how can we achieve multipolar expansion? The underlying assumption is that the charge-distribution

interaction reduces rapidly in space, and the total potential energy obtained by modeling the distribution with multipoles, containing a lesser number of atoms, is the same as the total potential energy obtained using all charges. As a result, we may express the potential as a sum of components that rely on the electrical moments: Higher-order moments approach 0 quicker as distance increases. At greater distances, the charges, not the times, take precedence, but if I set the bids to be neutral, the potential falls quicker. Furthermore, if I select the grain as neutral, the charge-charge interaction does not occur. In this approach, the error in computing potential energy is modest, but I still receive a distribution to restrict the number of interactions to be computed. The main idea is that the distance between the systems must be substantially higher than the size of the molecules; it is often used to tiny molecules because convergence is slow in any case.

### *Simulating the environment*

Solvents, in general water, play an important role in defining the tertiary structures of proteins and in their dynamic behavior, such as shielding any interactions between molecules. To simulate the influence of the solvent, explicit solvents can be employed, which are easy to model, exact in taking the effect of the solvent into account but computationally expensive, or implicit solvents, which are complicated to implement and whose correctness is still debatable. Water is modeled as a V-shaped molecule with  $sp^3$  hybridization in which two electron pairs are unpaired and two are bonded to hydrogen atoms; it has the volume of a 1.5 Armstrong sphere. The bond length between oxygen and hydrogen in water was determined experimentally and by quantum mechanical analyses to be roughly 0.97-0.99 Armstrong, with a bond angle of approximately  $106^\circ$ . These properties fluctuate rapidly due to polarization events or hydrogen bonds formed by the molecule with its surroundings.

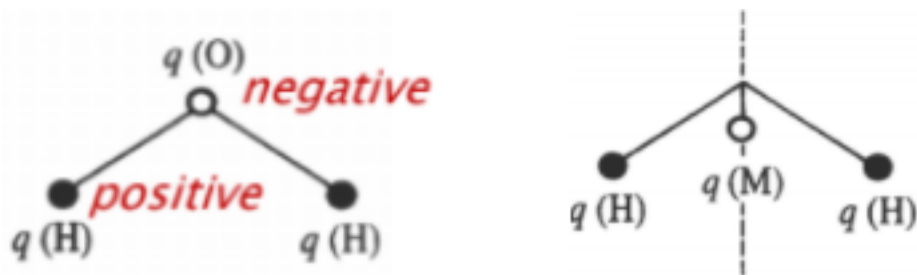


Figure 7: V-shape model of the water.

Water is considered as a medium and filler in explicit models, characterized by its geometry and the parameters of the force fields (e.g., Lennard-Jones), however many characteristics cannot be reproduced. The three-site model

(Berendsen's SPC and Jorgensen's TIP3P), with negatively charged oxygen and two positively charged hydrogens, is the most commonly employed; in this model, VdW interactions are computed exclusively with oxygen, and the dipole moment is underestimated. It is also feasible to utilize 4-site models, in which one of the particles must have no mass, or even 5 charges, which is heavier at a computational level to better estimate the dipole moment, to simulate certain features, particularly the dipole moment. The water models used in the simulations communicate with the force fields, which means that a water force field compatible with that used for the protein must be utilized, and which ones to use are determined by reviewing the literature. Water is treated as a rigid molecule in the simplest models, and only non-bonding interactions between oxygen atoms, such as Coulomb and Lennard-Jones interactions, are investigated. Unfortunately, when water molecules come into contact with other molecules, they become polarized, which should be accounted for in the simulation; however, highly computationally heavy models are necessary to do so; this polarization can be entered, for example, into the potential energy function. When a molecule is solved, the number of particles inside the simulation box grows significantly; consequently, a box that is not too large to simulate merely water and save computing cost, but also not too tiny to mimic immersion in a means, must be chosen. You may then mix in additional ions with the water, such as a physiological concentration of NaCl (0.15 M), which can sometimes drastically alter the simulation's outcome. To avoid mistakes when using Ewald's summation, it is best to maintain bulk neutrality, however this might still imply that there is a non-zero charge locally; However, if our protein has a protonation state, it is important to adjust for it, or to ensure that fluctuations in the charge inside the box are balanced by the entry of ions into the environment, as would occur in a physiological setting. Water has its own deformability and thus constraints are sometimes used to stiffen it.

The explicit models can mimic the behavior of biomolecules in the biological environment in a realistic manner, but they come at a high computing cost, and by making the molecules stiff, there is an extra cost due to constraints. As a result, implicit solvation models have been devised, in which the solvent is represented as a continuous medium with electrostatic, entropic, and viscosity properties that mimic those of the solvent. These approaches are mostly used to quantify free energy in solute-solvent interactions in structural and chemical processes such as protein folding and drug transport through biological membranes. Two models are commonly used:

- Surface areas accessible to solvent (SASA) models
- Continuous electrostatic models

It is possible to adjust and adapt the various methodologies, which, in general, allow for considerable reductions in processing costs and simulation speed, as well as statistical mistakes caused by insufficient modeling of the water structure. Although they are effective for modelling the behavior of biomolecules (for example, protein folding), they are still an approximation and have issues with parameterization and ionization effects.

Solvation free energy is defined as the energy needed to move a solute from a gas to a state immersed in a solvent, which is not only related to potential energy but also depends on the conformations that both the solute and the solvent can assume, with less possibility of organizing itself spatially by varying its entropic component. In general, it is the sum of the free energy change caused by the solute's transition from apolar to polar form ( $\Delta G_{elec}$ ) and the free energy change produced by the solute's entrance into the solvent ( $\Delta G_{np}$ ). The solvation process is hampered by the breaking of hydrogen bonds between water molecules as well as a decrease in entropy due to the smaller space accessible to water; however, it is aided by interactions between the solute and the solvent, even if the solute is non-polar and these interactions are less energetic than those between polar molecules. A non-polar solute's free energy of solvation is proportional to its solvent accessible surface area (SASA):

$$\Delta G_{np}(X) = \gamma A(X)$$

Where  $\gamma$  it is a parameter related to the surface tension.

The SASA is computed using a simulation in which a solvent molecule (probe) of a specific radius is forced to slide on the molecule: the SASA is the location of the spots occupied by the center of the probe.

SASA is commonly used to investigate the hydrophobic effect, which occurs when hydrophobic molecules in water approach and exclude water and is crucial in protein folding. These approaches, however, do not allow for the analysis of specific distance-dependent interactions.

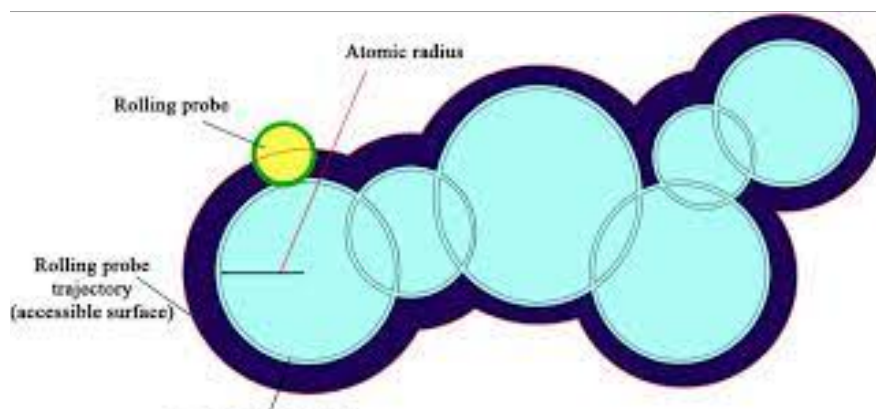


Figure 8: Solvent accessible surface area (SASA).

The GBSA model, which is based on SASA, is one of the most commonly used implicit models. In any event, the difficulty in getting parameters that describe water using implicit techniques implies that the free energy function produced differs significantly from that obtained using explicit methods, particularly in areas of energy minimums, which are thus of fundamental concern to our goals.

### *Energy minimization*

Before analyzing the dynamics of a system, the first step is to minimize its potential energy: the potential energy surface (PES - potential energy surface) is a multidimensional function of the system coordinates; if it consists of  $N$  atoms, there will be  $3N$  Cartesian coordinates and  $3N-6$  internal coordinates (bond lengths, bond angle, torsion angle), and the potential energy will assume a certain value depending on the value they assume. We are interested in relaxing the system and bringing it to at least a local minimum; if I did not do this, extremely strong forces may be produced that destabilize the system at the start of the dynamic analysis.

To identify the minimum, derivative or non-derivative approaches are utilized, beginning with knowledge of the expression of potential energy and initial coordinates, but in general, the coordinates of the system are adjusted at each step to reach a lower energy configuration. In general, the global minimum is not reached from the initial configuration; however, it is not certain that the active configuration of the molecule is the one found in the global minimum or the local minimum closest to the starting position: if there are points where the energy potential drops very steeply towards a global minimum, it is not clear that that location is the most populous because there may be local minima where the energy difference between surrounding places is significantly lower and hence more multiplicity exists.

- SIMPLEX non-derivative method, it is based on the development of a geometric shape with  $N + 1$  linked vertex, where  $N$  denotes the potential energy's dimensionality, and each vertex refers to a set of coordinates for which the potential energy may be determined.

If I have three dimensions, I'll use a tetrahedron and change the location of the vertex with the greatest energy (reflection, reflection and expansion, reflection and contraction) at each step, acquiring the next vertex until I find one with a low enough energy. It works well in the early stages, when we are far from the minimum, but it is inefficient in its vicinity.

The first order derivative of the potential energy function (i.e., the gradient) is utilized in derivative techniques because it reveals the direction of maximal variation of the function towards the maximum and its modulus informs me how steeply it changes. The second derivative informs me if the function is concave

or convex, which indicates where the function will change direction or whether we are at a stationary point.

As a result, the energy of the system may be reduced by moving each atom in accordance with the total force acting on it.

- Steepest Descend method, it is a first order approach that includes travelling in a direction parallel to the particle's total force. At this stage, you must select how far to advance in the specified direction: you may either execute a line search in the direction of the gradient or take arbitrarily lengthy steps.

The first stage of the line search is to select three places along the gradient line where the potential energy of the central point is smaller than the potential energy of the other two points. At this point, a parabola is used to interpolate the three points, and the point with the highest energy is substituted with the energy estimated in the minimum of the parabola. The procedure is repeated at this point: the gradient in the minimum of the line search will be perpendicular to the preceding direction, implying that the search is carried out in successive orthogonal directions. The amount to be moved is normally chosen, said step size, which is dynamically varied in the minimization: normally, the step size is increased until the minimum is exceeded, at which point it is progressively decreased; in this way, speed and precision in energy minimization are optimized simultaneously. This approach may require more steps for minimization, but it frequently involves fewer function evaluations and hence a reduced computing cost. Because the gradient's direction is governed by the strongest interatomic forces, it is an excellent tool for locating the areas of greatest energy in an initial configuration.

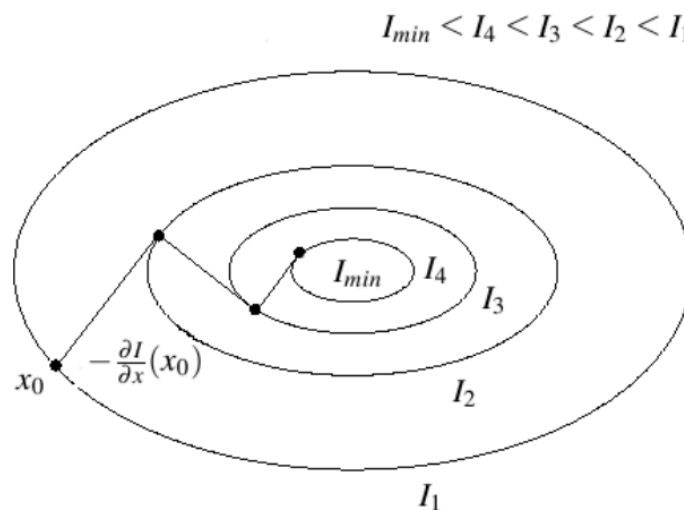


Figure 9: The steepest-descent direction requires calculation of the gradient.<sup>38</sup>

- To optimize minimization, the conjugate gradient approach overcomes the limitation of movement in orthogonal directions: in this situation, the gradients for successive steps are orthogonal while the directions in which we travel are conjugated. This indicates that the minimum of a quadratic function with M variables is obtained in at most M steps. It is also available as a second order approach, but it involves the inversion of the Hessian matrix, which is typically difficult and expensive to compute. As a result, it is employed in systems with less than a hundred atoms. The approach to be used must be determined by the type of system to be optimized for minimization.

For example, while we are far from the minimum, we can use the steepest descent technique, but when we go closer, we may utilize the conjugate gradient method. In certain circumstances, knowledge on energy reduction is sufficient to properly anticipate particular system features (all minimum energy configurations must however have been identified).

### *Molecular dynamics*

Molecular Dynamics (DM) is a technique in computational chemistry that allows you to simulate the motion of individual atoms in atomic or molecular systems. There are several methods for describing a system of particles that use Cartesian coordinates ( $r_3, v_3$ ) or even non-Cartesian coordinates (generalized coordinates  $q_3$  and generalized moments  $p_3$ ), such as the methods of Lagrange and Hamilton, which are designed for the dynamic treatment of particle systems.

Lagrange proposed defining the Lagrangian operator, which is valid for conservative systems, as the difference between kinetic and potential energy. The standard equation of motion in these words is:

$$\frac{d}{dt} \left( \frac{dL}{dv_i} \right) - \frac{dL}{dq_i} = 0$$

This approach has the benefit of being able to construct second order equations of motion in systems with non-Cartesian coordinate sets. Given the starting circumstances, the solution of the equation of motion is the system's trajectory, which may be described as the position and speed that the system's components adopt at succeeding instants. Hamilton modified the Lagrange equation to describe the system in terms of locations and momentum, which is commonly described as the Lagrangian derivative with respect to velocity for each i-th particle:

$$p_i = \frac{\partial L}{\partial v_i}$$

$$K = \sum_{i=1}^N \frac{p_i^2}{2m_i}$$

The Hamiltonian is therefore defined as the sum of kinetic and potential energy, where the coordinates are the positions and moments:

$$H(q, p) = K + U = K = \sum_{i=1}^N \frac{p_i^2}{2m_i} + U(q)$$

Compared to before, the system has been transformed in first order because we can write the equations of motion as:

$$\frac{dq_i}{dt} = \frac{\partial H}{\partial p_i} \quad \frac{dp_i}{dt} = -\frac{\partial H}{\partial q_i}$$

As with Lagrange's approach, the equation of motion in each set of coordinates that is equivalent to the standard Newton equation may be obtained. To see the formulation's trajectory, it can be reported in phase space (position-moments). A differential equation of the second order is transformed into two first order equations in  $6N$  coordinates. If I have a system of particles, I may create a Cartesian space with  $6N$  orthogonal dimensions, and the sequence of points in this system represents the molecule's route.

At this point, we can define the  $6N$ -dimensional vector  $x$  that defines the position in the state space, and because energy conservation holds, the Hamiltonian as a function of  $x(t)$  remains constant over time: this corresponds to defining a  $6N-1$  dimensional hypersurface in which the trajectory must remain. Positions and moments characterize the system's state in phase space.

A thermodynamic ensemble is a group of points in phase space that meet the requirements of a certain thermodynamic state. It is feasible to connect a macroscopic attribute of the ensemble  $A(p^N, q^N)$  to a microscopic one by using the probability density function  $\rho(p^N, q^N)$  of the considered set, where  $Q$  is the partition function:

$$A_{ensemble} = \iint dp^N dq^N A(p^N, q^N) \rho(p^N, q^N)$$

$$\rho(p^N, q^N) = \frac{e^{-\frac{H(p^N, q^N)}{k_B T}}}{Q}$$



$$Q = \iiint dp^N dq^N e^{-\frac{H(p^N, q^N)}{k_B T}}$$

The partition function is significant because it connects the microscopic thermodynamic variables that cannot be measured to the macroscopic state function that can be measured; as a result, the partition function is a comprehensive thermodynamic description of the system. Normally, calculating the integral is complex since all potential states must be sampled. Another approach is to run an M-step simulation to compute the temporal average of the observable of interest:

$$A_{time} = \lim_{\tau \rightarrow \infty} \frac{1}{\tau} \int_{t=0}^{\tau} A(p^N(t), q^N(t)) dt \approx \frac{1}{M} \sum_{t=1}^M A(p^N, q^N)$$

The ergodicity hypothesis is therefore utilized, in which the average on the ensemble and the average time coincide, but which is only valid if the system is allowed to evolve for a long enough period.

