POLITECNICO DI TORINO

Department of Mechanical and Aerospace Engineering

MASTER'S DEGREE IN BIOMEDICAL ENGINEERING

Academic Year 2020/2021

DETERMINATION OF THE BINDING SITES AND BINDING MODES OF GATASTATIN AND ITS NOVEL DERIVATIVES

Supervisors
Prof. J. A. TUSZYNSKI
Prof. M. A. DERIU

Candidate
Paola VOTTERO

December, 2021

ABSTRACT

Glioblastoma Multiforme (GBM) is among the most aggressive forms of brain tumours. The prognosis for GBM patients remains poor despite advances in therapies, as they have proven unable to prolong patient survival more than a few months.

The World Health Organization classification of GBM is a necrosis-predisposed grade IV cancer that is mitotically active, which means its cells are proliferating at a higher rate than normal tissue cells. During mitosis prophase, the cell prepares to divide by condensing its chromosomes and starting the formation of the mitotic spindle, of which microtubules are the main component. *In vivo* nucleation of microtubules begins with a complex called γ-TuRC (γ Tubulin Ring Complex) formed by accessory proteins recalled by γ tubulin, which is therefore essential for the formation of microtubules. As γ tubulin has been found to be overexpressed in GBM, identifying a compound that can inhibit its functions could benefit patients greatly.

The present work aims at identifying the binding sites and binding modes of gatastatin, a potential γ tubulin inhibitor, and its recently developed derivatives. Computational modelling, particularly molecular docking simulations, has proven a valuable tool to explore and compare the compounds' binding to γ and α-β tubulin.

ACKNOWLEDGEMENTS

CONTENTS

CONTENTS

1

INTRODUCTION

This introductory section is an overview of the organization of the work.

1. Introduction
2. Biological Background
3. Overview on Computational Methods
4. Modelling of Human Tubulin Isotypes
5. Molecular Docking
6. ADMET Properties Analysis
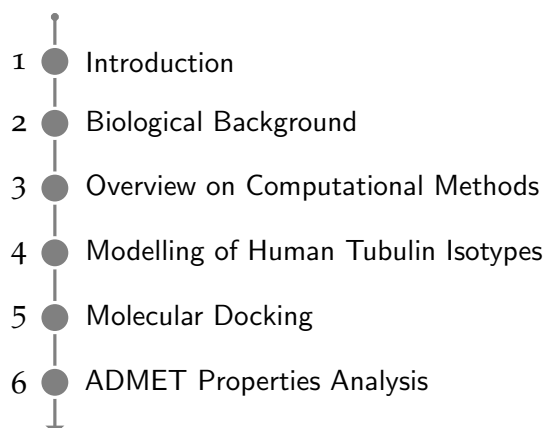
CHAPTER 1 is the present introductory section.

CHAPTER 2 describes the biological basis of the work. Microtubules structure and functions are first discussed, focusing on their essential role in cell mitosis which makes them a long-studied target for chemotherapics. The link between microtubules and γ tubulin is then explored, together with the potential for therapeutic applications targeting γ tubulin, especially for potential treatment of Glioblastoma Multiforme. Lastly, a brief overview of current therapeutic strategies in the treatment of GBM is provided.

CHAPTER 3 focuses on computational methods, with particular attention to homology modelling, molecular docking, and ADMET properties prediction. A general description of these methods is provided in this section.

CHAPTER 4 goes into the details of homology modelling and reviews the steps required to obtain human tubulin isotype models. An analysis of the differences between human and animal β tubulin in the colchicine binding site is provided.

CHAPTER 5 is devoted to describing the adopted molecular docking approaches, namely consensus and blind docking, and to analyse the scores of the resulting poses. The consensus docking protocol is explained in details.

CHAPTER 6 is dedicated to the analysis of the compounds' pharmacokinetics through ADMET properties analysis. The results from three different tools are compared.

CHAPTER 7 is a brief discussion on docking and ADMET prediction results. Where applicable, a comparison with available experimental data is provided.

BIOLOGICAL BACKGROUND

This section aims at providing a better understanding of the correlation between Glioblastoma Multiforme and γ tubulin and the possible therapeutic benefit that would derive from targeting it.

Microtubules (MT) are part of cytoskeletal filaments in cells. They branch off from the centrosome, near the cell nucleus, and develop as a network up to the sub-membrane region. Being connection structures, they are involved in many cell life processes as motility, signal sensing, cell organisation and structural strength, intracellular transport, and chromosome segregation during mitosis.

From a structural, biophysical and chemical point of view, cytoskeletal filaments can be defined *biopolimers*: they are assemblies of monomers (that in this case are proteins) linked by non-covalent bonds. The nature of the monomers characterises the filament: microtubules are composed by heterodimers of α and β tubulin; both are globular subunits of approximately 50 kDa of mass and around 4 nm long. The interaction between α and β tubulin inside the heterodimer is high energy, despite not being a covalent one, so chemical treatments are required to break it.

The dimers organise in proto-filaments that in turn place side by side to form a hollow cylinder structure with an external diameter of about 25 nm, as schematised in figure 1.

α and β tubulin are part of the so-called tubulin superfamily *including γ, δ, ϵ, ν, and others. Some of these tubulins are widespread among eukaryotes, but the superfamily members other than α, β and γ are still being explored [44].*
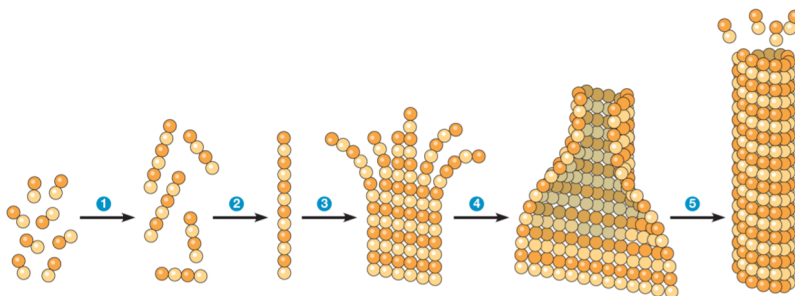
Figure 1: Microtubules assembly from α-β tubulin dimers.