Molecular Dynamics (MD) is a theoretical and computational technique for estimating the average attributes of a system by sampling the microstates of a given ensemble sequentially through time. It is a deterministic approach, which implies that the same results may be acquired by repeatedly running the same simulation. The essential concept is to solve Newton's equation of system motion, for which the potential energy is known owing to an appropriate force field. It's utilized to figure out the equilibrium or transport features of organic or inorganic systems.

The MD is based on Newton's second equation, which, when the force is expressed as the inverse of the gradient of the potential, becomes:

$$-\frac{dV}{dr_i} = m_i \frac{d^2 r_i}{dt^2}$$

To compute the trajectory, the beginning conditions must be supplied: the position ones are provided via the input file, while the velocity ones are generally calculated starting from a Maxwell-Boltzmann distribution at system temperature. Because potential energy is a function of the locations of all the atoms in the system, the equations of motion cannot be solved analytically; they must be computed numerically using an algorithm that fits the following criteria:

- It must conserve the energy and the moment.
- It must be computationally efficient.
- It must allow integration for a long time (of the order of tens of

nanoseconds).

- They assume that position, velocity, and acceleration can be approximated by their Taylor expansion.

There are several approaches for integrating Newton's equations of motion in physics simulations.

The velocity is initially computed at  $t + \frac{\delta t}{2}$  and then utilized to determine the location at  $t + \delta t$  in the Leap Frog technique. In this manner, the speed exceeds the location and vice versa. The benefit of this technique is that the velocities are computed directly, but not at the same time as the locations. Instead, the Position Verlet approach calculates the positions of time  $t + \delta t$  using positions and accelerations at time  $t$  and positions at time  $t - \delta t$ . It has a little storage demand and is straightforward; nonetheless, it is inaccurate and does not explicitly compute the speed.

Finally, the Velocity Verlet is a version of the preceding one that requires just the position, speed, and acceleration that correspond to the same time step to be stored.

The time step must be set so that it is less than the system's lowest period of oscillation, which is generally less than a tenth of the fastest harmonic oscillator present. If it is too little, the simulation and sampling of the state space take too long; if it is too vast, there is no chance of not considering some interactions, bringing some particles too near to collide or produce significant repulsion forces, destabilizing the system.

The time-step is often set at 1-2 fs; however, this depends on the sort of system under consideration. Higher vibrational frequencies are related to hydrogens because of their lower masses, and covalent bonds because of their greater rigidity; thus, to increase the time step, the mass of hydrogens or bonds in the bonds can be increased, i.e., it establishes a distance around which the covalent bond can oscillate (bond terms are eliminated which originate fast oscillations as they are very rigid). In molecular mechanics, there is no time, just step numbers; time occurs only in molecular dynamics in the form of speed.

To integrate the equations of motion, initial velocity conditions must be provided, and in addition to taking them from a Boltzmann distribution at the desired temperature, they can be assigned to a low temperature and then start a heating phase of the system by performing a weak type thermal coupling; the simulation continues by binding the protein until the temperature considered is reached and then continuing a dynamics of the entire system (without restrictions). In any event, the NVT or NPT settings at constant temperature best mimic actual systems. It is required to include the solution equations (integrators) of the other equations that regulate temperature and pressure, which we shall refer to as thermostats and barostats.

Kinetic energy is employed to adjust the temperature of the system. You can modify it using a velocity shift factor  $\lambda = \sqrt{\frac{T_{new}}{T(t)}}$ , but this does not enable us to sample the state space well: in fact, it has been demonstrated that it is important to restrict the direct speed fluctuations in order to make modest mistakes. As a result, keeping the system linked to an external bath at a constant temperature equal to the desired one is an option. In this manner, the system grows in temperature toward that of the bath using a diminishing exponential of time constant  $\tau$ . The smaller it is, the faster the temperature change is, and the coupling is strong, allowing the temperature to quickly converge to the target temperature; if  $\tau = \delta t$ , the weak coupling corresponds with the speed scaling process.  $\frac{\delta t}{\tau}$  is normally 0.0025.

It is feasible to connect two subsystems in a different way, with two separate bathrooms rather than a single bathroom. It is possible that it has a very little protein in comparison to the amount of water in which it is housed. Because the solvent dominates the speed distribution, there is the phenomena of hot solvent and cool solute, or the coupling of the single system may not regulate the temperature of the solute. In this instance, the two subsystems must be coupled independently.

There are weak couplings with pressure that are comparable to and different from those with temperature. Pressure is more related to particle positions than particle velocities and fluctuates much more than temperature even between non-physical values; in particular, it is linked to the virial, which is the product of the positions and the derivative of the potential energy with respect to them, a quantity that varies much more rapidly with position with respect to internal energy: You must just look at the average value of the pressure while looking at the oscillations in the volume of the box to determine if it is stabilizing (In an isothermal-isobaric system, pressure remains constant with volume fluctuation) or the density of the water until it reaches  $1000 \frac{kg}{m^3}$ , it makes no sense to look at the pressure punctually since it is a feature of the entire.

A macroscopic system maintains its pressure by shifting its volume, which occurs in the simulation when the volume of the box varies. The amount of fluctuation is proportional to the medium's isothermal compressibility:

$$\kappa = \frac{1}{V} \left( \frac{\partial V}{\partial P} \right) \tau$$

The isothermal compressibility of the medium indicates how much the material reacts to pressure fluctuation and hence how much it will fluctuate: the more incompressible the substance, the more rigid it is, i.e., it will fluctuate more but change less on average. Heating the system from 0K to 300K takes on the order of picoseconds because, despite the weak coupling, the speed variation is so significant that it leads me to have a system in temperature in maximum 5 ps: 43

events that cannot be controlled can occur, which can be a problem because protein folding lasts on the order of microseconds. As a result, position restraints are created that, like a spring, confine each atom to its beginning position and oscillate around a reference position to obtain a probable initial distribution of velocity at the required temperature.

This prevents unmanageable destructuring during the initial period of system heating. The final velocities are basically the same in terms of distribution as the initial velocities, but the orientation of the velocity vector for each atom is no longer random at the conclusion of the heating.

Instantaneous and macroscopic attributes, that is, the averages of the properties of the microstates that populate the studied macrostate, may be determined using molecular dynamics analysis.

The average temperature is one of them.

The RMSD is a summation on the particles, a function of simulation 'time' that offers information on how much the structure deviates from its conformation as a reference, and an index that identifies a conformational change over time with a number. In general, the one mentioned before has a pattern that begins at zero, climbs abruptly, varies greatly, and finally stabilizes, stably or not. It is not feasible to declare if two locations at the same height have the same structure a priori, but simply that they are equally far from the start.

To determine whether the structure has attained convergence, compute it with regard to the conclusion of the simulation: if it approaches zero (roughly) before the last simulation moment, the structure has reached convergence.

The RMSF is a time-averaged sum that expresses the quadratic distance averaged across frames from the reference conformation. It emphasizes the protein's mobile parts, which are the variations of the molecule's components (for example of the various residues).

The radius of gyration is a quadratic mean of the difference in position between each atom and the structure's center of mass at a given moment. The mean radius of the molecule if it were a sphere, which is used to assess molecular size but only makes sense if the protein has a globular form. It, for example, allows you to visualize how proteins change size when active.

### *Free energy*

The most essential quantity in thermodynamics is free energy, which is generally written as Helmholtz free energy in NVT sets or Gibbs free energy in NPT settings. Molecular dynamics samples the space of states using deterministic methods by solving Newton's equation, but it does not allow us to calculate the free energy because it does not provide an adequate sampling of all the important configurations for this calculation: we never know how effectively we will explore the space of states since the energy curve generally includes numerous

minima and, depending on the existence of the energy gap to get out of the minimum, only one gets examined, the one closest to the original configuration. Free energy and entropy are hard to determine using finite simulations since they need knowledge of the system's partition function or sampling the state space entirely. However, we are more interested in free energy differences, which may be described as a function of a partition function relationship. Assume we have two states, X and Y, and we wish to know the difference in free energy as Helmholtz free energy (connected to the partition function Q) between the two:

$$\Delta A = A_y - A_x = -k_B T \ln \left( \frac{Q_y}{Q_x} \right) = k_B T \ln \left\langle e^{-\frac{H_y - H_x}{k_B T}} \right\rangle_x$$

Rewriting to remove the partition function. The difference in free energy may be stated as an average of the ensemble configurations corresponding to the initial state (X). However, if the two states X and Y are not superimposed in space, the calculation of the free energy difference will be inaccurate because we will not be sampling the state space of Y adequately when we simulate X: there is a large approximation error in not calculating the approach path from X to Y. To address this, you may inject an intermediate state, or rather n intermediate states that are small enough to stabilize the reaction and overlaid to decrease this inaccuracy; in essence, I invent states that do not exist (non-physical) and connect the two routes without any unknown points from X to Y. In practice, we design a parameter that alters the energy equation potential such that when it is zero, the state X is simulated and when it is unitary, the state Y is simulated; I then multiply this parameter by all the parameters of the equation that represents the free energy. This is equivalent to doing non-physical simulations on intermediate states. These approaches are known as thermodynamic perturbation, and they may be implemented as follows:

- Thermodynamic integration, in which n simulations are performed with  $\lambda$  variable from 0 to 1. At the output the software tells me how the Hamiltonian varies as a function of lambda and the free energy difference between the two states is equal to the integral of the curve in exit, that is the area underneath.
- Slow growth, which is used much less because it is unstable; makes  $\lambda$  vary within the same simulation by returning the free energy variation from the initial state as a function of  $\lambda$ . This method is in series while the thermodynamic integration is in parallel, i.e., it can be optimizing by running simulations in parallel on multiple computers.

These approaches are often employed for ligand-receptor affinity, in which the non-covalent interaction between molecules is represented. They are referred to as "old-school" since they do not allow for visualization of the route. In this process, free energy is related to the equilibrium constant according to the relationship:

$$\Delta G = -RT \ln K$$

Where T is the Kelvin temperature and R is the universal gas constant ( $R = 8.314 \frac{J}{mol \cdot K}$ ). The law of mass action can be stated as follows:

$$K_d = \frac{[ligand] \cdot [protein]}{[ligand - protein complex]}$$

As a result, the lower the value of  $K_d$ , the stronger the ligand-protein interaction. The difference between enthalpic and entropic contributions is denoted by  $\Delta G$  so, because entropic contributions are related to degrees of freedom, a larger negative value of  $\Delta G$  suggests a stronger binding affinity because a high entropy equilibrium is preferred.

Consider the interaction of two ligands  $L_1$  and  $L_2$  with a receptor R: the relative bond affinity between the two ligands may be calculated as  $\Delta G = \Delta G_2 - \Delta G_1$ .

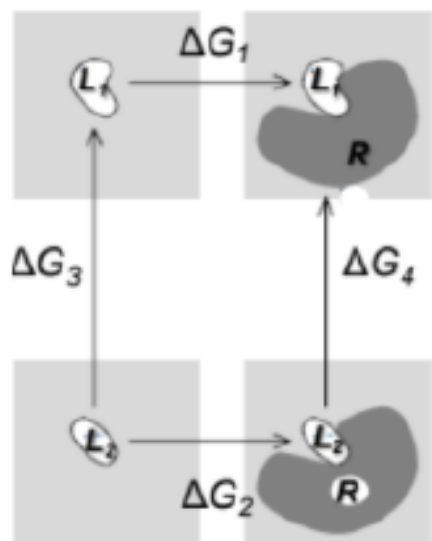


Figure 10: Old-school methods scheme.

In theory, the values  $G_2$  and  $G_1$  might be computed by modeling the ligand-receptor association process; however, in many circumstances, this would result in rearrangement of the receptor, ligand, and solvent, making proper phase space sampling problematic. Because free energy is a state function, its variation must be zero in a thermodynamic cycle:

$$\Delta G_2 - \Delta G_1 = \Delta G_4 - \Delta G_3$$

where  $\Delta G_3$  denotes the free energy difference between the two ligands in solution and  $\Delta G_4$  the difference between the two intermolecular complexes. These free energy fluctuations are independent of structural changes and may be assessed using experimental approaches such as *in silico*. As a result, the difference in free energy may be determined more easily by modelling non-physical paths, especially if the two ligands have comparable structures: this method is known as thermodynamic cycle perturbation.

### *Molecular docking*

Protein-protein, enzyme-substrate, protein-nucleic acid, drug-protein, and nucleic acid-drug interactions all play important roles in many important biological processes, including signal transduction, cellular transport and regulation, and cell control. gene expression, enzyme inhibition, and multi-domain protein assembly These interactions frequently result in the formation of stable protein-protein or protein-ligand complexes, which are required for biological functions. The tertiary structure of proteins is required to comprehend the methods of interaction and molecule affinity. However, obtaining complex structures using experimental methods such as X-ray crystallography or nuclear magnetic resonance is difficult and expensive (NMR).

To reduce these costs and save time, software that can highlight the aforementioned interactions has been developed over time. These are molecular docking capable.

Molecular docking is a computer simulation procedure that is widely used to predict the conformation of a receptor-ligand complex, where the receptor is typically a protein, or a nucleic acid molecule and the ligand is typically a small molecule or another protein.

Structure-based virtual screening is a method commonly used to find new compounds for a protein target of particular interest to the researcher, which can estimate the probability that a compound will bind to the protein with greater affinity with the help of docking and various scoring functions.

Unfortunately, this technique has a limitation in that different poses can produce similar docking-scores: this makes it impossible to distinguish between correct and incorrect poses and thus eliminate false positives from the study.

Consensus docking attempts to overcome this limitation by reducing scoring function errors and combining the results of different scoring algorithms in a consensus scheme, allowing for the prediction of compound interactions in a specific target using more than one scoring algorithm.

The agreement is based on the hypothesis that the quality of a pose resulting from docking calculations, defined as correspondence with experimental data, increases as the number of methods that predict it increases.

To begin docking, select the target you want to investigate; then, the three-dimensional structure of the protein is investigated using techniques such as X-ray crystallography. This method involves the isolation, purification, and crystallization of proteins by measuring the electron density inside the crystal, which allows the atomic positions in three dimensions to be deduced. The results are saved in a database ([www.rcsb.org](http://www.rcsb.org)) that contains a large number of protein structures. NMR (nuclear magnetic resonance) is another technique used; in this case, the result is a 2D structure. Aside from the 3D structure, the binding site and the ligand interactions within it must be known. Docking can be stiff, which means that both the ligand and the protein are rigid; only rotation and translation are evaluated, omitting any sort of flexibility. The flexible docking corresponds to the ligand's flexibility inside the receptor, which cannot be overlooked. Docking can be accomplished using one of two methods:

- Docking is done manually by the operator. The hypothesis about the ligand's mode of interaction must be postulated in this manner.
- Automatic docking: does not involve human abilities on the part of the operator; this computation is delegated to software, which positions the ligand inside the receptor automatically. The validity of the orientation is assessed using the score, which is a numerical number that helps you to determine whether or not an orientation is interesting. It is further subdivided into two major groups:
  - Exhaustive Docking: the results show all of the layout's conformations.
  - Stochastic Docking: In the findings, only certain conformations will be assessed at random. The findings acquired are unique and cannot be replicated.

Once the docking computation is complete, the quality of the postures generated, or the best poses may be evaluated. This depends on whether you want to attain "greatest hits" or the best conformation. Each pose has a binding score, given as Gibbs Free Energy  $\Delta G$ , which may be used to rank and prioritize the poses. It should be remembered, however, that the  $\Delta G$  supplied by docking software is only a scoring function and can only provide an indication of how favorable a position is. Docking can provide some ideas that should be tested in silico with other methods and empirically.