Tubulin polymerises in a *head-to-tail* fashion, with the α subunit of a dimer binding the β subunit of the following one.

Therefore, MTs are polar structures with a fast-growing plus end, where the β subunit is exposed, and a slow-growing minus-end exposes α tubulin.

Microtubules polymerisation is regulated by a molecule called *guanosine triphosphate* (GTP), whose structure is shown in figure 2. It is a purine nucleoside triphosphate exerting various functions in the cell, from being an energy source (similarly to ATP) for protein synthesis and gluconeogenesis to acting as a signal transducer. Concerning microtubules, GTP is responsible for the mechanism known as *dynamic instability*, i.e. the alternation of assembly and disassembly cycles.
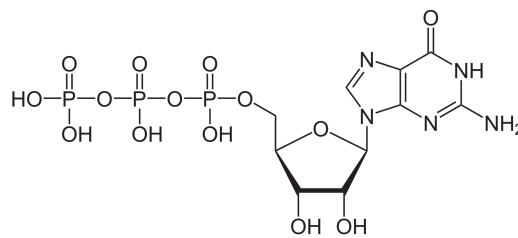
Figure 2: Chemical structure of GTP.

The α-β heterodimer hosts a GTP molecule at the dimerization interface in a non-exchangeable site. Upon formation of the dimer, another GTP binds to the exposed β subunit in the exchangeable site located at the inter-dimer interface, as shown in figure 3. This second GTP molecule, being more exposed to solvent, will be hydrolysed to GDP during or shortly after polymerisation; GTP hydrolysis is thought to provide MTs with the flexibility needed to undergo rapid turnover cycles polymerisation and depolymerisation. Indeed, hydrolysis leads to a change in the angle of the longitudinal assembly, which destabilises lateral interactions between adjacent proto-filaments. The formed microtubule structure can be stabilised by binding a GTP cap to the exposed β tubulin.

Above a critical α-β tubulin concentration, heterodimers spontaneously assemble into microtubules *in vitro*, resulting in MTs having different diameters because varying numbers of adjacent protofilaments form them.
Conversely, microtubule nucleation *in vivo* is initiated from a ring-like template of γ tubulin, which guides the polymerisation of MTs formed by exactly 13 proto-filaments by exposing as many γ tubulin units. The nucleation process likely occurs

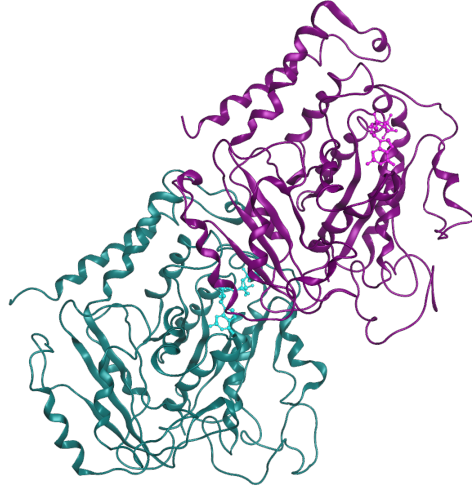*γ tubulin is another member of the tubulin superfamily. It shares 34% similarity with β tubulin.*

Figure 3: Structure of the α-β tubulin heterodimer in complex with GTP at the non-exchangeable site and GDP at the exchangeable site on β.

through the recruiting of α-β dimers via α-γ interactions, which are regulated by GTP similarly to the regulation of α-β polymerisation [33]. Thanks to the spatial organisation of the γ tubulins inside the complex, the heterodimers are arranged to support lateral contacts between forming adjacent proto-filaments. These γ tubulin complexes, called γ-TuCs, are able to promote microtubule nucleation at concentrations below those required for spontaneous assembly; moreover, they define the cellular sites for microtubule nucleation and stabilise the MT minus end (the one exposing α tubulin).

γ tubulin recruits the so-called γ tubulin complex proteins (GCPs), GCP2 and GCP3 to form a tetrameric 2:1:1 complex called the small γ tubulin complex, or γ-TuSC. In higher eukaryotes, γ-TuSC recruits other GCP paralogues, namely GCP4, 5, and 6, to form the stable γ tubulin ring complex (γ-TuRC), whose structure is shown in figure 4.

Among their various functions, perhaps the most critical role of microtubules is the formation of the mitotic spindle, as they comprise the most abundant components of the spindle apparatus. During mitosis, the already duplicated chromosomes condense and attach to spindle fibres that pull one copy of each chromosome to opposite sides of the cell, resulting in two daughter nuclei. The cell may then divide by cytokinesis to produce two daughter cells. During the prophase of the mitosis, the centrosome duplicates into two microtubules organising centres,
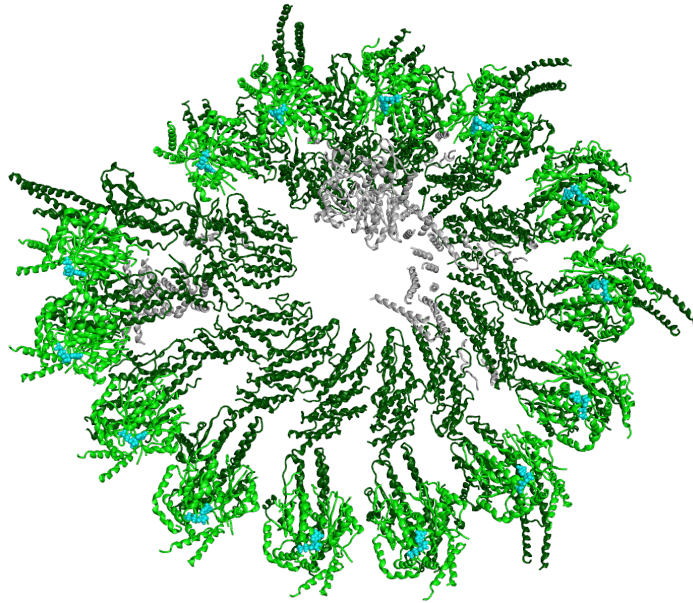
Figure 4: Model of γ-TuRC from PDB entry 6V6S [1]; GCPs, γ tubulin units, and GDP molecules are depicted.

which recruit γ-TuRC units in the pericentrosomal regions to polymerise MTs and stabilise their minus end so that they are anchored to the centrosome.