### *Docking simulation*

Simulating the docking procedure is a significantly more difficult task. The protein and the ligand are physically separated in this manner. After a given number of



"moves" in its conformational space, the ligand finds its place in the active site of the protein. The motions include rigid body transformations like translations and rotations, as well as internal ligand structure modifications like rotation and torsion. Because each motion in the ligand's conformational space incurs a total energy cost for the system, the total energy of the system is recalculated after each move. The initial condition for docking is the structure of the protein. X-ray crystallography or, more rarely, NMR of proteins are employed to determine the structure of proteins. The protein structure and a database of possible ligands are fed into the algorithm. Docking success is determined by two factors: the search algorithm and the scoring function.

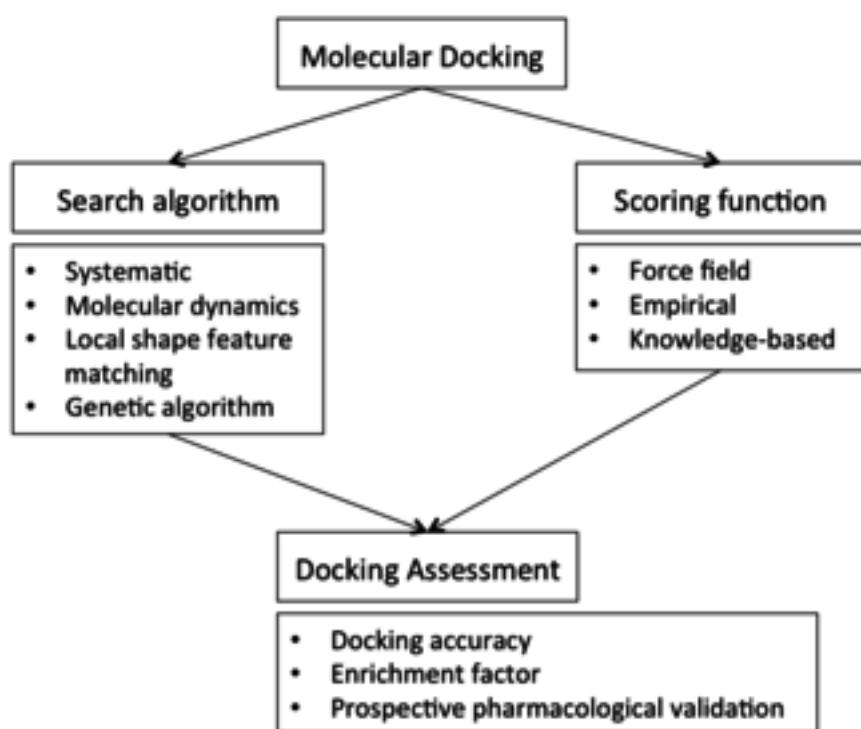


Figure 11: Block diagram of molecular docking.

### *Search algorithm*

In theory, the research space includes every potential orientation and conformation of the protein associated to the ligand. However, with current computational resources, it is impossible to exhaustively explore all of the research space, including all possible molecule distortions (molecules are dynamic and can have a set of conformational states) and all possible ligand orientations relative to the protein to a given level of granularity. The majority of docking algorithms in use incorporate a flexible ligand and some attempts to

bind to a protein's flexible receptor. Each "snapshot" of the couple is referred to as a posture.

There are several search techniques for the ligand and the receptor, which include torsion research on rotatable boundaries, either systematic or stochastic. The novel low-energy conformations are "developed" using molecular dynamics simulations and genetic algorithms.

### *Flexibility of the ligand*

The conformations of the ligand can be created in the absence of the receptor, to which docking will then be applied [32], or they can be formed on the fly in the presence of the receptor [33], or with a rotational flexibility of each dihedral angle utilizing fragment-based docking [34]. Force field energy evaluation is frequently employed to choose energy-efficient conformations, although knowledge-based approaches can also be applied [35-36].

### *Flexibility of the receptor*

In recent years, computer capability has risen dramatically, allowing for the application of more advanced approaches and computationally expensive processes. However, receptor flexibility remains a difficult issue. The primary cause of this difficulty is the huge number of degrees of freedom that must be considered in this sort of computation. Ignoring them, on the other hand, yields minimal outcomes [37].

Many experimentally established static structures of the same protein in various conformations are employed to simulate the receptor's flexibility [38]. To construct an acceptable protein energy conformation, information about the protein structure of the amino acid chains around the binding cavity is needed [39-40].

### *Score function*

The majority of the scoring functions are based on the molecular mechanics force field, which evaluates a pose's energy; a low (negative) energy suggests a stable system and hence a probable binding contact. A different technique would be to generate statistical potential for interactions from a huge database of protein-ligand complexes, such as the Protein Data Bank, and then analyze the results.

A vast number of protein structures and high affinity ligands have been resolved thanks to X-ray crystallography, however low affinity ligands remain elusive due to the fact that they are less stable and hence more difficult to crystallize. The score functions derived from this data may accurately replicate high affinity ligands while also providing believable docking for non-binding ligands. This

results in a huge number of false positives, such as a ligand binding to a protein, which does not occur during the experimental phase.

To reduce the number of false positives, recalculate the energy of the best postures using (possibly) more accurate, but computationally more intensive approaches like the Generalized Born or Poisson-Boltzmann methods. [41]

### *ADMET properties and PK*

Anyone working in drug research understands how important in vitro Absorption, Distribution, Metabolism, Excretion and Toxicity (ADMET) and in vivo pharmacokinetic (PK) investigations are for identifying candidate compounds for the pre-clinical and clinical stages. Indeed, it is the ADMET features that enable people involved in pharmaceutical development to comprehend the true potential of a molecule to become a medicine via efficacy and toxicological studies. As soon as feasible, it is therefore recommended to add those chemical-physical features that will ensure a favorable ADMET/PK profile in the future. The analysis of collected data and the development of computational approaches are providing significant impetus to in-silico ADMET/PK, resulting in a significant improvement in drug discovery procedures. Although the work of Hansch, Iwasa, and Fujita [42] at Pomona College and Chou and Jurs [43] at the University of Pennsylvania, which led to the development of an algorithm for calculating logP (CLOGP), can certainly be considered one of the first computational methods that could be used to establish a relationship between the physicochemical and ADMET properties of molecules. Lipinski [44] will then use ClogP and other simple chemical-physical characteristics to establish the well-known "rule of five." Lipinski's research, based on a database of phase II clinical candidates, claims that when the ClogP is 5, the molecular weight is 500 Dalton, the number of hydrogen bond donors (OH and NH) is < 5, and the number of hydrogen bond acceptors (O and N) is < 10, better absorption and permeability are more likely. Lipinski's rule has become one of the primary metrics of "drug-likeness" due to its speed and ease of computation. Subsequent analyses revealed that the chemical space of produced compounds differs depending on the historical time, therapeutic region, and kind of biological targets, but also on elements inherent in research laboratory culture. To attempt to quantify what Leeson and Springthorpe [45] noticed regarding the danger associated with the candidacy of compounds with ClogP more than 4.5-5, the notion of Lipophilicity Ligand Efficiency (LLE) was proposed, where  $LLE = pIC_{50}$  (or  $pK_i$ ) - ClogP. (or ClogD). According to Leeson's research, clinical candidates should have an LLE greater than 5, which means that for a molecule with a potency of 10 nM, the highest ClogP value should be < 3. Several writers later validated the relevance of lipophilicity in connection to good ADME qualities by introducing other "rules." Young et al. [46-47] presented two metrics in 2010 and 2011: SFI (Solubility Forecast Index) and PFJ

(Property Forecast Index), with  $SFI = ClogP + (\text{Number of aromatic rings})$  and  $PFI = \text{chromLogD}_{pH7.4} + (\text{Number of aromatic rings})$ . Also striking in this investigation is the strong link between the sum of two chemical-physical factors and the determined ADMET characteristics. If the PFI index is less than 5, 67% of the candidates tested in Young's research exhibit acceptable values for solubility, plasma protein binding (albumin), interaction with cytochromes, intrinsic clearance, affinity for hERG, and low promiscuity. Only permeability improves as PFI increases. It should be emphasized that in the case of the PFI index, a chromatographic measurement of logD is recommended over that calculated by the octanol / water partition, because poorly soluble and lipophilic compounds do not usually enable exact readings with the distribution techniques. Young also shows how the estimated chromLogD values match the experimentally measured values well. This enables the PFI to be used during the forecast phase as well. Other simple "rules" comparable to those already described have been reported in the literature: for example, Gleeson (rule 4/400: molecular weight < 400 and  $ClogP < 4$  minimizes ADME risks) [48], or Veber (excellent bioavailability if the bonds rolling stock < 10 and the polar surface area (PSA) < 140 Å<sup>2</sup>) [49]. To overcome some of the drawbacks inherent in simple "rules," such as the discontinuity caused by cut-off values, Birketon et al. [50] proposed a model called Quantitative Estimate of Drug Likeness (QED) in which seven different chemical-physical properties (MW, ALOGP, HDB, HBA, PSA, ROTB, AROM, ALERTS) [51] and the presence of potentially toxic groups are compared with those of drugs in current use to obtain QED is a step advance in the effort to harmonize diverse chemical-physical characteristics, allowing molecules with a less-than-ideal quality to be tolerated if all other parameters are within optimal limits. In addition to the application of these principles and models, the "mapping" of the physico-chemical characteristics of the molecules in question, together with reference molecules, utilizing the study of the key components has proven valuable in our research effort.

## *Absorption*

The process through which a medicine enters the circulation is known as absorption. There are several delivery methods available, however the two most popular are intravenous and oral. When a medicine is delivered intravenously, the absorption phase is bypassed since the substance reaches circulation instantly. Many medications, however, are administered orally since it allows patients to self-administer them. When a xenobiotic is consumed, it goes via the gastrointestinal tract, then to the liver through the portal circulation, and finally to the systemic circulation, where it might be transported to the site of action. Small compounds often cross membranes during this process, sometimes passively, but most commonly via drug transporters, which are proteins. In many parts of the pharmacokinetic trip, drug transport can be a significant

component of a drug's disposition, and preclinical studies should be undertaken to offer information on how a drug interacts with various transporters - as either substrates or inhibitors.

Many variables can influence medication absorption, including molecular weight, topological polar surface area (TPSA), solubility, ionization, and other physicochemical features. Importantly, absorption data can be useful in understanding how much of the medicine reaches the circulation following oral treatment. After oral absorption, the first-pass impact (among other factors) will eventually influence bioavailability.

### *Distribution*

The reversible transfer of a medicine from one area in the body to another is referred to as distribution. Radiolabeled in vivo ADME investigations, such as quantitative whole-body autoradiography (QWBA), micro autoradiography (mARG), and tissue dissection, can provide drug developers with a comprehensive picture of drug concentration in numerous tissues and organs throughout time.

Other in vitro investigations can assist in piecing together the finer features of a compound's distribution. Permeability assays, for example, can describe a compound's ability to enter cells, drug transporter studies can help identify proteins responsible for moving a drug into (uptake) and out of (efflux) cells, and plasma protein binding (PPB) studies can quantify the extent of binding to plasma proteins, which can limit the amount of free drug available for therapeutic action or interaction with transporters or enzymes.

### *Metabolism*

Metabolism is the process by which more lipophilic xenobiotic substances are converted into hydrophilic metabolites that can be excreted from the body. A drug's metabolism involves enzymes, and numerous research investigations may be required to discover important metabolites and relevant metabolic pathways.

To confirm significant participants in a medication's metabolism and fulfill regulatory submission requirements, a few primary drug metabolism studies are undertaken in vitro. Metabolic stability studies to predict a drug's in vivo half-life, metabolite characterization and identification across species to elucidate metabolites formed and determine if any are unique to humans or disproportionately higher in humans than preclinical species, and reaction phenotyping studies to provide insight into which enzymes are responsible for metabolism are among those being conducted.

When a sponsor conducts animal research, they have frequently identified metabolic pathways, enzymes, and metabolites from previous in vitro data and

may utilize animal ADME studies to validate decisions and increase the association between in vitro predictive data and in vivo/clinical outcomes. Metabolite identification studies, which use LC-MS or radiolabeled compounds to identify and perhaps quantify metabolites in plasma and excreta from treated animals at different time periods, are a common component of an in vivo ADMET package. Metabolite identification may then be repeated during clinical trials, plasma, urine, and other bodily fluids from treated people can be tested using the same procedures to offer supportive data on which human metabolites are discovered clinically.

### *Excretion*

Excretion is the permanent removal of a substance from the body. In most situations, all drug-related material, including the parent drug and metabolites, is excreted from the body. It is critical to identify the most essential excretory pathways. The medicine is frequently eliminated by the kidney (urine) or liver (bile/feces), although it can also be excreted through perspiration, tears, or breath.

In vivo excretion studies can help to discover the route(s) of a compound's excretion as well as describe drug-related material clearance while monitoring drug and metabolite exposure in plasma and other compartments.

Radiolabeled compounds are used in animal mass balance experiments to assess a drug's excretion route and rate. Quantitative examination of urine, feces, (in certain circumstances) expired air, and corpse provides a full picture of how and at what rate a chemical is removed from the body. Other supportive research can give information to further investigate biliary excretion (bile duct cannulation technique), lymphatic partitioning rate, milk excretion, and other topics.

### *Toxicity*

The toxicity profile of a medicine is one of the most critical components of drug approval and real prospective usage. As a result, many ADMET indicators have been devised to define it: cardiotoxicity, AMES toxicity, maximum Therapeutic and tolerated dose, Rat acute toxicity, hepatotoxicity and some others. Some are obtained from animal-based testing, whereas others are human toxicity predictions.

### *QSAR model*

Ligand-based techniques are predicated on the idea that comparable molecules have similar biological characteristics. Virtual screens based on ligand-centered approaches allow for the discovery of new modulator scaffolds.

by using the structural features of known active molecules. Such approaches are especially effective when the three-dimensional structure of the biological target(s) under investigation has not yet been experimentally determined using X-ray crystallography, NMR spectroscopy, or homology modeling. Such approaches, however, plainly rely on the chemical space coverage of already known compounds. These well-documented methodologies range from pharmacophore modeling to similarity searches (SS) and Quantitative Structure-Activity Relationships (QSAR) [54-55].

The quantitative structure-activity relationship, abbreviated QSAR (Quantitative Structure-Activity Relationship), is an analytical application that may be used to assess the relationship between a particular molecule's biological activity and its chemical-physical or structural properties. When a molecule penetrates a cell membrane, its behavior is governed by features unique to the molecule. Intermolecular forces, hydrophobicity, polarity, electrostatic and steric interactions all play a role in a drug's interactions with its biological equivalents. This mathematical model is made up of a regression, which represents the statistical correlations between several variables. However, Corwin Hansch, a pioneer in pharmaceutical research, used this technique to develop an equation that connects biological activity to electrical properties and the hydrophobicity of a sequence of structures:

$$\log \left( \frac{1}{C} \right) = k_1 \log P - k_2 (\log P)^2 + k_3 \sigma + k_4$$

Where:

- C is the minimum effective dose
- P is the partition coefficient given by  $P = \frac{[\text{drug in octanol}]}{[\text{drug in water}]}$
- $\sigma$  is the Hammett constant which depends on inductive and resonance effects
- $k_x$  of the constants obtained from the regression analysis
- Log P represents a measure of the hydrophobicity of a drug, or a measure of its ability to pass through a membrane

The primary goal of QSAR is to observe biological reactions to a group of compounds, quantify them, and determine the statistical link between the activity and the molecular structure.

As a consequence, equations, pictures, or models in both 2D and 3D that connect biological reactions to medication chemical structure are obtained.





## MOLECULAR DOCKING

### *Receptor*

In the realm of molecular modeling, it is a technique that predicts the preferred orientation of one molecule towards another when they bind together to create a stable complex.

Using the score functions, for example, knowledge of the preferred orientation may be used to forecast the strength of a protein-ligand connection or binding between two molecules.

After determining this, the initial step was to download the 3D of Hsp90 $\alpha$  (isotype AA1, Alpha A1 Antibody) from the UniProt database (code P07900). MOE QuickPrep functionality with default settings was used to prepare the protein. This included structural error correction, hydrogens addition, partial charge calculation, 3D optimization of the H-bond network (Protonate3D), deletion of water molecules beyond 4.5 Å from any receptor or ligand atom, and restrained energy minimization of ligand and pocket residues within 8 Å from the ligand. Following that, all water molecules were removed and were not used in further computations.

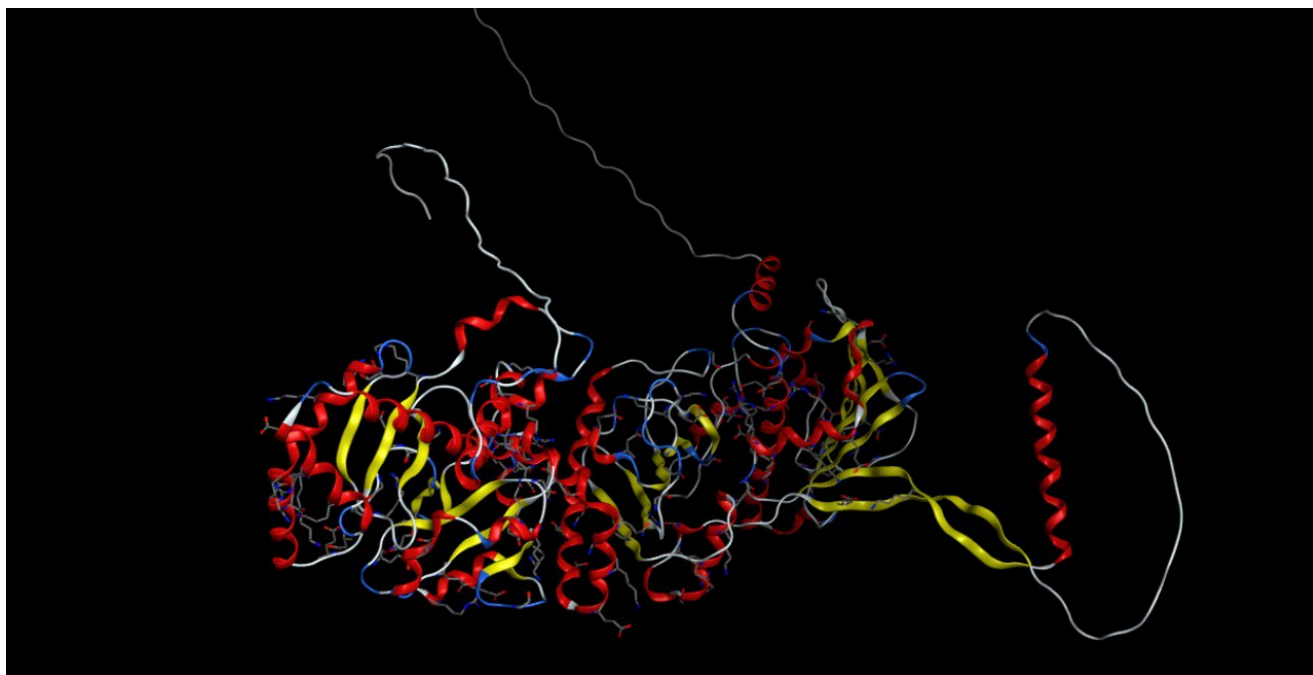


Figure 12: Crystal structure of Hsp90 alpha.

## Ligands

PubChem was used to download the structures of four ligands of interest:

Rifabutin, Geldanamycin, Alvespimycin, and Tanespimycin.

All ligands were loaded into an MOE database before proceeding into docking simulations. Two stages were taken in accordance with the MOE's database preparation guidelines:

1. Database washing to address any structural mistakes in ligands acquired from outside sources.
2. Energy minimization.

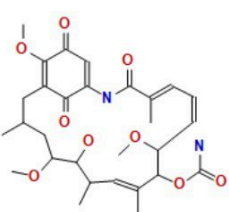
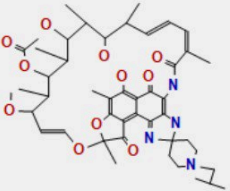
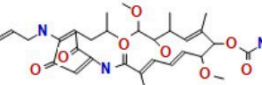
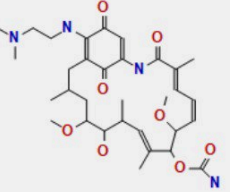
	mol	number
1		1
2		2
3		3
4		4

Figure 13: MOE's database of the compounds of interest.

## Methods

We were able to accomplish consensus docking thanks to DockBox. It is a computational package that allows you to integrate numerous docking and scoring algorithms using different consensus procedures. Using DockBox lets us to employ a novel consensus docking approach known as score-based consensus docking (SBCD), which enhances the docking prediction success rates of individual docking algorithms.

After performing a simple molecular docking on Moe with the four compounds of interest, it was discovered that Geldanamycin produced the best results in terms of S value, or binding affinity, which is the strength of the binding interaction between a single biomolecule and its ligand/binding partner. The results of MOE's Site Finder tool were utilized to do this. It is a geometric technique that does not rely on energy models to identify potential active sites in receptors. It is based on shapes, which are piecewise linear curve families in the Euclidean plane that are related with the shape of a limited collection of points [52]. The approach finds regions of dense atomic packing and then filters out implausible places such as protrusions, inaccessible regions, or those that are very solvent-exposed. The candidate sites are then rated based on their aminoacidic composition and their Propensity for Ligand Binding (PLB) score, as implemented in [53]. Six of the 40 Site Finder findings had a positive PLB score, with the first three scoring above 1, as shown in table 1.

As a result of the use of Moe, four databases were created. The first contained 154 compounds from the Zinc database, with a similarity to the drug Geldanamycin ranging from 88% to down, the second 133 compounds from the Zinc database, with a similarity to the drug Tanespimycin ranging from 89% to down, and the third 102 compounds from the Zinc database, with a similarity to the drug Alvespimycin ranging from 96% to down and the last having 109 compounds from the Zinc database with a resemblance to the drug Rifabutin ranging from 93% to down.

The use of DockBox enabled us to employ three molecular docking software programs, Vina, AutoDock, and Moe, to achieve even more precise results. Furthermore, the utilization of Compute Canada has become essential from now on.

Compute Canada, in collaboration with regional organizations WestGrid, Compute Ontario, Calcul Québec, and ACENET, is driving research innovation forward by installing cutting-edge advanced research computing (ARC) hardware, storage, and software solutions. These collaborators work together to deliver critical ARC services and infrastructure to Canadian researchers and collaborators across all academic and industry sectors.

The world-class staff of more than 200 professionals at Compute Canada, which is hired by 34 partner universities and research institutes across the nation, provides direct support to research teams. Compute Canada is a proud national and international advocate for Canadian expertise in ARC.

Several steps were completed thanks to the employment of some bash scripts:

1. It was initially essential to convert the extension of the .mol2 files obtained from MOE, which included all of the compounds, to .csv.
2. Go through the proteins folder, which contains the protein .pdb files, and generate the target folder and target.csv file.
3. Some requirements have to be entered into the configuration file. In terms of AutoDock, ga\_run = 20 and spacing = 0.4 Å, where the first is the number of autodock runs, i.e. the number of final postures intended, and spacing is grid spacing. For Vina, num modes = 20, which is a goal number of final postures. Finally, leaving the default settings was sufficient for Moe.
4. The implementation of the script "prepare sites", which produces sites using the MOE site finder, is appreciated. We might have also supplied the sites manually. The format of the file should be targetID, center, size, site, which correspond to the ID relating to the target, that is, the protein, the x, y, and z coordinates of the center of the box, the box dimensions, chosen as 30, 30, 30, and the number that identifies the site, provided in ascending order based on the PBL value.
5. The procedure begins once you have prepared the job folders and to submit folder so that you may execute everything at once. There are 3 levels to run it. you can use the help manual to go through your options. You can change the level based on the number of ligands to dock multiple ligands (on all targets) in a single job, and you can also specify the number of ligands to run for each submitted job. DockBox also allows you to rescore the generated docking poses using other scoring techniques and review the results using different consensus docking/scoring procedures.
6. Once the required time has elapsed, you need to use additional bash scripts specialized to this purpose to extract the results and the best poses, with the relative S values.

The top-scoring poses from different docking motors were compared in terms of RMSD in the adopted consensus, and only the comparable ones, that is, those that differed by less than a certain RMSD threshold, were maintained. If the RMSD between the poses is less than or equal to a predefined threshold

(default: 2.0), `extract_dbx_best_poses` deems it a consensus and returns the appropriate pose. Finally, the new score-based consensus docking (SBCD) technique evaluates all of the poses created by all of the programs at the same time and compares the poses that were ranked highest by different scoring functions during the rescoring stage. A consensus is obtained, as in the CD technique, if the top postures predicted by the scoring functions are comparable (default:  $\text{RMSD} < 2.0 \text{ \AA}$ ).

### *Rescoring*

It is feasible to establish a consensus score by rescoring binding poses with several functions and aggregating the scores, which leads to better accuracy when compared to single scoring functions [56].

Several methods for combining individual scores into a consensus score have been proposed, including establishing a pass/fail cut-off for each function and counting how many functions pass for a specific ligand (voting), averaging the ranks (rank-by-rank) or binding affinities (rank-by-number) obtained using different scoring functions, or combining the scores using fitted coefficients (regression) [56]. Docking result rescoring is faster than more complicated approaches for estimating binding energies, such as endpoint or alchemical free-energy calculations, and hence more appropriate for large-scale VS campaigns.

## Results

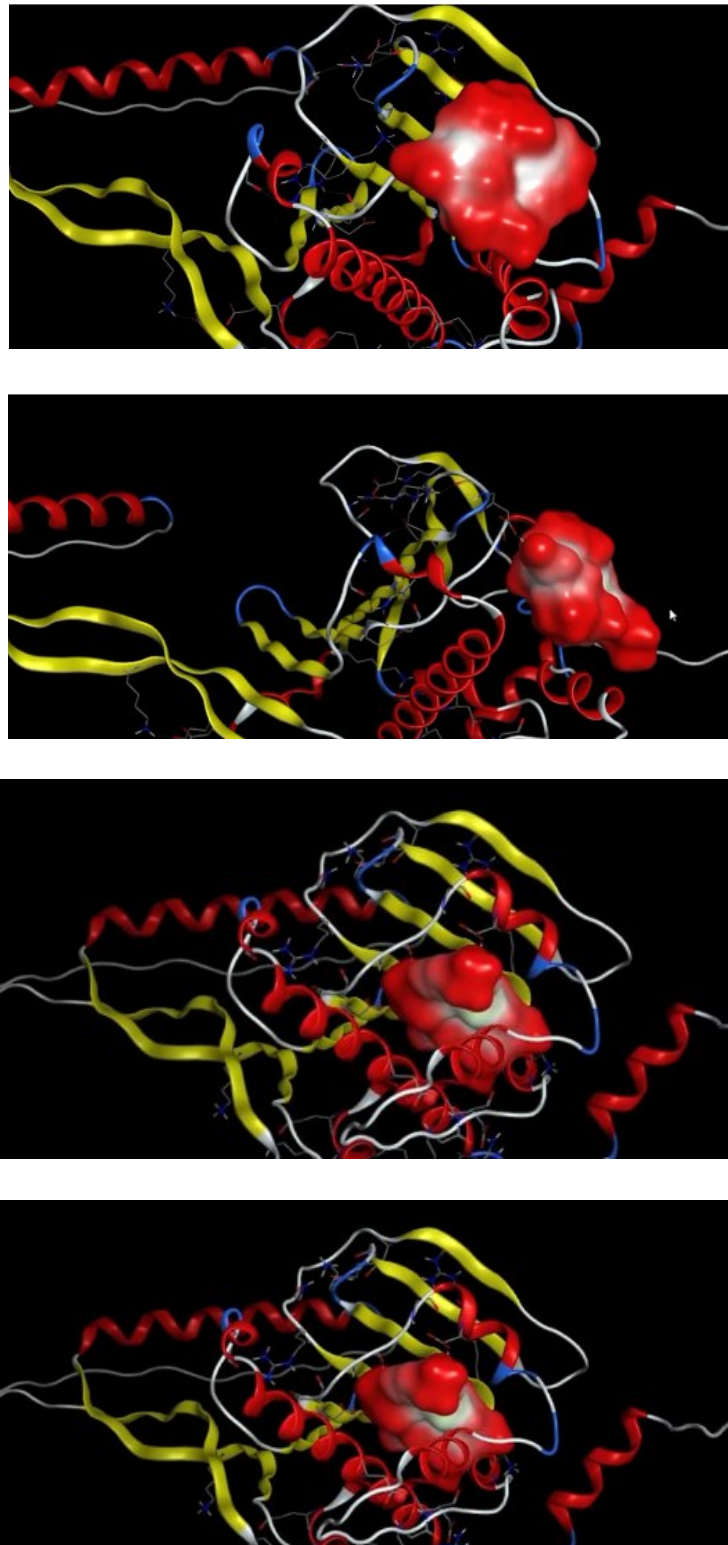


Figure 14: Hsp90 alpha binding Geldanamycin, Alvespimycin, Tanespimycin, and Rifabutin.

In the previous images, from top to bottom, N-terminal domain of human Hsp90 alpha binding Geldanamycin, Alvespimycin, Tanespimycin, and Rifabutin can be seen.

Table 1 contains five of the best score, in terms of binding affinity, of three other molecular docking programs: Vina, AutoDock, and Moe. The best binding site for each score is indicated in the second column.

<b>LigID</b>	<b>Site</b>	<b>Score</b>
Tanespimycin	autodock.site1	-13.320
Alvespimycin	autodock.site1	-12.690
Geldanamycin	autodock.site1	-11.450
Rifabutin	autodock.site1	-7.680



Table 1: Molecular docking outputs.

## MOLECULAR DYNAMICS

Molecular dynamics simulations are primarily used to aid in the study of protein dynamics at various timescales. In particular, molecular dynamics methodologies allow the molecular systems under investigation, which may be envisioned as a sequence of interacting atoms and molecules, to change over time and are studied to aid in interpreting the findings of biophysical experiments and modeling studies. The trajectories created by MD experiments may be solved numerically by integrating Newton's equation of motion, where forces operating between atoms and molecules, as well as their energy contributions, are measured using force fields (if molecular systems in MD experiments are described according to molecular mechanics). MD may be utilized to investigate the flexibility of receptor-binding sites in the ligand-protein binding posture.

Geometrical analysis methods (such as H-bond distance or RMSD calculations) and free-energy based scoring algorithms can be used to identify essential structural and/or energetic aspects directing ligand-protein interaction [57]. Molecular dynamics was used in this thesis to calculate RMSD and extract the trajectory for visualization and other post-processing purposes. Some procedures were also required in this case:

1. In most circumstances, we do not require information from water, thus we delete the water and ions from the parameter file to be consistent with the trajectory. We're outputting the trajectory without the water.
2. The real simulation starts now. It will be feasible to produce the RMSDs and retrieve the trajectories in the end.

The RMSD graphs generated for each complex, Hsp90-Geldanamycin, Hsp90-Alvespimycin, Hsp90-Tanespimycin, and Hsp90-Rifabutin, are presented below. It is difficult to evaluate RMSD charts. In reality, we know that the RMSD does not take directionality into account, thus it can only identify if the structure has attained convergence, that is, if the graph flattens, I can finish simulating; if it does not flatten, I am out of equilibrium and cannot compute set attributes. On very complex systems, it is possible that a flat function will not reach an equilibrium, but if the RMSD continues to oscillate for an ideally infinite time, I will have to stop and establish this condition myself, looking at the stabilization of the average value and more than anything else that is freely spreading.