An increase in tubulin turnover during mitosis indicates that microtubule dynamics is pivotal to the mitotic process: their ability to rapidly assemble and disassemble is key to the formation of the mitotic spindle, as the microtubules array present in interphase cells disassembles and the free tubulin subunits are recruited by γ-TuRC and polymerise to form the spindle [2, 3].

Given their essential role in cell duplication, both microtubules and free tubulin have long been targeted by chemotherapeutic anti-mitotic agents that act by disrupting normal MTs dynamics by either stabilising or destabilising them. Any failure in microtubule assembly during spindle formation or the subsequent chromosome segregation phase leads typically to mitotic arrest and cell death.

Some examples of well-known anti-mitotic agents are taxanes and *Vinca* alkaloids; both target primarily β tubulin. Several human β tubulin isotypes exist: information from the Human Protein Atlas [4] on their expression and tissue distribution in normal cells is summarised by table 3.

The taxanes in clinical use include paclitaxel (*Taxol*, by Bristol-Myers Squibb), which is a natural product, and docetaxel (*Tax-*

otere, by Sanofi-Aventis), a semi-synthetic analogue. These drugs exert an anticancer effect by stabilising MTs [50].

Vinca alkaloids such as *vinblastine*, *vinorelbine*, *vindesine*, and *vincristine* are naturally extracted from the Madagascar periwinkle plant. Their mechanism of action is to arrest dividing cells in metaphase by binding free tubulin and preventing its polymerization into microtubules [54].
Lastly, *colchicine* is a natural pharmaceutical alkaloid. Despite its clinical use in cancer treatment being limited by toxicity, it can block mitotic cells in metaphase. Unlike taxanes, colchicine's mechanism of action is microtubule destabilisation: it binds to free tubulin heterodimers (establishing most of the interactions with β tubulin), forming a stable complex that at low concentrations arrests microtubule growth and at higher concentrations promotes microtubule depolymerisation.

These compounds are effective anticancer agents, as they tend to impact dividing cells more. However, being the structural components of microtubules, α and β tubulin are abundant proteins and make approximately 2.5% of the cell's total protein content, so these agents affect both cancerous and normal cells by binding tubulin indiscriminately and lead to often severe side effects. These unwanted effects may be reduced by increasing the drugs' specificity for tubulin within cancerous cells only.

For some tumours, a possible way to discriminate between normal and cancerous cells may be to target γ tubulin instead of the α-β dimer: since γ tubulin exerts an essential role in mitosis, inhibiting its functions would have an anti-mitotic effect as well. γ has been found to be overexpressed in *glioblastoma multiforme* and breast lesions and carcinomas, while in normal tissue cells, it is less abundant than α-β, making less than 1% of the total tubulin content of the cell. Therefore, a compound having a strong specificity for γ tubulin would have fewer off-target interactions.

In their study from 2015, Chinen *et al.* identified a potential γ tubulin inhibitor. Based on γ and β tubulin similarity, they screened a library of β tubulin colchicine site-binding compounds, namely colchicine itself, nocodazole, plinabulin, and glaziovianin A (AG1), to assess whether some of them were able to bind γ tubulin as well. AG1 was found to bind to both γ and β tubulin with similar affinity, so it was derivatised to look for selective γ tubulin binders. The AG1 derivative *gatastatin* seemed to bind to γ tubulin with a 12-fold higher affinity than

for the α-β dimer. The results of their drug-binding assay for colchicine, AG1 and gatastatin are shown in table 1. The lower the dissociation constant K_d, the higher the binding affinity; K_d values were calculated from changes in tryptophan fluorescence of the proteins from three experiments. Results were then averaged, and the table reports the mean value $\pm$ standard deviation.

Chinen *et al.* suggested that gatastatin's mode of action is blocking GTP binding to γ tubulin, thus inhibiting its microtubule-nucleation activity [21].

Table 1: Results of the drug-binding analysis by Chinen *et al.*

Compound	$K_d{}^{\alpha\beta}$ (μM)	$K_d{}^{\gamma}$ (μM)	$K_d{}^{\alpha\beta}/K_d{}^{\gamma}$
Colchicine	17.5 ± 2.7	196.4 ± 47.9	0.09
AG1	51.9 ± 36.4	85.3 ± 22.8	0.61
Gatastatin	42.5 ± 36.7	3.6 ± 1.3	11.81

Microscale thermophoresis is a technology for the biophysical analysis of interactions between biomolecules, based on the detection of a temperature-induced change in fluorescence of a target as a function of the concentration of a non-fluorescent ligand.

A *microscale thermophoresis* assay comparing colchicine and gatastatin binding to α-β tubulin was also performed in 2019 by Qian Wang[1]. Three and six replicates were performed for colchicine and gatastatin, respectively, and their averaged values are reported by figure 5. According to these results, colchicine and gatastatin bind the tubulin heterodimer with higher affinities than what was reported by Chinen *et al.*, but their ratio is similar in the two cases. Results of the two studies are compared in table 2.
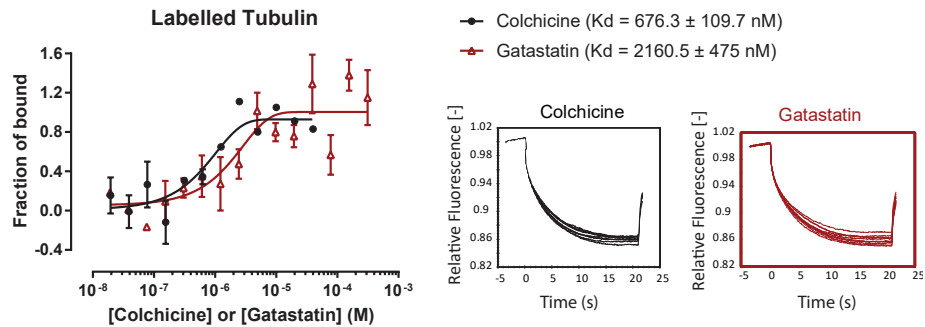

Figure 5: Results of the drug-binding assay by Q. Wang.

A previous study by Friesen *et al.* had already investigated the binding of colchicine-site binders to γ tubulin and, conversely to the results by Chinen *et al.*, they found that colchicine

---

[1] *Personal communication.*

Table 2: Comparison of drug-binding assays results.

Compound	$K_d{}^{\alpha\beta}$ (μM)	
	Chinen *et al.*	Qian Wang
Colchicine	17.5 ± 2.7	0.676 ± 0.11
Gatastatin	42.5 ± 36.7	2.16 ± 0.47

seems to bind with equal affinity to α-β and γ tubulin. Based on the high structure similarity (75%) between β and γ tubulin, they also identified a homologous region on γ tubulin corresponding to the binding site of colchicine on β tubulin [31].

Two main isoforms of human γ tubulin exist: γI, encoded by gene TUBG1, and γII, encoded by gene TUBG2. They share over 97% sequence identity, so any pharmacological agent targeting one isoform would likely target both [60]. Mouse studies by Yuba-Kubo *et al.* showed that mouse γ tubulin-expressing genes are orthologs of human TUBG1 and TUBG2, and it emerged from knockout experiments that γI, conversely to γII, is essential to proper mitotic division [61]. In addition to this, the studies by Katsetos *et al.* showed that both TUBG1 and TUBG2 are overexpressed in glioblastoma [37, 39], therefore the analysis of the binding sites will be focused entirely on tubulin γI.

Katsetos *et al.* found that also tubulin βIII is overexpressed and forms complexes with γ tubulin in glioblastomas, and suggested that aberrant expression and interactions of tubulins γ and βIII may be linked to malignant changes in glial cells [38].

Moreover, a differential gene expression study by Drs A. Wang and G. Zhang found isotypes βIIa and IVa to be differentially expressed in glioblastoma compared to healthy brain tissue. They also analysed the protein-protein interactions (PPIs) of the differentially expressed genes products and identified a protein interaction network made by thirty proteins, among which tubulin βIIa and βIVa are predicted to be relatively important [58].

PPIs are high specificity physical contacts between two or more protein molecules.

More on GBM

Gliomas are the most commonly occurring form of brain tumour; GBM is the most malignant one. The typical onset age is around 62-64 years, but it can occur at any age, including paediatric populations. The vast majority of cases, approximately 90%,

9

occur *de novo* as a primary tumour, while less than 5% of cases are secondary tumours progressing from different neural cells.

Despite considerable research investment over the last thirty years, the life expectancy of GBM patients has not been improved, and the 5-year survival rate is below 5%.

The tumour heterogeneity and its location in the brain make its treatment complex and very costly. To date, only a few therapies are approved for treating glioblastoma; current approaches include surgical resection followed by chemo and radiotherapy. Due to the risk of postoperative neurologic deficits and the infiltrative tendency of the disease, complete tumour resection is often not possible. Together with its aggressivity and the development of drug resistance, this leads to tumour recurrence.

Currently, the pharmacological treatment of choice is temozolomide (TMZ), an oral drug approved in 2005, but in the past also taxanes have been proposed [42]. TMZ is an alkylating agent that targets DNA and induces its methylation, leading to cell death. Being a small and lipophilic compound, TMZ is a rather efficacious treatment because it is able to cross the blood-brain barrier. However, a large percentage of patients do not respond favourably to it.

Many strategies have been explored to improve brain drug delivery, such as chemical modification, efflux transporter inhibition, direct intra-tumoural injections, and blood-brain barrier disruption through ultrasound. The major challenge in using these approaches is to deliver the drug successfully without compromising brain function. Other emerging strategies include nanomedicines, and immunotherapy [36].

Table 3: β tubulin isotypes distribution.

Isotype	Encoding gene	Tissue specificity	Cell type specificity
βI	TUBB	Constitutive	All cells
βIIa	TUBB2A	Brain and nerves	Horizontal cells and suprabasal keratinocytes
βIIb	TUBB2B	Brain and nerves	Müller glial cells, horizontal cells, distal tubular cells
βIII	TUBB3	Brain and nerves	Neurons
		Testis	Late spermatids
		Pancreas	Endocrine and epithelial cells
βIVa	TUBB4A (or TUBB4)	Brain	Horizontal and Müller glial cells
		Testis	Sertoli cells
		Adrenal gland	
βIVb	TUBB4B (or TUBB2C)	Testis	Early and late spermatids spermatocytes
βV	TUBB6	Low tissue specificity	Trophoblasts (placenta)
βVI	TUBB1	Blood, lymphoid tissue	Erythroid cells
βVIII	TUBB8	Blood, testis	Germ cells

OVERVIEW ON COMPUTATIONAL METHODS

This section presents an overview on some computational methods and their applications, followed by a more detailed discussion on the ones that were most used during the project.

3.1 ● Molecular Mechanics

3.2 ● Homology Modelling

3.3 ● Molecular Docking

3.4 ● ADMET Properties
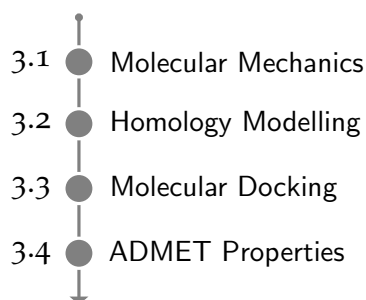
The advantage of computational models is that they maximise the gathering of information. The avail of a model is not precisely reproducing reality but providing means of understanding the mechanisms that bring to a particular result.

Biological systems are defined to be complex. The term *complexity* indicates the presence of *emergent properties*, which depend on the interaction between the parts composing the system: the outcome of a complex system thus depends on the interaction between its components. It is unpredictable unless one can model both its parts and their interactions in a dynamic way.