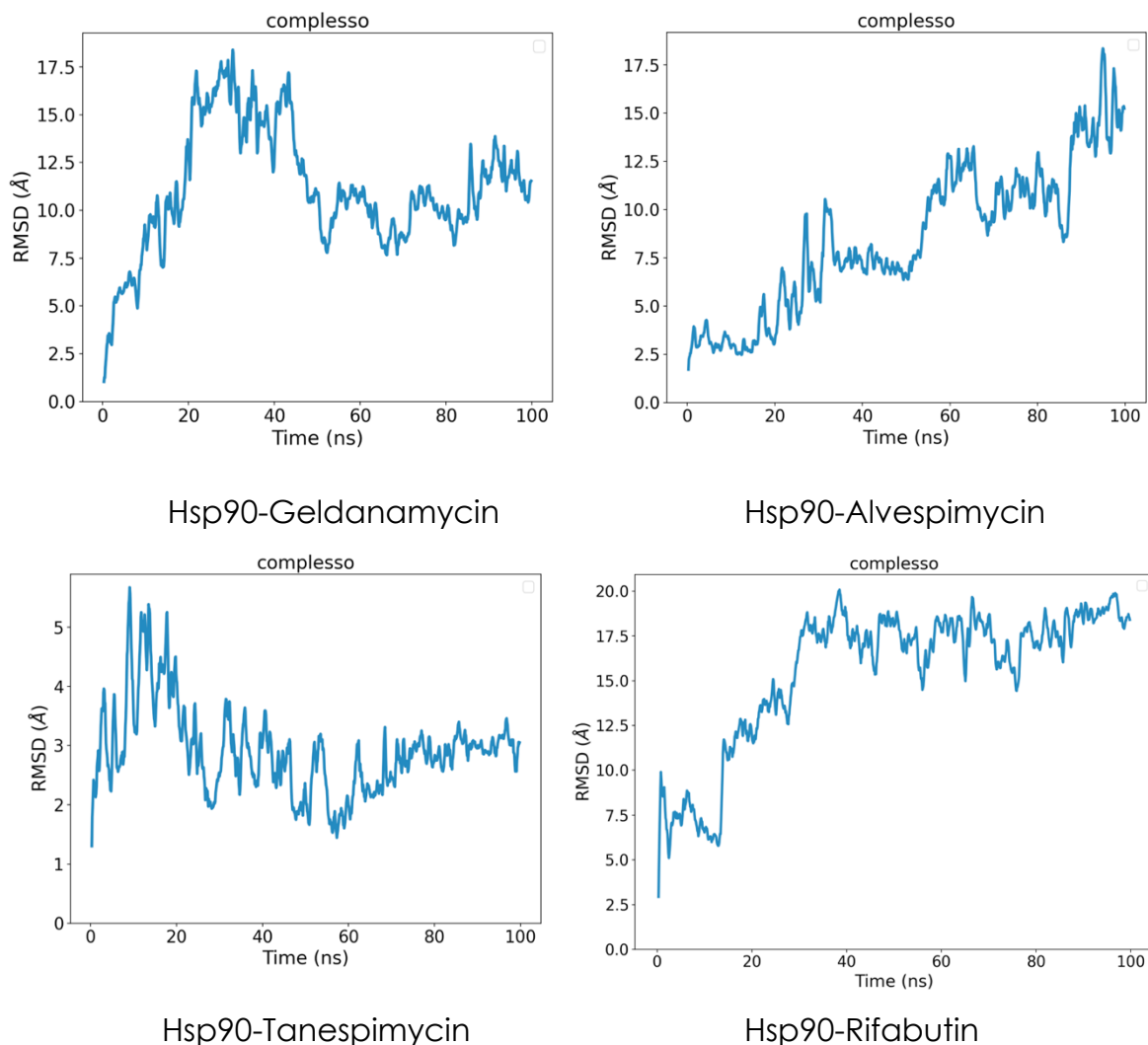


Figure 15: RMSD plots of the complexes of interest.

When it reaches a plateau there are configurations that are equally distant from the initial configuration, but this does not mean that they are equal to each other, but that a conformational equilibrium has been reached in order to calculate the macroscopic properties of interest.

The RMSD tells us precisely the trend of the simulation in terms of conformational convergence, in the sense that the RMSD is defined as the average with respect to the distance of the protein with respect to a representative configuration that we choose, in the X axis we have time, because an RMSD value is associated with each instant of time, which depends on the difference between the configuration's distance at instant T and the representative one. Because it may be organized more freely in a vacuum, it initially establishes an equilibrium, but this does not occur in water (hydrophobic and hydrophilic). Because we are

discussing two structures with the same configuration that may be overlaid, 0 begins at 0. It is therefore a measure of similarity; it is a metric.

Furthermore, at the end of the molecular dynamics, we get a trajectory, which is not infinite, so we can't explore the complete space of states, so I'd have to run a lot of simulations to get a "bundle" of trajectories and explore it better. In truth, molecular dynamics is a deterministic system.

Important information is obtained from the analysis of the trajectory (or trajectories) obtained from a molecular dynamics experiment:

- RMSD: As already seen previously, it represents the proximity, in structural terms, to the native structure. A distance between 3 and 4 Å with respect to the initial structure is generally used as an indicator of the achievement of folding.
- SASA change: The folding process is typically accompanied by a significant decrease in the Solvent Accessible Surface Area (SASA). The SASA, therefore, or alternatively the radius of gyration, is an indicator of the folding process.
- Folding landscape: since folding is a multidimensional problem, analyzes based on one-dimensional reaction coordinates offer an incomplete picture of the folding process. The coordinates typically used are the RMSD, the turning radius and the percentage of native contacts.



## ANALYSIS OF ADMET PROPERTIES

### *Material and methods*

By using the ADMET Predictor 10.2 by Simulations Plus and the online resources SwissADME [60] and pkCSM [61] it was possible to make a prediction of the pharmacokinetic, drug-likeness and physicochemical descriptors properties of the tested compounds.

### *ADMET Predictor Simulations Plus*

ADMET Predictor is a machine learning software application that made a prediction of over 175 attributes such as solubility, logP, pKa, CYP metabolism sites, and Ames mutagenicity. The most recent version combines market leading ADMET modeling with compound design, data analysis, SAR, and cheminformatics features to help scientists in computational chemistry, medicinal chemistry, DMPK, and other fields.

In a few steps you will get the forecasts of interest:

- Load a SMILES file containing the desired ligands.
- Choose the desired properties and begin the calculations.
- Move to Excel to examine the table results in a more straightforward and plain manner, owing to the usage of the Excel spreadsheet given as output.

It is more complete than the other tools since it provides some intriguing cumulative scores relating to the compound's chance of failing as a medicine owing to ADMET issues, as well as associated codes that clarify what these issues are:

- TOX Risk: a value from 0 to 6 reflecting the amount of expected toxicity issues.
- CYP Danger: a risk associated with the oxidation of cytochromes P450. The score ranges from 0 to 6, indicating the number of possible issues that the substance may have owing to metabolism by one or more of the five main cytochrome P450 isoforms.
- Absn Risk: a score between 0 and 8, indicating the likelihood of the compound having oral absorption issues.
- MUT Danger: a value ranging from 0 to 5.4 reflecting the risk of mutagenicity.

- RuleOf5: a score indicating the quantity of possible problems with passive oral absorption that a compound is expected to have based on Lipinski's Rule of Five.

All of these data are combined in the ADMET Risk Prediction.

ADMET Predictor Simulations Plus is used to compare the other two tools because is considered as the golden standard.

### SwissADME

SwissADME is a free web tool that provides free access to a pool of quick yet rigorous prediction models for physicochemical qualities, pharmacokinetics, drug-likeness, and medicinal chemistry friendliness, including proprietary approaches like the BOILED-Egg, iLOGP, and Bioavailability Radar.

In comparison to the state-of-the-art of free web-based tools for ADME and pharmacokinetics (e.g., pk-CSM14 and admetSAR15), and aside from unique access to proficient methods (e.g., iLOGP16 or the BOILED-Egg17), SwissADME strong points include, but are not limited to different input methods, computation for multiple molecules, and the ability to display, save, and share results per individual molecule or SwissADME is also incorporated into the SwissDrugDesign workspace. One-click interoperability provides access to various CADD tools developed by the SIB Swiss Institute of Bioinformatics Molecular Modeling Group, such as ligand-based virtual screening (SwissSimilarity), biotarget prediction (SwissTargetPrediction), molecular docking (SwissDock), bioisosteric design (SwissBioisostere), and molecular mechanics (SwissParam) [58].

A molecular sketcher is included in the input section, allowing the user to import, create, and change a 2D chemical structure before transferring it to a list of molecules. This list represents the calculation's real input. It is editable like plain text, allowing you to input or paste SMILES. One input molecule per line is characterized by a SMILES and optionally a name separated by a space in the list. If the name is not entered, SwissADME will generate an identification for you. When the computation is finished, the output panel is filled one molecule at a time. This allows you to view the findings of the first compounds without having to wait for the complete list to be processed. There are numerous parts in this molecular panel:

- The first section, which comprises the two-dimensional chemical structure and canonical SMILES, identifies the chemical form on which the predictions were made. In addition, the bioavailability radar is shown for fast drug similarity assessment. Lipophilicity, size, polarity, solubility, flexibility, and saturation are the six physicochemical qualities evaluated. Adapted descriptions defined a physicochemical range on each axis,

which was shown as a pink region within which the molecule's radar plot must fall entirely to be termed drug-like. More information on the descriptors may be obtained using the radar. This pink area shows the optimal range for each property (lipophilicity: XLOGP3 between 0.7 and +5.0, size: MW between 150 and 500 g/mol, polarity: TPSA between 20 and 1302, solubility: log S less than 6, saturation: fraction of carbons in the sp<sup>3</sup> hybridization less than 0.25, and flexibility: no more than 9 rotatable bonds).

- Following this part, the depiction of the physicochemical characteristics begins, which are as follows:
  - Formula
  - Molecular weight
  - Number of heavy atoms
  - Number of heavy aromatic atoms
  - Fraction of carbons in the sp<sup>3</sup> hybridization
  - Number of rotatable bonds
  - Number of hydrogen bond acceptors
  - Number of hydrogen bond donors
  - Molar refractivity, i.e., the measure of the total polarizability of a mole of the given compound
  - TPSA

### *Lipophilicity*

Lipophilicity is traditionally described by the partition coefficient between n-octanol and water (log Po/w). Because of the vital relevance of this physicochemical characteristic in pharmacokinetic drug discovery, it gets its own section in SwissADME. Many computers approaches for estimating log Po/w have been devised, with varying degrees of success on various chemical sets. Multiple predictors are commonly used to either pick the best accurate approaches for a specific chemical series or to achieve consensus estimation. To improve prediction accuracy by consensus log Po/w, the models underpinning the predictors should be as varied as feasible. SwissADME provides access to five freely available predictive models: XLOGP3, an atomistic method with corrective factors and a knowledge-based library; WLOGP, our own implementation of a purely atomistic method based on Wildman and Crippen's fragmental system; MLOGP, an archetype of topological method based on a linear relationship with 13 molecular descriptors

implemented. SILICOS-IT, a hybrid technique based on 27 fragments and 7 topological descriptors; and iLOGP, our proprietary physics-based method based on solvation free energies in n-octanol and water derived using the Generalized-Born and solvent accessible surface area (GB/SA) model. iLOGP was tested against two drug or drug-like external sets and outperformed six well-established predictors<sup>16</sup>. The arithmetic mean of the values anticipated by the five recommended approaches is the consensus log Po/w [58].

### *Water solubility*

In terms of water solubility, having a soluble molecule considerably simplifies several drug development operations, particularly handling and formulation. Furthermore, for oral administration discovery initiatives, solubility is a crucial feature determining absorption. A medicine intended for parenteral administration must also be highly soluble in water in order to deliver a sufficient amount of active component in the tiny volume of such pharmacological dosage. SwissADME includes two topological approaches for predicting water solubility. The first is an application of the ESOL model, while the second is an adaptation of Ali et al.. Both vary from the foundational universal solubility equation in that they do not include the melting point parameter, which is difficult to estimate. The decimal logarithm of the molar solubility in water is used to calculate all expected values (log S). SwissADME also offers solubility in mol/l and mg/ml units, as well as qualitative solubility classes.

### *Pharmacokinetics*

Individual ADME behaviors of the chemical under inquiry are evaluated using specialized models, the predictions of which are provided in this section. A multiple linear regression model, for example, seeks to predict the skin permeability coefficient ( $K_p$ ).  $K_p$  was discovered to be linearly linked to molecule size and lipophilicity ( $R^2 = 0.67$ ). The lower the log  $K_p$  (in cm/s), the less permeant the molecule is to the skin.

The BOILED-Egg model, a straightforward graphical categorization model, is used to estimate passive human gastrointestinal absorption (HIA) and blood-brain barrier (BBB) permeability. Other binary categorization methods are offered, which focus on a particular small molecule's proclivity to be a substrate or inhibitor of proteins that influence crucial pharmacokinetic behaviors. Knowledge of compounds that are substrates or non-substrates of the permeability glycoprotein (P-gp, suggested to be the most important member among ATP-binding cassette transporters or ABC-transporters) is essential for evaluating active efflux through biological membranes, such as from the gastrointestinal wall to the lumen or from the brain. P-gp has an important function in protecting the central nervous system (CNS) from xenobiotics. P-gp is

also overexpressed in certain tumor cells, which contributes to multidrug-resistant cancers.

It is also necessary to understand how chemicals interact with cytochromes P450 (CYP). This isoenzyme superfamily plays an important role in drug removal via metabolic biotransformation. It has been proposed that CYP and P-gp can process small compounds synergistically to promote tissue and organism protection<sup>44</sup>. It is estimated that 50 to 90% of medicinal compounds are substrates of one or more of the five main isoforms (CYP1A2, CYP2C19, CYP2C9, CYP2D6, CYP3A4). Inhibition of these isoenzymes is undoubtedly a significant source of pharmacokinetics-related medication-drug interactions, which can result in toxic or other undesired side effects due to decreased clearance and buildup of the drug or its metabolites.

Several CYP isoform inhibitors have been found. Some impact various CYP isoforms, whilst others are selective for certain isoenzymes. It is consequently critical for drug development to anticipate the likelihood that a chemical would produce substantial drug interactions by inhibiting CYPs, as well as to establish which isoforms are impacted.

SwissADME can predict if a compound is a P-gp substrate or an inhibitor of the most major CYP isoenzymes. The models answer "Yes" or "No" if the molecule under examination is more likely to be a P-gp substrate or not (respectively inhibitor or non-inhibitor of a given CYP) [58].

### *Drug-likeness*

"Drug-likeness" quantifies the likelihood of a chemical becoming an oral drug in terms of bioavailability. Structure or physicochemical examinations of development compounds progressed enough to be regarded oral drug candidates revealed drug-likeness. This concept is commonly used to filter chemical libraries in order to avoid compounds having features that are most likely incompatible with an acceptable pharmacokinetics profile. This SwissADME area provides access to five alternative rule-based filters, each with a different set of attributes within which the molecule is classified as drug-like. The Lipinski filter was the first to use the rule-of-five. The techniques of Ghose, Veber, Egan, and Muegge were modified. Multiple estimations provide consensus views or the selection of techniques best suited to the end-unique user's requirements in terms of chemical space or project-related demands. The Abbot Bioavailability Score is similar, but it attempts to estimate a compound's likelihood of having at least 10% oral bioavailability in rats or detectable Caco-2 permeability. This semi-quantitative rule-based score identifies four types of compounds with probability of 11%, 17%, 56%, or 85% based on total charge, TPSA, and violation of the Lipinski filter [58].



The goal of this section is to help medicinal chemists with their everyday drug development efforts. Two complimentary pattern recognition algorithms can be used to identify possibly troublesome segments. PAINS (pan assay interference chemicals, also known as frequent hitters or promiscuous compounds) are molecules with substructures that exhibit robust response in assays regardless of the protein target. If such moieties are detected in the molecule under review, SwissADME issues warnings.

Furthermore, they used Structural Alert, which is a list of 105 fragments identified by Brenk et al. to be potentially hazardous, chemically reactive, metabolically unstable, or to have traits that cause poor pharmacokinetics. Flying over the "question mark" symbol displayed after the fragment list in SwissADME allows you to get a chemical description of the problematic fragments discovered in a specific molecule. This is done for both the PAINS and Brenk filters. Brenk et al. discovered that majority of the remaining compounds meet requirements for "lead likeness" by using these and other physicochemical filters to construct screening libraries. This idea is comparable to drug-likeness, but it focuses on the physicochemical boundaries that define a good lead, i.e., a molecular entity that can be optimized. Leads are, by definition, subjected to chemical changes that would most likely enhance their size and lipophilicity. As a result, leads must be smaller and less hydrophobic than drug-like compounds. Because it is critical for a chemist to determine whether a particular molecule is acceptable to begin lead optimization, they devised a rule-based technique for lead likeness in addition to structural filters.