Complex systems organise hierarchically. Molecular modelling encompasses theoretical and computational methods to predict a system's macroscopic properties by describing its organisation and behaviour at a microscopic level.
The common feature of these methods is to describe molecular systems at an atomistic level.
Atomistic resolution models may be accomplished by:

- Explicitly modelling sub-atomic particles. At the quantum level, the model has a resolution that allows the evaluation of electronic interaction. The most accurate solution is obtained by solving Schrödinger equation (the so-called *ab initio* method). However, this is nearly impossible for any system that is larger than the hydrogen atom. Opportune

> simplifications make small structures (several hundreds of atoms) manageable.
>
> - Treating atoms as the fundamental unit. This approach is known as *Molecular Mechanics* (MM). It uses Newton's mechanics to model molecular systems by applying classical dynamics and kinematics to mass-spring systems (which model atoms and bonds, respectively).
>   This approach has the advantage that one can use it to study small molecules as well as large biological assemblies, from hundreds of thousands to millions of atoms.

The immediate extension of Molecular Mechanics is *Molecular Dynamics* (MD), which allows the analysis of the dynamic evolution of a system. It is based on the resolution of Newton's motion equations for a system of interacting atoms.

MD is a computational approach to calculate the mean properties of a system by sampling *microstates* over time in a specific *statistical ensemble*. A microstate is a specific arrangement of a molecular system: a system's mechanical or microscopic state is defined by atom locations (either Cartesian or internal generalised coordinates) and momenta. These coordinates exist in a multi-dimensional space called *phase space*. A single point in the phase space describes the state the system is in. A set of points in the phase space that satisfy a specific thermodynamic state is called a thermodynamic or statistical ensemble.

Figure 6 summarises the main steps of a molecular dynamics simulation. New positions, velocities and a new potential energy value are calculated by integrating Newton's equation starting from the initial positions, velocity distribution and potential energy functions of the atoms. Molecular dynamics is thus a deterministic method, which means that its present state entirely determines the system's future state. The output of the simulation is a trajectory. Different trajectories exist for a single system; the basic idea is that if the simulation is long enough (or multiple simulations with different initial conditions are performed), all these trajectories are able to sample the phase space. By getting all possible trajectories, one can identify the ones that are crossed by a higher number of trajectories as the most probable states that will contribute the most in determining the macroscopic properties of the system.

For proteins, the system's initial configuration is obtained from experimental data of X-ray crystallography or nuclear magnetic resonance. These data are collected in databases such as the *Protein Data Bank* (PDB). On a practical level, to have a model
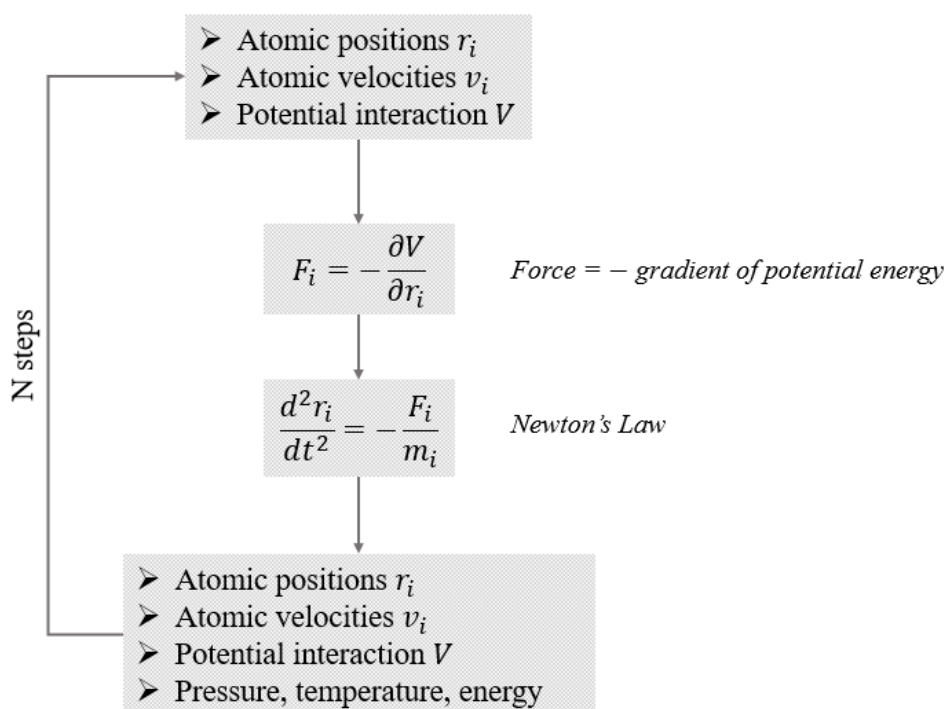
Figure 6: Molecular dynamics flowchart.

of a protein's structure means identifying the positions of all its atoms with respect to an arbitrary reference. PDB format files (with .pdb extension) are widely used to store information on the molecular structure of proteins. A typical .pdb file contains a header with bibliographic information, then lists Cartesian coordinates of all atoms and data on how the atoms are connected, as shown in figure 7. HETATM means *heteroatom* and indicates atoms not belonging to the protein but, for example, to a ligand (or solvent). The second column enumerates atoms, the third one identifies them, and the fourth one identifies the monomer the atom belongs to. The last three columns are atomic coordinates expressed in Å. The CONECT label introduces the section of the PDB file that describes the structure's topology, i.e. how atoms are bound. Atom 1, for example, is bound to atoms 2, 6 and 13.

If one wants to analyse a protein that is a potential target for clinical application, it is best to work on the human protein structure. However, it is not always available: the structures of some proteins are hard to resolve. If the protein sequence is known, one can overcome this problem by creating a homology

```
HETATM  1  C   1  −1.576  −0.433   0.004
HETATM  2  C   2  −1.301   0.777   0.643
HETATM  3  C   3  −0.072   1.410   0.444
HETATM  4  C   4   0.898   0.848  −0.394
etc.etc.

CONECT  1  2  6  13
CONECT  2  1  3  14
etc. etc.

END
```

Figure 7: Example of PDB file main sections.

model of the protein. *Homology modelling* is also known as comparative modelling; its purpose is to obtain an atomic-resolution model of the target protein starting from its aminoacidic sequence and an experimental three-dimensional structure of a related homologous protein.