One of the most important components of CADD operations is assisting in the selection of the most promising virtual molecules to be synthesized and tested in biological assays or other investigations. In this selection procedure, synthetic accessibility (SA) is an important criterion to consider. Obviously, medicinal chemists are the best at determining SA for a reasonable number of compounds. When there are too many molecular structures to evaluate, *in silico* estimate can be employed for pre-filtering. Ertl and Schuffenhauer provided a fingerprint-based technique for SA estimation, but included closed-source information regarding fingerprint definition, which prohibits a simple implementation in our open-source tool. As a result, we developed our own fragmental approach by analyzing over 13 million chemicals that are promptly deliverable by suppliers. We hypothesized that the most common molecular fragments (FP2 bits, see Computational Methods) in this vast collection suggest a likely high SA, whereas uncommon fragments signal a complex synthesis. For a specific molecule, the fragmental contributions to SA are averaged and adjusted by Ertl and Schuffenhauer parameters indicating size and complexity, such as macrocycles, chiral centers, or spiro functions. The SA Score, after normalization, varies from 1 (extremely easy) to 10. (very difficult). We retrieved

two previously released SA test sets to evaluate the performance of the approach developed for SwissADME. External molecules were used in both sets, and their difficulty of synthesis was graded on a scale of 1 to 10 by nine and four medicinal chemists, respectively. After that, the mean expert score may be compared to an in-silico SA Score. The predictive power of all three approaches appears to be highly reliant on the test set. Indeed, the SAs of set 1, which were smaller and reviewed by more chemists, proved to be significantly more robustly predicted than those of set 2. Human judgement of synthetic complexity is obviously subjective and is dependent on the expertise and experience of individual chemists. However, significant linear correlation and small errors, particularly with the SwissADME SA Score, which outperformed the reference methods on both sets with smaller errors and equal or higher linear correlation coefficients, show how this simple and fast methodology can aid in molecule prioritization [58].

### *BOILED-Egg graph*

The BOILED-Egg is a simple approach for predicting two major ADME characteristics at the same time: passive gastrointestinal absorption (HIA) and brain access (BBB). It only uses two physicochemical descriptors (WLOGP and TPSA, for lipophilicity and apparent polarity). The yolk (i.e., the physicochemical area for highly probable BBB permeability) and the white are included in the egg-shaped categorization plot (i.e., the physicochemical space for highly probable HIA absorption). Both compartments are not mutually exclusive, and the outside grey zone represents chemicals with projected poor absorption and brain penetration. While the BOILED-Egg has a broad chemical space predictive power, it is limited to passive penetration through the gastro-intestinal wall and BBB.

We have a global assessment of passive absorption (inside/outside the white), passive brain access (inside/outside the yolk), and active efflux from the CNS or to the gastrointestinal lumen by color-coding: blue dots for P-gp substrates (PGP+) and red dots for P-gp non-substrate (PGP): this allows for intuitive evaluation of passive gastrointestinal absorption (HIA) and brain penetration (BBB) in function on the same graph [58].

### *pkCSM*

pkCSM is a revolutionary technology that uses distance-based graph signatures to predict and optimize small-molecule pharmacokinetic and toxicity parameters. 30 predictors were developed as a result of the adaptation of the Cutoff Scanning concept to represent small-molecule structure and chemistry in order to represent and predict their pharmacokinetic and toxicity properties:

absorption (seven predictors), distribution (four predictors), metabolism (seven predictors), excretion (two predictors), and toxicity (10 predictors). Given a collection of input molecules, two sets of descriptors are produced and integrated for use in the following machine learning step: generic molecule attributes and a distance-based graph signature. The first significant component of the pkCSM signature pertains to molecular features, which include:

- a toxicophore fingerprint
- an atomic pharmacophore frequency count
- lipophilicity (log P), molecular weight, surface area, and the number of rotatable bonds are all examples of general molecular characteristics

The ADMET properties prediction on the pkCSM platform is split into two groups of highly predictive models: 14 regression models that aim to predict a numeric quantification of the pharmacokinetic or toxicity property and 16 classification models that classify the output into two different classes [59].

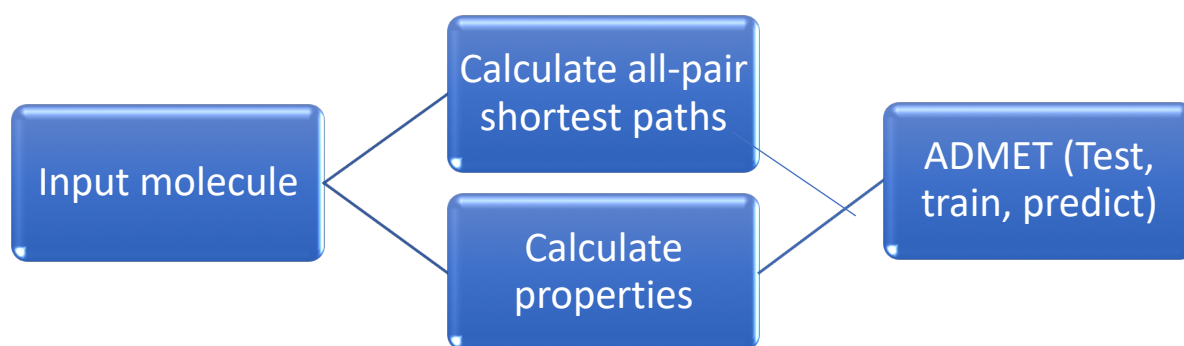


Figure 16: pkCSM operating diagram.

Some components in these five categories, such as absorption, distribution, metabolism, excretion, and toxicity, strike the eye due to their significance. As previously stated, the log S value, i.e., the 10-based logarithm of the molar concentration, is supplied here as well. A binary (yes/no) value is supplied for P-glycoprotein substrate and P-glycoprotein I and II inhibitor. The skin permeability, for which the log Kp value is given. A log Kp value greater than 2.5 suggests that the skin is relatively permeable. Caco-2 permeability predicts the logarithm of the human epithelial colorectal cancer cell line in the absorption region. To

estimate oral medication absorption, a monolayer of these cells is employed as an in vitro model of the human intestinal mucosa, followed by the apparent permeability coefficient ( $\log P_{app}$ , where  $P_{app}$  is given in cm/s); it is deemed high if the projected values are larger than 0.90. The expected proportion of the substance absorbed through the small intestine is provided. A substance with less than 30% absorbance is expected to be poorly absorbed.

In the distribution section, we observe the fraction unbound ( $F_u$ ) in plasma, as well as the previously mentioned BBB permeability, reported as  $\log BB$ , i.e., the logarithmic ratio of brain-to-plasma drug concentrations. Compounds with  $\log BB < -1$  are projected to be poorly dispersed to the brain, but  $\log BB > 0.3$  suggests a fast crossing of the barrier, as is the  $VD_{ss}$  (Volume of Distribution at a steady state):  $\log VD_{ss}$ . The estimated value is deemed low if it is less than -0.15 ( $VD_{ss} < 0.71$  L/kg) and high if it is greater than 0.45 ( $VD_{ss} > 2.81$  L/kg). The CNS permeability is determined as the  $\log PS$ , which is the blood-brain permeability-surface area product. The two parameters, BBB permeability and CNS permeability are interpreted in the same way, but are measured differently. BBB permeability is measured as the ratio of brain to plasma concentration. CNS permeability, on the other hand, is a more direct measurement, as the brain is perfused in situ with the compound that is injected into the carotid artery, thus bypassing/shielding the systemic effect of distribution of the compound in other part of the body. It could be used to understand in detail, as far as the brain is concerned, what effect the compound has. However, both come from in vivo measurements.

In terms of metabolism, a binary value (yes/no) indicates whether the molecule is likely to be an inhibitor of the cytochrome P450 isoforms CYP1A2, CYP2C19, CYP2C9, CYP2D6, and CYP3A4 or a substrate of the CYP2D6 and CYP3A4 isoforms. If a substance inhibits CYP450 at a concentration less than 10  $\mu$ M, it is termed a CYP450 inhibitor (50% inhibition).

Last but not least, we have the toxicity section, which gives the mutagenicity prediction (AMES toxicity), as well as the T. Pyriformis toxicity is measured in  $\log \mu$ g / L; a value greater than -0.5 is deemed poisonous. The Minnow toxicity, derived as the logarithm of  $LC_{50}$ , moreover the Rat  $LD_{50}$ , provided in mol/kg. Values less than -0.3 ( $LC_{50} < 0.5$  mM) indicate severe acute toxicity. A crucial component for toxicity and medication administration is also found here: the Maximum Recommended Tolerated Dose (MRTD): on a logarithmic scale,  $MRTD \leq 0.477$   $\log$  (mg/kg/day) is considered low, while MRTD beyond 0.477 is considered excessive. Not to mention the chronic rat oral toxicity, provided in  $\log$  (mg/kg<sub>bw</sub>/day), hepatotoxicity, skin sensitization, and cardiotoxicity (hERG I and II inhibitors).

## Results

The data acquired from the three ADMET predictors stated accurately earlier will be shown in this paragraph. A comparison will be conducted between them (pkCSM, ADMET Predictor Simulations Plus and SwissADME) as well. The data comes from the analysis of the pharmacokinetic properties of the drugs Geldanamycin, Tanespimycin, Alvespimycin and Rifabutin.

Before evaluating the various predictors, the first focus will be on the graphs of interest provided by SwissADME.

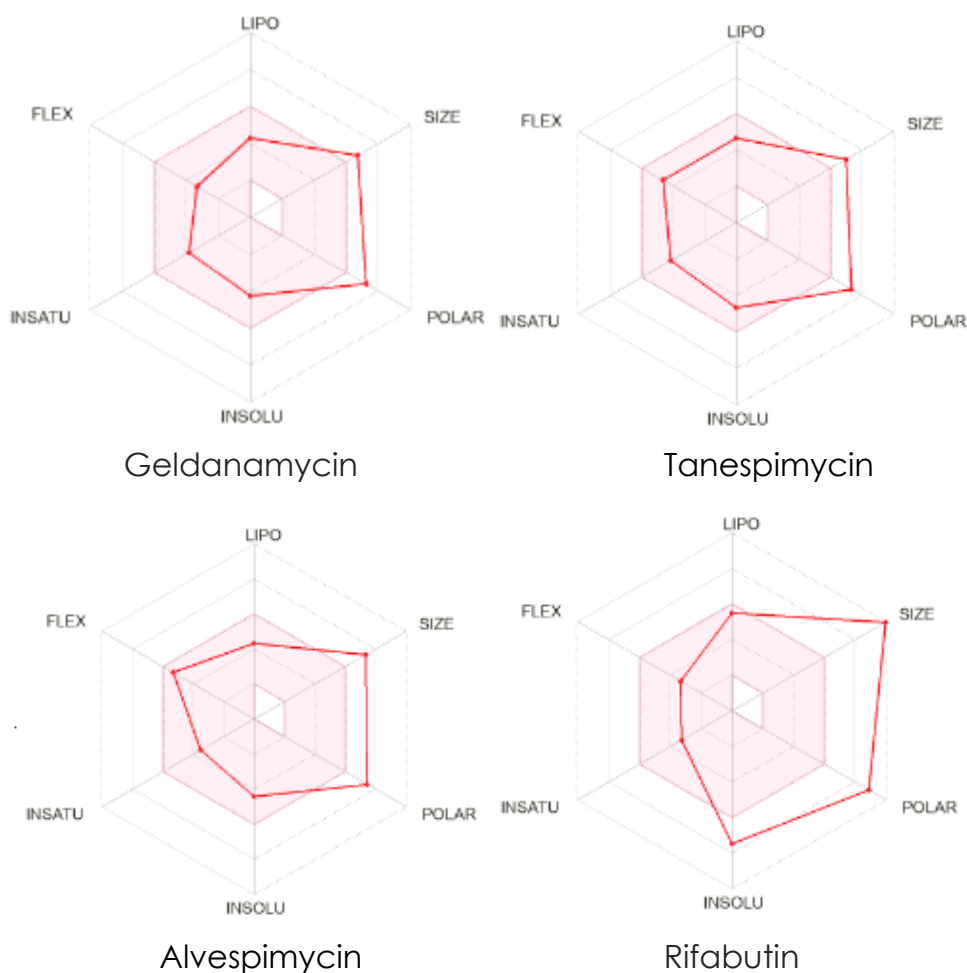


Figure 17: Graphs by SwissADME.

As already illustrated previously in paragraph about SwissADME, the colored zone is the appropriate physicochemical area for oral bioavailability (considered lipophilicity, molecular weight, polarity, solubility, flexibility and saturation). Geldanamycin, Tanespimycin and Alvespimycin are predicted not orally

bioavailable, because they have a very high molecular weight and they are too polar. Rifabutin, on the other hand, exhibits the same issues with the added of being insoluble.

This is not the only graph provided by SwissADME; the online tool also calculates Boiled-EGG.

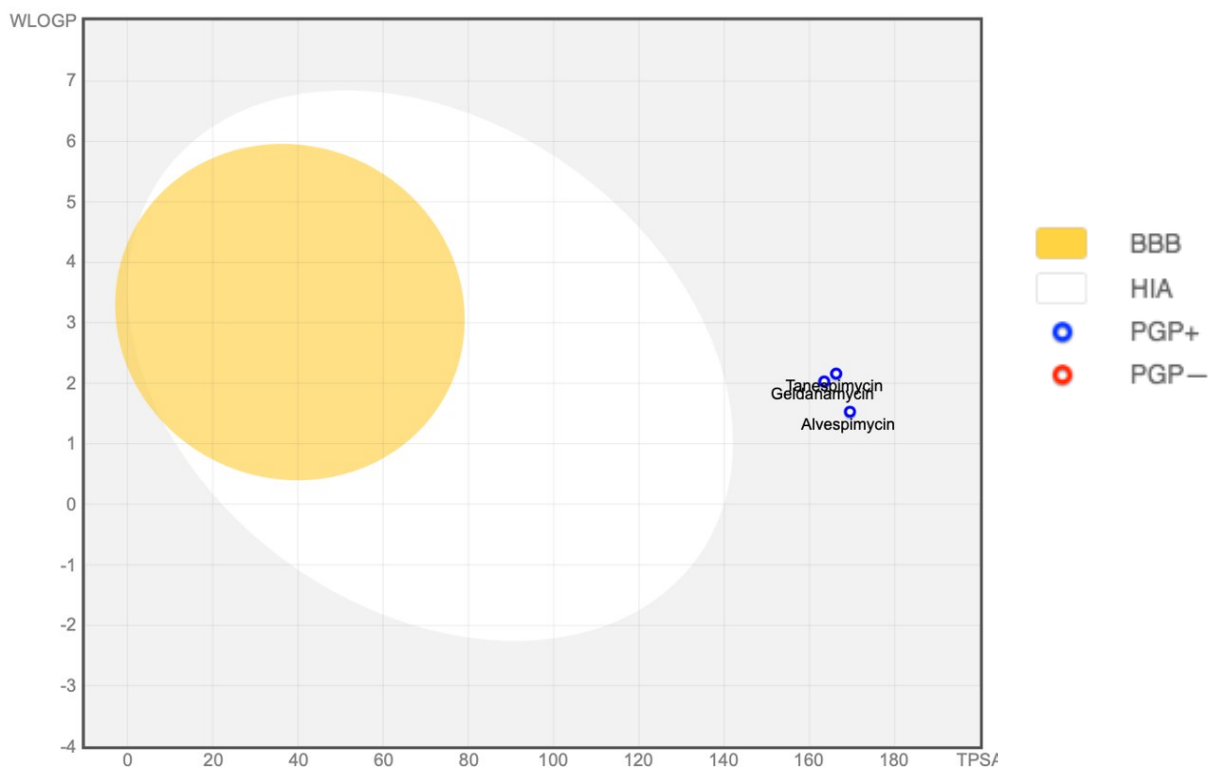


Figure 18: Boiled-EGG.

In this instance, all drugs are projected to be non-absorbed, non-brain penetrant (outside the Egg), and pumped out of the brain (blue dot). One molecule is anticipated not to be absorbed since it is outside of the plot's range (Rifabutin with a TPSA of 209.04).