Once the target structure is known, several methods can be applied to analyse compounds that can be potential therapeutic agents and their interactions with the target. *Molecular Docking* is a computational approach to predict the most probable binding mode of a compound, referred to as *ligand*, in a protein called *receptor*. Its two main goals are to predict the geometry of the interaction, i.e. the binding pose, and estimate the binding energy.

If different potential drug candidates are available, they can be analysed based on their pharmacokinetics. *Pharmacokinetics* (PK) is the study of the disposition of a drug; in simpler words, it is "what the body does to the drug" upon administration: the pharmacokinetics of a drug determines whether it is able to get to the site of action. It includes the so-called ADME processes:

- **A**bsorption;
- **D**istribution in blood plasma and in tissues;
- **M**etabolism or biotransformation: it transforms a molecular species into another one;
- **E**xcretion.

These four processes, together with **T**oxicity, make up the ADMET properties that can be used to predict whether a compound would be a viable therapeutic agent. They are, in fact, key elements that determine the safety, uptake, elimination, metabolic

behaviour, and effectiveness of drugs. ADMET analysis is thus a critically important step in drug discovery and development.

3.1 MOLECULAR MECHANICS

As previously stated, molecular mechanics describes molecular systems with classical mechanics. It calculates the energy of molecules using empirical functions containing parameters. The set of parameters needed to define the potential energy of a system is called *forcefield*. The system's energy is thus described as a function of the spatial coordinates of all its constitutive particles. The forcefield also includes electrostatic interactions; this way, it is possible to evaluate the behaviour of the system with respect to a particular environment, in a specific conformation, or towards another molecule.

The most widely used forcefields for modelling macromolecules are AMBER, CHARMM and GROMOS.

3.1.1 *Potential energy function*

The particles that form the system are considered spherical masses of a given radius. Several parameters are used to define a particle, namely the geometry of the system and the atom type.
Each particle will be located in space with respect to a reference. The potential energy function depends on particles' locations in three-dimensional space and their properties, and it can be written as the sum of two fundamental contributions:

$$V = V_{bond} + V_{non-bond}$$

Where *bond* refers to *bonded interactions* and *non-bond* to *non-bonded interactions*.
Bonded interactions are parameters that model the effects of an interatomic bond. They are:

- Bonds
- Angles
- Dihedrals
- Improper dihedrals

Non-bonded interactions are:

- Van der Waals
- Coulomb forces
- Hydrogen bonds

Bonded interactions

A bonded interaction describes a covalent interatomic bond. Molecular mechanics models *bonds* as springs so that the associated potential energy term is a harmonic potential. It is a function of the bond length:

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

Where k_l is the force constant, l_0 is the reference length of the bond, i.e. the length it assumes when every other term of the forcefield is zero; l is the equilibrium bond length.
This parabolic model is a good approximation around l_0. However, the farther from the minimum, the worse it reflects reality: when atoms are separated by an infinite distance, according to the model, energy would be infinite as well, while in reality, it should increase with l then start decreasing again and come back to zero because of the bond rupture. Alternative models exist, which offer a better approximation far from the minimum, such as cubic and Morse.

Bonds model interactions between two atoms. *Angles* consider relative motion between three atoms that describe an angle on a plane. Since atom types identify bond angles, a correction term is needed that adds an energy penalty when the angle changes. Several models exist, but the most common is a harmonic potential around an equilibrium angle.

Relative motion between four bonded atoms is modelled by *dihedral* (or torsional angle) terms. They describe the relative rotation of two planes, which are identified by bonds. A sinusoidal function describes the dihedral contribute to the potential energy function.
These terms are called *proper dihedrals* because they contemplate an accessible rotation of 360° of an atom with respect to the other three. Another set of dihedrals, called *improper dihedrals*, must be considered for atoms involved in a cyclic structure. Improper dihedral terms are needed to maintain the planarity of external atoms bond to cycles. They are again modelled with a harmonic potential.
Bonded terms of the potential energy function can be written as:

$$V_{bond} = V_{bonds} + V_{angles} + V_{dihedrals} + V_{impropdihedrals}$$

Non-bonded interactions

Bonded interactions have less of an impact on the potential energy function because they are less numerous. Apart from co-valent bonds, however, each particle's behaviour is influenced by the interaction with its surroundings: if an atom is moved from its position, not only bonded interactions terms must change but also those that describe interactions with all other particles in the system. The latter are called non-bonded interactions; they model van der Waals and Coulomb electrostatic interactions.

van der Waals interactions can occur between particles not having net charge. They are short-range interactions if compared with electrostatics because their effect is strong within $1 - 2$ nm. The model describes that each particle interacts with other atoms and reaches an equilibrium distance corresponding to an energy minimum.

Short-range interactions further divide into very short-range interactions, repulsive, which avoid atoms interpenetration, and "long" (up to 1.5 nm) range, also known as London dispersion forces, which are attractive.

Lennard-Jones potential models well the physical behaviour of van der Waals forces. Some forcefields such as AMBER use a modified version of this function to model also hydrogen bonds.

Electrostatic interactions are calculated using Coulomb's law:

$$V(r) = \sum_{i=1}^{N} \sum_{j=1}^{N} \frac{q_i q_j}{4\pi\varepsilon_0 \varepsilon_r r_{ij}}$$

Electrostatic interaction is the most difficult to address in molecular mechanics simulations. Since it is a long-range interaction (acts over 10 nm or more), there will be a huge number of interactions to calculate in a molecular system, which originates a computational difficulty.

Several solutions have been proposed to make a particle "blind" beyond a certain distance in order to reduce the number of calculations:

- Plain cutoff: non-bonded interactions are set to zero beyond the cutoff distance. It is an inaccurate and, therefore, outdated method.
- Shifted or Switched potential: these approaches apply modifications to the electrostatic term of the potential energy function in order to make it null at the cutoff distance.

- Cell multipole method: exploits calculation of electric moments of a higher order as we move further away from the particle under analysis.
- Ewalds summations: it is the fastest method and therefore the most used one. Several numerical implementations exist.

3.1.2 *Environment simulation*

Biological macromolecules exist in an aqueous environment. Water has a fundamental role in determining the molecules' structure and function.

There are two main reasons why the solvent is difficult to model. The first is water's very peculiar properties that make it hard to describe with a classical model, leading to bigger errors in the model compared to macromolecules. Another reason is the huge number of water molecules to simulate.

Two different approaches have been proposed to simulate water:

- *Explicit solvent*: water molecules are explicitly simulated and included in the model. This approach is accurate but computationally expensive. One of the most widely used explicit models is the *three-site* model (such as TIP3P by Jorgensen) that describes water as composed of three particles.
- *Implicit solvent*: these methods use a continuous medium to represent water. Damping factors are then added to simulate the solute's diffusion in water. Another property the model has to recreate is the solvent's effect on the surface of the molecule it contacts: when a protein comes into contact with water, as an example, hydrophobic portions move to its core while hydrophilic ones move towards the surface. This effect is maintained in explicit models thanks to partial charges that express a polar characteristic. Implicit models can reproduce it based on *Solvent Accessible Surface Area* (SASA) or by modelling electrostatics at a continuous level.

3.1.3 *Energy minimization*

A molecular mechanics model aims to "draw" the *Potential Energy Surface* (PES). PES is a multi-dimensional function of

molecular coordinate systems. It is also called potential energy landscape.

It is essential to characterise its stationary points, particularly the *points of minimum*. There is a great number of minima in a landscape, and they can be local or global. To localise minimum points is very important because equilibrium points are where the biological system tends to work: a local minimum can correspond to a protein's active state.

An *energy minimisation algorithm* can be used to reach a minimum point that is near where the structure is on the energy landscape. Several approaches exist; they are divided into *non-derivative* and *derivative* methods.

Non-derivative methods

The SIMPLEX method is based on constructing a geometric figure having $N + 1$ interconnected vertices, where N is the dimensionality of the potential energy function under analysis. For a two-dimensional function, as an example, the figure will be a triangle. Each vertex corresponds to a specific coordinate set for which energy can be computed.
The geometric figure has a set of allowed moves such as reflection, contraption, expansion, *et cetera*, through which it approaches the minimum point. This method is not particularly efficient, so it is usually applied when the structure is far from a minimum.

Derivative methods

These methods overcome the limitation of non-derivative ones that do not contain any information on the slope. They are further divided into:

- First order derivative methods, such as *Steepest Descent* and *Conjugate Gradient* algorithms. They impose modifications in atomic coordinates to move the system towards lower energies. The starting point of each iteration is derived from the previous one.
  Steepest Descent and Conjugate Gradient are often used in sequence.
- Second order derivative methods, such as *Newton-Raphson* and *L-BFGS*. These algorithms are based on the inversion of the Hessian matrix and involve a relatively strong

approximation. The Hessian matrix contains the second-order derivatives of the potential energy. They are used for relatively small systems ($1,000 - 5,000$ atoms) because the Hessian inversion has a high computational cost.

First-order algorithms are faster, especially if the structure is far from the minimum. Second-order methods are potentially more complete because by diagonalising the Hessian matrix, one obtains information on all the function's minima but are not always applicable due to the high computational power they require.

The reached energy minimum not necessarily corresponds with an active state of the molecule. Nevertheless, it can be useful to bring the structure to a minimum point if one wants to study the dynamics of a system because there is a derivative relationship between potential energy and force. To be in an energy minimum means to have lower forces on the atoms. Forces acting on atoms are important because when the dynamics simulation begins, they define the acceleration (i.e. the speed variation) at the beginning of the trajectory, as the system starts moving from a standstill. If there are very intense forces, the atoms will have high accelerations, which can rupture the structure.

3.2 HOMOLOGY MODELLING

The primary structure of a protein, i.e. its amino acid sequence, largely defines its three-dimensional conformation. Therefore, two proteins sharing similar sequences will have similar structures and functions as the structure is oriented to a specific function.
Homology modelling is based on the assumption that a similar sequence means a similar structure. This method provides means of reliably predicting the tertiary structure of a protein.

Methodology

The comparative modelling process can be broken down into four main steps:

- Template selection
- Alignment
- Model construction

- Quality assessment

The first requirement for the creation of a homology model is a *template*, i.e. the three-dimensional structure of a protein whose sequence shares a certain amount of similarity with the target one and is, therefore, likely to have a similar structure. A good starting point is 40% sequence identity (i.e. 40% of amino acids are identical in the template and target sequences). With a sequence identity percentage below 40%, the model's applicability and accuracy decrease, and just the general folding of the target can be predicted. With higher similarity, especially above 60%, the accuracy of the model is comparable with crystallography. If the identity percentage is above 80-90%, the task becomes trivial.

The target and template sequences are aligned to locate common subsequences that are then used as guides for the modelling process: structural features of identical subsequences are maintained in the final model. More than a template can be used, as in the segment-matching approach, which divides the target into a series of short segments and matches each to its own template.

Highly variable regions such as loops are less likely to match a template and are modelled by loop modelling. Loops are the most susceptible to modelling inaccuracies that occur more frequently when the target and template have low sequence identity. Loop databases exist, but the coordinates determined by loop modelling are generally less accurate than those obtained from copying the coordinates of a known structure, especially if the loop is ten residues or longer.

Sidechains are modelled based on the template. Rotamer libraries can be used for this purpose; rotamers are positional isomers of side chains with the right torsional angles for the backbone torsions.

Lastly, an energy minimisation step can be helpful to relieve strains, solve atom clashes, and bring the structure towards an energy minimum.

Validation

Several tools to assess the models' quality exist. The SAVES server [5], provided by UCLA-DOE Institute, offers some:

- ERRAT [22] differentiates between correctly and incorrectly determined regions of protein structures. It uses statistical

methods to identify model errors, which lead to more randomised distributions of the different atom types with respect to actual structures.