### Absorption

Compounds	ESOL	ALI	SILICOS-IT	pkCSM	Simulation Plus
Geldanamycin	$3.23 \cdot 10^{-2}$	$5 \cdot 10^{-3}$	$3.13 \cdot 10^{-2}$	$8.03 \cdot 10^{-3}$	$1.42 \cdot 10^0$
Tanespimycin	$1.25 \cdot 10^{-2}$	$9.66 \cdot 10^{-4}$	$5.92 \cdot 10^{-2}$	$2.62 \cdot 10^{-2}$	$5.41 \cdot 10^{-1}$
Alvespimycin	$2.34 \cdot 10^{-2}$	$3.65 \cdot 10^{-3}$	$5.85 \cdot 10^{-2}$	$1.06 \cdot 10^{-2}$	$3.52 \cdot 10^0$
Rifabutin	$2.82 \cdot 10^{-5}$	$5.15 \cdot 10^{-6}$	$7.71 \cdot 10^{-5}$	$6.61 \cdot 10^{-1}$	$1.26 \cdot 10^{-1}$

Table 2: Water solubility expressed in mg/mL.

As seen in the table above, Rifabutin is poorly soluble, in fact it produces the worst results, Alvespimycin is moderately soluble, Tanespimycin and Geldanamycin are moderately soluble for the first two techniques of prediction, although soluble for SILICOS-IT.

Keeping in mind that Insoluble < -10 < Poorly < -6 < Moderately < -4 < Soluble < -2 < Very < 0 < Highly.

As for P-glycoprotein substrate prediction, all compounds show themselves as substrates.

All 4 drugs show 2 violations of the Lipinski's rule of five, corresponding to the fact that they have a molecular weight (MW) > 500 g/mol, number of N or O atoms > 10 (H-bond acceptors).

### *Distribution*

Because SwissADME does not give information on distribution parameters, we will examine those shared by the other two predictors: Volume of distribution (VDss) and fraction unbound (Fu).

Compounds	pkCSM	Simulation Plus
Geldanamycin	-0.69	0.77
Tanespimycin	-0.76	0.77
Alvespimycin	-0.52	0.91
Rifabutin	1.33	0.78

Table 3: Volume of distribution.

Compounds	pkCSM	Simulation Plus
Geldanamycin	0.32	0.84
Tanespimycin	0.31	0.77
Alvespimycin	0.48	0.76
Rifabutin	0.10	0.54

Table 4: Fraction unbound.

Compounds	pkCSM	Simulation Plus
Geldanamycin	-1.38	-0.62
Tanespimycin	-1.17	-0.67
Alvespimycin	-1.14	-0.56
Rifabutin	-1.79	-0.94

Table 5: logBB.

Compounds	pkCSM
Geldanamycin	-3.14
Tanespimycin	-3.18
Alvespimycin	-3.32
Rifabutin	-3.22

Table 6: Log PS (CNS penetration).

We will utilize the guidelines provided by pkCSM to analyze the findings shown above: VDss is low if less than -0.15 and high if greater than 0.45 and a log BB less than -1 implies poor distribution to the brain, whereas substances with log BB greater than 0.3 are projected to cross the BBB easily.

Outside of Rifabutin, all other drugs have a low VDss, indicating that the higher the VD is, the more of a drug is distributed in tissue rather than plasma. It is susceptible to renal failure and dehydration.

Most drugs in plasma will be in an equilibrium condition between being unbound or bound to serum proteins. The degree to which a specific drug binds proteins within blood may impair its efficacy, since the more that is bound, the less efficiently it may penetrate cellular membranes or disseminate. In the instance of pkCSM, Alvespimycin has the highest Fu, although Geldanamycin has the highest Fu in Simulation Plus.

All of the compounds investigated had a poor distribution to the brain.

The findings of pkCSM deviate significantly from the numbers reported by Simulations Plus in all the parameters analyzed.

pkCSM offers a log PS value, which may be a more accurate estimate of the capacity to reach the brain, logPS, often known as CNS permeability.

Compounds with logPS > -2 are expected to permeate the Central Nervous System (CNS), whereas those with logPS < -3 are thought to be unable to do so. Nobody can penetrate the CNS.

### Metabolism

In the table below, the green color corresponds to No, it is not a substrate of the listed enzymes, while the red color to Yes, it is a substrate. The same goes for the inhibitor table. The only parameter that differs is the inhibition of CYP3A4, which for SwissADME in the case of Geldanamycin, Alvespimycin, and Rifabutin corresponds to a no, contrary to what pkCSM and Simulation Plus reveal. The drug with the highest CYP Risk score is Rifabutin.

Cytochrome P450 (CYP2D6/CYP3A4) is a vital detoxifying enzyme found mostly in the liver. It oxidizes xenobiotics to aid in their elimination. Many medications are destroyed by cytochrome P450s, whereas others might be stimulated by them. Inhibitors of this enzyme can interfere with medication metabolism and are thus not recommended. As a result, it is critical to evaluate a compound's



potential to inhibit the cytochrome P450. If the dose required to achieve 50% inhibition is less than 10  $\mu$ M, a substance is termed cytochrome P450 inhibitor.

Compounds	CYP1A2	CYP2C19	CYP2C9	CYP2D6	CYP3A4
Geldanamycin	●	●	●	●	●
Alvespimycin	●	●	●	●	●
Tanespimycin	●	●	●	●	●
Rifabutin	●	●	●	●	●

Table 7: Substrates of the enzymes CYP1A2, CYP2C19, CYP2C9, CYP2D6 and CYP3A4.

Compounds	CYP1A2	CYP2C19	CYP2C9	CYP2D6	CYP3A4
Geldanamycin	●	●	●	●	● ▲
Alvespimycin	●	●	●	●	● ▲
Tanespimycin	●	●	●	●	●
Rifabutin	●	●	●	●	● ▲

Table 8: Inhibitors of the enzymes CYP1A2, CYP2C19, CYP2C9, CYP2D6 and CYP3A4.

Legend: ▲ SwissADME, ● Simulation Plus

Cytochrome P450 (CYP2D6/CYP3A4) is a vital detoxifying enzyme found mostly in the liver. It oxidizes xenobiotics to aid in their elimination. Many medications are destroyed by cytochrome P450s, whereas others might be stimulated by them. Inhibitors of this enzyme can interfere with medication metabolism and are thus not recommended. As a result, it is critical to evaluate a compound's potential to inhibit the cytochrome P450. If the dose required to achieve 50% inhibition is less than 10  $\mu$ M, a substance is termed cytochrome P450 inhibitor.

#### Excretion

Compounds	pkCSM
Geldanamycin	0.95
Tanespimycin	1.08
Alvespimycin	-0.42
Rifabutin	-0.94

Table 9: Total clearance.

Drug clearance occurs predominantly as a mix of hepatic clearance (liver metabolism and biliary clearance) and renal clearance (excretion via the kidneys). Total Clearance is measured in  $\text{log}(\text{mL}/\text{min}/\text{kg})$ . The prediction results show that the total clearance of Tanespimycin is the highest followed by Geldanamycin, Alvespimycin and Rifabutin, this indicates that bioavailability of Tanespimycin is the highest.

Compounds	pkCSM
Geldanamycin	No
Tanespimycin	No
Alvespimycin	No
Rifabutin	No

Table 10: Renal OCT2 substrate.

The results also reveal that all drugs may not be substrates for the organic cation transporter 2 (OCT2), a renal cation transporter that plays a key role in drug elimination via the kidney. Based on the above results, we may conclude that the demons travel via the kidneys in a mechanism other than OCT2.

### *Toxicity*

Since SwissADME does not give toxicity estimations, only Simulations Plus and pkCSM are reported. Rifabutin, unlike the other three compounds, induces skin sensitivity, which can result in allergic responses such as allergic contact dermatitis. Rifabutin and Alvespimycin, on the other hand, produce respiratory sensitivity.

They all exhibit reproductive toxicity, which is a critical regulatory endpoint classified as developmental toxicity. Reproductive toxicity refers to any parameter that interferes with an organism's reproductive methods, such as negative effects on sexual organs, performance, ease of conception, and any developmental toxicity suffered by the progeny.

None of the compounds are hERG I and II inhibitor, but all of them cause hepatotoxicity and reproductive toxicity.

In terms of Ames toxicity, pkCSM forecasts that no molecule would be mutagenic, but Simulations Plus predicts that Rifabutin and Alvespimycin will be Ames positive. Three of them are suspected of being mutagenic, namely Geldanamycin, Rifabutin and Tanespimycin.

The predicted concentration in units of  $\text{mg}/\text{L}$  of a certain substance that will kill 50% of a population of minnows after 96 hours of exposure is shown in the table below.

<b>Compounds</b>	<b>pkCSM</b>	<b>Simulation Plus</b>
Geldanamycin	4.797	0.0030
Tanespimycin	4.436	0.0010
Alvespimycin	4.677	0.0020
Rifabutin	3.375	0.0001

Table 11: Minnow toxicity.

Maximum Recommended Tolerated Dose (MRTD) represents the dangerous dose threshold in humans, whether it is greater or less than 3.16 mg/kg/day ( $\approx 0.5$  in logarithmic scale).

<b>Compounds</b>	<b>pkCSM</b>	<b>Simulation Plus</b>
Geldanamycin	0.117	Below
Tanespimycin	0.210	Below
Alvespimycin	0.322	Above
Rifabutin	-0.105	Below

Table 12: Maximum Recommended Tolerated Dose.

The table below shows the risk associated with predicted toxicity traits a score in the 0-6 range indicating the number of potential toxicity problems a compound might have, with the related code connected, where rat stands for acute rat toxicity, Xr for carcinogenicity in rat, Xm for carcinogenicity in mice and MUT for Ames positive.

<b>Compounds</b>	<b>TOX_Risk_Score</b>	<b>TOX_Risk_Code</b>
Geldanamycin	1.28	Rat; Xr
Tanespimycin	2.00	Rat; Xr
Alvespimycin	2.07	Rat; Xr; MUT
Rifabutin	1.00	Xr+;Xm-

Table 13: TOX Risk predictions by Simulation Plus.

The last table shows the Full ADMET Risk scores and codes, where HBD indicates H-bond donors, HBA H-bond acceptors, ch charge, Kow lipophilicity, Peff permeability, Vd volume of distribution, rat acute rat toxicity, Xr for carcinogenicity in rat, Xm for carcinogenicity in mice and MUT for Ames positive, 1A2 high clearance by CYP1A2 and CL high microsomal clearance.

<b>Compounds</b>	<b>FULL_Risk_Score</b>	<b>FULL_Risk_Code</b>
Geldanamycin	4.95	Size; HBD; HBA; ch; Peff; rat; Xr
Tanespimycin	6.54	Size; HBD; HBA; ch; Peff; rat; Xr
Alvespimycin	7.03	Size; HBD; HBA; ch; Peff; rat; Xr;; MUT
Rifabutin	8.50	Size; HBD; HBA; ch; Peff; rat; Xr; Xm+; 2D6-; 3A4+; CL-; Vd-; Kow+

Table 14: Full ADMET Risk predictions by Simulation Plus.

## CONCLUSION

Geldanamycin (GA) is the progenitor of Hsp90 inhibitors. This chemical can bind to the ATP site at the N-terminal domain of Hsp90, blocking ATP binding and disrupting the ATP-dependent conformation of a wide range of client proteins involved in signal transmission, cell cycle control, and hormone response.

Although GA has strong anticancer benefits, it also has severe hepatotoxicity; also, preclinical animal tests have revealed low solubility. GA similar chemicals have been developed to address these issues, despite producing positive outcomes in terms of binding affinity in molecular docking.

As previously stated, the methoxy group in position C-17 is a likely reason of GA toxicity since it is particularly reactive to the nucleophilic groups found in biological molecules. By replacing an allylamine group for this group, 17-allylamino-17-dimethoxygeldanamycin (17-AAG), or Tanespimycin, was developed, which is less reactive towards nucleophilic groups and less hepatotoxic. Furthermore, 17-AAG has been found to impair Hsp90's chaperone function and to cause ubiquitin-dependent proteasomal degradation of a number of oncoproteins.

In preclinical studies, these Hsp90 inhibitors demonstrated clear chemo preventive effects against a variety of tumor cell lines, accelerated the degradation of multiple oncogenic Hsp90 client proteins. Some Hsp90 inhibitors, such as 17-AAG, have entered phase II clinical studies.

However, numerous clinical studies using Hsp90 inhibitors indicated a variety of deleterious consequences, as evidenced by pharmacokinetic property analyses.

It is the one that exhibited the best values in terms of binding affinity following the molecular docking procedure, according to the results of this thesis work.

LigID	Site	Score
Tanespimycin	autodock.site1	-13.320

Table 15: Best molecular docking score.

The administration of geldanamycin and 17-allylamino-17-demethoxygeldanamycin (17-AAG), which is Alvespimycin, as Hsp90 inhibitors has run into a number of issues, including hepatotoxicity and formulation issues, necessitating the continuous research of novel molecules. Preclinical studies demonstrate that 17-des-methoxy-17-N,N-dimethylaminoethylamino-geldanamycin (17-DMAG) is efficacious against breast cancer, lung cancer, and melanoma xenografts, is orally active, and has high bioavailability. The crystal structure of the N-terminal domain of human Hsp90 alpha complexed with 17-DMAG reveals the compound's precise interactions with the ATP binding site. Besides that, simulations of the conformational changes that convert the macrocyclic ring from free to bound suggest that a geldanamycin analog with

a constrained cis-amide bond in the ground state would bind without the large energy and entropy loss required for the protein-induced conformational change, resulting in a significant increase in affinity. These findings provide a structural framework for the creation of conformationally restricted Hsp90 inhibitors [59]. Considering that 17-AAG has a lower toxicity profile than GA and it is also moderately soluble, it is a more appealing clinical prospect.

Rifabutin has demonstrated that in the molecular docking process, it binds preferably in the C-terminal domain of Hsp90, rather than the ATP binding pocket in the N-terminal domain of Hsp90, thus distinguishing itself from the other inhibitors of Hsp90 alpha illustrated in this thesis work.

However, It is just minimally soluble and also shows the highest full ADMET risk score. Further validation might come from investigations on the biological effectiveness of Rifabutin's activities as a Hsp90 inhibitor using various biological approaches illustrated in this thesis work. Following molecular dynamics analyses, this drug was also shown to attain superior stability in terms of conformational convergence than the other compounds evaluated.

<b>Compounds</b>	<b>FULL_Risk_Score</b>	<b>FULL_Risk_Code</b>
Rifabutin	8.50	Size; HBD; HBA; ch; Peff; rat; Xr.; Xm+; 2D6-; 3A4+; CL-; Vd-; Kow+

Table 16: Highest full ADMET risk score and the related full risk codes.

## BIBLIOGRAPHY

- [1] URL: <https://upbiotech.wordpress.com/2019/04/30/il-folding-proteico/>.
- [2] URL: <https://www.biopills.net/heat-shock-protein/>.
- [3] [http://unica2.unica.it/biotecnologie/index2.php?option=com\\_docman&task=doc\\_view&gid=381&Itemid=218](http://unica2.unica.it/biotecnologie/index2.php?option=com_docman&task=doc_view&gid=381&Itemid=218).
- [4] Pearl, L. H., & Prodromou, C. (2006). Structure and mechanism of the hsp90 molecular chaperone machinery. *Annual Review of Biochemistry*, 75(1), pp. 271–294. URL: <https://doi.org/10.1146/annurev.biochem.75.103004.142738>.
- [5] Wandinger, S. K., Richter, K., & Buchner, J. (2008). The hsp90 chaperone machinery. *Journal of Biological Chemistry*, 283(27), pp. 18473–18477. URL: <https://doi.org/10.1074/jbc.r800007200>.
- [6] Solit, D. (2008). Development and application of hsp90 inhibitors. *Drug Discovery Today*, 13(1-2), pp. 38–43. URL: <https://doi.org/10.1016/j.drudis.2007.10.007>.
- [7] Pearl, L. H., Prodromou, C., & Workman, P. (2008). The hsp90 molecular chaperone: An open and shut case for treatment. *Biochemical Journal*, 410(3), pp. 439–453. URL: <https://doi.org/10.1042/bj20071640>.
- [8] Stragliotto Stefano, "In silico development of new inhibitors of the Hsp90 chaperone of potential therapeutic interest". PhD thesis, University of Padua, 2011.
- [9] Avendaño, C., & Menéndez, J. C. (2015). Other nonbiological approaches to targeted cancer chemotherapy. *Medicinal Chemistry of Anticancer Drugs*, pp. 493–560. URL: <https://doi.org/10.1016/b978-0-444-62649-3.00011-9>.
- [10] Pacey, S., Gore, M., Chao, D., Banerji, U., Larkin, J., Sarker, S., Owen, K., Asad, Y., Raynaud, F., Walton, M., Judson, I., Workman, P., & Eisen, T. (2010). A phase II trial of 17-allylamino, 17-demethoxygeldanamycin (17-aag, tanespimycin) in patients with metastatic melanoma. *Investigational New Drugs*, 30(1), pp. 341–349. URL: <https://doi.org/10.1007/s10637-010-9493-4>.
- [11] Neckers, L. (2002). Hsp90 inhibitors as novel cancer chemotherapeutic agents. *Trends in Molecular Medicine*, 8(4). URL: [https://doi.org/10.1016/s1471-4914\(02\)02316-x](https://doi.org/10.1016/s1471-4914(02)02316-x).
- [12] Liu, D., Hu, J., Agorreta, J., Cesario, A., Zhang, Y., Harris, A. L., Gatter, K., & Pezzella, F. (2010). Tumor necrosis factor receptor-associated protein 1 (TRAP1) regulates genes involved in cell cycle and metastases. *Cancer Letters*, 296(2), pp. 194–205. URL: <https://doi.org/10.1016/j.canlet.2010.04.017>.
- [13] Caplan, A. J., Jackson, S., & Smith, D. (2003). Hsp90 reaches new heights. *EMBO Reports*, 4(2), pp. 126–130. URL: <https://doi.org/10.1038/sj.embor.embor742>.
- [14] Pernas, S., & Tolaney, S. M. (2019). Her2-positive breast cancer: New therapeutic frontiers and overcoming resistance. *Therapeutic Advances in Medical Oncology*, 11. URL: <https://doi.org/10.1177/1758835919833519>.

- [15] Kelland, L. R., Sharp, S. Y., Rogers, P. M., Myers, T. G., & Workman, P. (1999). DT-diaphorase expression and tumor cell sensitivity to 17-Allylamino,17-demethoxygeldanamycin, an inhibitor of heat shock protein 90. *JNCI Journal of the National Cancer Institute*, 91(22), pp. 1940–1949. URL: <https://doi.org/10.1093/jnci/91.22.1940>.
- [16] Gaspar, N., Sharp, S. Y., Pacey, S., Jones, C., Walton, M., Vassal, G., Eccles, S., Pearson, A., & Workman, P. (2009). Acquired resistance to 17-allylamino-17-demethoxygeldanamycin (17-aag, tanespimycin) in glioblastoma cells. *Cancer Research*, 69(5), pp. 1966–1975. URL: <https://doi.org/10.1158/0008-5472.can-08-3131>.
- [17] Kummar, S., Gutierrez, M. E., Gardner, E. R., Chen, X., Figg, W. D., Zajac-Kaye, M., Chen, M., Steinberg, S. M., Muir, C. A., Yancey, M. A., Horneffer, Y. R., Juwara, L., Melillo, G., Ivy, S. P., Merino, M., Neckers, L., Steeg, P. S., Conley, B. A., Giaccone, G., Murgo, A. J. (2010). Phase I trial of 17-dimethylaminoethylamino-17-demethoxygeldanamycin (17-DMAG), a heat shock protein inhibitor, administered twice weekly in patients with advanced malignancies. *European Journal of Cancer*, 46(2), pp. 340–347. URL: <https://doi.org/10.1016/j.ejca.2009.10.026>.
- [18] Butler, L. M., Ferraldeschi, R., Armstrong, H. K., Centenera, M. M., & Workman, P. (2015). Maximizing the therapeutic potential of hsp90 inhibitors. *Molecular Cancer Research*, 13(11), pp. 1445–1451. URL: <https://doi.org/10.1158/1541-7786.mcr-15-0234>.
- [19] O'Brien, R. J., Lyle, M. A., & Snider, D. E. (1987). Rifabutin (ansamycin LM 427): A new rifamycin-S derivative for the treatment of mycobacterial diseases. *Clinical Infectious Diseases*, 9(3), pp. 519–530. URL: <https://doi.org/10.1093/clinids/9.3.519>.
- [20] O'Brien, R. J., Geiter, L. J., & Lyle, M. A. (1990). Rifabutin (ansamycin LM427) for the treatment of pulmonarymycobacterium aviumcomplex. *American Review of Respiratory Disease*, 141(4\_pt\_1), pp. 821–826. URL: [https://doi.org/10.1164/ajrccm/141.4\\_pt\\_1.821](https://doi.org/10.1164/ajrccm/141.4_pt_1.821).
- [21] Holmberg, S. D. (1998). Possible effectiveness of clarithromycin and Rifabutin for cryptosporidiosis chemoprophylaxis in HIV disease. *JAMA*, 279(5), pp. 384. URL: <https://doi.org/10.1001/jama.279.5.384>.
- [22] Fichtenbaum, C. J., Zackin, R., Feinberg, J., Benson, C., & Griffiths, J. K. (2000). Rifabutin but not clarithromycin prevents cryptosporidiosis in persons with advanced HIV infection. *AIDS*, 14(18), pp. 2889–2893. URL: <https://doi.org/10.1097/00002030-200012220-00010>.
- [23] Trepel, J., Mollapour, M., Giaccone, G., & Neckers, L. (2010). Targeting the dynamic hsp90 complex in cancer. *Nature Reviews Cancer*, 10(8), pp. 537–549. URL: <https://doi.org/10.1038/nrc2887>.
- [24] Ren, X., Li, T., Zhang, W., & Yang, X. (2022). Targeting heat-shock protein 90 in cancer: An update on combination therapy. *Cells*, 11(16), pp. 2556. URL: <https://doi.org/10.3390/cells11162556>.



- [25] Zagouri, F., Bournakis, E., Koutsoukos, K., & Papadimitriou, C. A. (2012). Heat shock protein 90 (HSP90) expression and breast cancer. *Pharmaceuticals*, 5(9), pp. 1008–1020. URL: <https://doi.org/10.3390/ph5091008>.
- [26] Miyata, Y., Nakamoto, H., & Neckers, L. (2013). The therapeutic target hsp90 and cancer hallmarks. *Current Pharmaceutical Design*, 19(3), pp. 347–365. URL: <https://doi.org/10.2174/138161213804143725>.
- [27] URL: <https://www.icr.ac.uk/about-us/our-achievements/our-scientific-discoveries/targeting-hsp90>.
- [28] Alcazar, A., & Cid, C. (2009). High cytotoxic sensitivity of the oligodendrocyte precursor cells to hsp90 inhibitors in cell cultures. *Experimental Neurology*, 216(2), pp. 511–514. URL: <https://doi.org/10.1016/j.expneurol.2008.12.022>.
- [29] Chen, L., Li, J., Farah, E., Sarkar, S., Ahmad, N., Gupta, S., Larner, J., & Liu, X. (2016). Cotargeting hsp90 and its client proteins for treatment of prostate cancer. *Molecular Cancer Therapeutics*, 15(9), pp. 2107–2118. URL: <https://doi.org/10.1158/1535-7163.mct-16-0241>.
- [30] Gupta, A., Bansal, A., & Hashimoto-Torii, K. (2020). HSP70 and hsp90 in neurodegenerative diseases. *Neuroscience Letters*, 716. URL: <https://doi.org/10.1016/j.neulet.2019.134678>.
- [31] McLean, P. J., Klucken, J., Shin, Y., & Hyman, B. T. (2004). Geldanamycin induces HSP70 and prevents  $\alpha$ -synuclein aggregation and toxicity in vitro. *Biochemical and Biophysical Research Communications*, 321(3), pp. 665–669. URL: <https://doi.org/10.1016/j.bbrc.2004.07.021>.
- [32] Kearsley, S. K., Underwood, D. J., Sheridan, R. P., & Miller, M. D. (1994). Flexibase: A way to enhance the use of molecular docking methods. *Journal of Computer-Aided Molecular Design*, 8(5), pp. 565–582. URL: <https://doi.org/10.1007/bf00123666>.
- [33] Friesner, R. A., Banks, J. L., Murphy, R. B., Halgren, T. A., Klicic, J. J., Mainz, D. T., Repasky, M. P., Knoll, E. H., Shelley, M., Perry, J. K., Shaw, D. E., Francis, P., & Shenkin, P. S. (2004). Glide: A new approach for rapid, accurate docking and scoring. 1. method and assessment of Docking Accuracy. *Journal of Medicinal Chemistry*, 47(7), pp. 1739–1749. URL: <https://doi.org/10.1021/jm0306430>.
- [34] Zsoldos, Z., Reid, D., Simon, A., Sadjad, S. B., & Johnson, A. P. (2007). EHiTS: A new fast, exhaustive flexible ligand docking system. *Journal of Molecular Graphics and Modelling*, 26(1), pp. 198–212. URL: <https://doi.org/10.1016/j.jmglm.2006.06.002>.
- [35] Wang, Q., & Pang, Y.-P. (2007). Preference of small molecules for local minimum conformations when binding to proteins. *PLoS ONE*, 2(9). URL: <https://doi.org/10.1371/journal.pone.0000820>.
- [36] Klebe, G., & Mietzner, T. (1994). A fast and efficient method to generate biologically relevant conformations. *Journal of Computer-Aided Molecular Design*, 8(5), pp. 583–606. URL: <https://doi.org/10.1007/bf00123667>.

- [37] Cerqueira, N. M., Bras, N. F., Fernandes, P. A., & Ramos, M. J. (2009). Madamm: A multistaged docking with an automated Molecular modeling protocol. *Proteins: Structure, Function, and Bioinformatics*, 74(1), pp. 192–206. URL: <https://doi.org/10.1002/prot.22146>.
- [38] Totrov, M., & Abagyan, R. (2008). Flexible ligand docking to multiple receptor conformations: A practical alternative. *Current Opinion in Structural Biology*, 18(2), pp. 178–184. URL: <https://doi.org/10.1016/j.sbi.2008.01.004>.
- [39] Hartmann, C., Antes, I., & Lengauer, T. (2009). Docking and scoring with alternative side-chain conformations. *Proteins: Structure, Function, and Bioinformatics*, 74(3), pp. 712–726. URL: <https://doi.org/10.1002/prot.22189>.
- [40] Taylor, R. D., Jewsbury, P. J., & Essex, J. W. (2003). FDS: Flexible ligand and receptor docking with a continuum solvent model and soft-core energy function. *Journal of Computational Chemistry*, 24(13), pp. 1637–1656. URL: <https://doi.org/10.1002/jcc.10295>.
- [41] Feig, M., Onufriev, A., Lee, M. S., Im, W., Case, D. A., & Brooks, C. L. (2003). Performance comparison of generalized born and Poisson methods in the calculation of electrostatic solvation energies for protein structures. *Journal of Computational Chemistry*, 25(2), pp. 265–284. URL: <https://doi.org/10.1002/jcc.10378>.
- [42] URL: <http://www.garfield.library.upenn.edu/classics1986/A1986D404600001.pdf>
- [43] Ashton, M. J., Jaye, M. C., & Mason, J. S. (1996). New Perspectives in lead generation II: Evaluating molecular diversity. *Drug Discovery Today*, 1(2), pp. 71–78. URL: [https://doi.org/10.1016/1359-6446\(96\)89091-x](https://doi.org/10.1016/1359-6446(96)89091-x)
- [44] Lipinski, C. A., Lombardo, F., Dominy, B. W., & Feeney, P. J. (2001). Experimental and computational approaches to estimate solubility and permeability in drug discovery and development settings 1PII of original article: S0169-409X(96)00423-1. the article was originally published in *Advanced Drug Delivery Reviews* 23 (1997) 3–25. 1. *Advanced Drug Delivery Reviews*, 46(1-3), pp. 3–26. URL: [https://doi.org/10.1016/s0169-409x\(00\)00129-0](https://doi.org/10.1016/s0169-409x(00)00129-0)
- [45] Leeson, P. D., & Springthorpe, B. (2007). The influence of drug-like concepts on decision-making in medicinal chemistry. *Nature Reviews Drug Discovery*, 6(11), pp. 881–890. URL: <https://doi.org/10.1038/nrd2445>
- [46] Hill, A. P., & Young, R. J. (2010). Getting physical in drug discovery: A contemporary perspective on solubility and hydrophobicity. *Drug Discovery Today*, 15(15-16), pp. 648–655. URL: <https://doi.org/10.1016/j.drudis.2010.05.016>
- [47] Young, R. J., Green, D. V. S., Luscombe, C. N., & Hill, A. P. (2011). Getting physical in Drug Discovery II: The impact of chromatographic hydrophobicity measurements and aromaticity. *Drug Discovery Today*, 16(17-18), pp. 822–830. URL: <https://doi.org/10.1016/j.drudis.2011.06.001>

- [48] Gleeson, M. P. (2008). Generation of a set of simple, interpretable ADMET rules of thumb. *Journal of Medicinal Chemistry*, 51(4), pp. 817–834. URL: <https://doi.org/10.1021/jm701122q>
- [49] Veber, D. F., Johnson, S. R., Cheng, H.-Y., Smith, B. R., Ward, K. W., & Kopple, K. D. (2002). Molecular properties that influence the oral bioavailability of drug candidates. *Journal of Medicinal Chemistry*, 45(12), pp. 2615–2623. URL: <https://doi.org/10.1021/jm020017n>
- [50] Bickerton, G. R., Paolini, G. V., Besnard, J., Muresan, S., & Hopkins, A. L. (2012). Quantifying the chemical beauty of drugs. *Nature Chemistry*, 4(2), 90–98. <https://doi.org/10.1038/nchem.1243>
- [51] URL: [http://chim.it/sites/default/files/chimind/pdf/2018\\_2\\_16\\_ca.pdf](http://chim.it/sites/default/files/chimind/pdf/2018_2_16_ca.pdf)
- [52] Edelsbrunner, H., Kirkpatrick, D., & Seidel, R. (1983). On the shape of a set of points in the plane. *IEEE Transactions on Information Theory*, 29(4), pp. 551–559. URL: <https://doi.org/10.1109/tit.1983.1056714>.
- [53] Soga, S., Shirai, H., Kobori, M., & Hirayama, N. (2007). Use of amino acid composition to predict ligand-binding sites. *Journal of Chemical Information and Modeling*, 47(2), pp. 400–406. URL: <https://doi.org/10.1021/ci6002202>.
- [54] Stahura, F., & Bajorath, J. (2005). New methodologies for ligand-based virtual screening. *Current Pharmaceutical Design*, 11(9), pp. 1189–1202. URL: <https://doi.org/10.2174/1381612053507549>.
- [55] Ripphausen, P., Nisius, B., & Bajorath, J. (2011). State-of-the-art in ligand-based virtual screening. *Drug Discovery Today*, 16(9-10), pp. 372–376. URL: <https://doi.org/10.1016/j.drudis.2011.02.011>.
- [56] Preto, J., & Gentile, F. (2019). Assessing and improving the performance of consensus docking strategies using the DockBox package. *Journal of Computer-Aided Molecular Design*, 33(9), pp. 817–829. URL: <https://doi.org/10.1007/s10822-019-00227-7>.
- [57] Pinzi Luca, “Computational approaches in polypharmacology”. PhD thesis, University of Modena and Reggio Emilia, 2018.
- [58] Daina, A., Michielin, O., & Zoete, V. (2017). SwissADME: A free web tool to evaluate pharmacokinetics, drug-likeness and medicinal chemistry friendliness of small molecules. *Scientific Reports*, 7(1). URL: <https://doi.org/10.1038/srep42717>.
- [59] Jez, J. M., Chen, J. C.-H., Rastelli, G., Stroud, R. M., & Santi, D. V. (2003). Crystal structure and molecular modeling of 17-DMAG in complex with human hsp90. *Chemistry & Biology*, 10(4), pp. 361–368. URL: [https://doi.org/10.1016/s1074-5521\(03\)00075-9](https://doi.org/10.1016/s1074-5521(03)00075-9).
- [60] URL: <http://www.swissadme.ch/index.php>.
- [61] URL: <http://biosig.unimelb.edu.au/pkcsml/prediction>.
- [62] URL: <https://pubchem.ncbi.nlm.nih.gov/>.