- Verify3D assesses the model quality by comparing it to its amino-acid sequence using a 3D profile. The 3D profile of a structure is a position-dependent scoring matrix computed from the atomic coordinates of the structure. Each residue position in the model is characterised by its environment and is represented by a row of twenty numbers, or *3D-1D scores*, indicating the statistical preferences for the environment of each of the 20 amino acids. The environment of a residue is defined by its buried area, the fraction of sidechain area covered by polar atoms, and the local secondary structure [28]. The software computes 3D-1D scores for the model and compares it to good quality structures. The Verify3D test is passed if at least 80% of the amino acids score >= 0.2 in the 3D-1D profile.

- PROCHECK [41] provides a detailed check on the stereochemistry of a protein structure by analysing residue-by-residue and overall structure geometry. One of the most informative and used outputs is the Ramachandran plot: a good quality model is expected to have over 90% of the amino acids in the most favoured regions.

- WHATCHECK [34] is derived from some of the protein verification tools of the WHAT IF software. It extensively checks many stereochemical parameters of the residues in the model.

Another useful metrics is the QMEAN score, available on the Swiss-Model website [6] by Biozentrum at the University of Basel. QMEAN stands for *Qualitative Model Energy ANalysis* and is a combined scoring function. The QMEAN4 score is a linear combination of four statistical potential terms: two distance-dependent interaction potentials of mean force, one at residue-level and one on all-atom types, are used to assess long-range interactions; the other two terms are a solvation potential and a torsion angle potential.

QMEAN6 score features two additional terms describing the agreement of the predicted and observed secondary structure and solvent accessibility, respectively [18, 19]. Both global scores are initially in a range $[0, 1]$, with one being good. They are then transformed into Z-scores to compare them with high-resolution X-ray structures.

3.3 MOLECULAR DOCKING

Docking tries to find the energetically most feasible three-dimensional arrangement of two molecules in close contact with each other and predict their binding energy. As previously stated, when investigating potential drugs' interactions with targets, the two molecules will be a *ligand* and a *receptor* (i.e. the protein).
So, molecular docking has two main goals:

- Predicting the geometry of the interaction using a *search algorithm*;
- Estimating the binding energy.

Inputs for a molecular docking simulation are:

- Three-dimensional protein structure model. One can consider more than one structure to improve predictions.
- The ligand structure: it can be 2D or 3D independently because, unlike the protein, most of the time, the ligand is assumed to be flexible in docking, so all possible conformations of the ligand will be explored *a priori* by the docking program.
- Some information about the location and the size of the binding site. This is not a fundamental requirement because blind docking approaches exist. However, it is usually better to already have some information about the binding site in order to focus on the relevant region of the interaction space.

After running a docking calculation, the output is a set of the most probable 3D structures for the complex. These are usually called ligand *poses* because the receptor is generally fixed. Each pose is associated with a *binding score* which can then be used to rank and prioritise the poses.

The score is given as *Gibbs Free Energy* ΔG. It can be related to the experimentally derived dissociation constant K_d with the following equation:

$$\Delta G = RT \ln(K_d)$$

Where R is the universal gas constant and T is the temperature in Kelvin degrees. The law of mass action can be expressed as:

$R = 8.314 \frac{J}{mol \cdot K}$

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

So the smaller the value of K_d, the stronger the ligand-protein interaction will be.

Moreover, ΔG is the difference between enthalpic and entropic contributions:

$$\Delta G = \Delta H - T\Delta S$$

Entropic contributions have to do with the degrees of freedom; it is best for the equilibrium to have high entropy, so a more negative value of ΔG indicates a higher binding affinity.

However, it should be noted that the ΔG reported by a docking software is no more than a scoring function and can just give an idea of how preferable a pose is. Docking can provide some hypotheses that should be assessed with a thorough *in silico* analysis with other tools and experimentally.

3.3.1   *Docking programs*

There are several docking programs commercially or freely available with different characteristics.

These programs differ in many ways:

- Protein representation
- Flexibility
- Search method
- Scoring function

Protein representation

The first thing one has to think of is how to represent the system: if the aim is to predict where a ligand can fit on the protein, then maybe just the surface of the target is relevant, and the internal residues can be neglected.

The first docking software, named DOCK, developed by J. Janin and S. Wodak in the early 80s and designed for protein-protein docking, used an *explicit representation* of the protein. This means it described each atom by its three-dimensional coordinates; this method has now fallen into disuse due to its computational and time requests.

A more efficient representation method is the *grid-based* one: it consists of the discretisation of the protein onto a grid whose resolution is decided directly by the grid spacing. It allows a more efficient sampling and scoring of the ligand poses. The sampling can be performed with an equally spaced grid by applying Fast Fourier Transform (FFT). AutoDock and AutoDock Vina use this representation method.

Another quite commonly used approach is the representation by constructing the *molecular surface* of the protein. The molecular surface (MS) is also well-used to visualise the protein and have an idea of the cavities inside the structures since binding sites are more likely to be located inside hydrophobic cavities. The MS is typically obtained by considering all atoms' van der Waals radii and by rolling a solvent probe sphere on this van der Waals surface. The MS will be the locus of the points where the sphere gets in contact with the surface. The molecular surface method is exploited in docking to locate cavities by filling them with the so-called gap spheres. These spheres are then used to place the ligand atoms during the sampling. The idea is to place ligand atoms one by one in the centre of the spheres while keeping interatomic distance reasonable. This way, the exploration of the interaction space is limited and makes for more efficient sampling. MOE (Molecular Operating Environment) uses this approach.

Flexibility

Most docking programs implement almost full flexibility for the ligand while keeping the target completely rigid except for thermal fluctuation. As previously stated, all possible ligand conformations are explored by the software prior to proceeding with the docking simulation. This exploration is computationally feasible thanks to the small size of the ligand; protein flexibility requires different approaches that include *soft docking* and *induced-fit docking*.

Soft docking changes the coefficient of the repulsive part of the van der Waals interaction by attenuating the Lennard-Jones repulsion parameters to allow the ligand to penetrate the protein surface slightly, thus permitting steric clashes. After the docking, the structure can be equilibrated by relaxing it through minimisation, allowing the protein to adapt to the binding.

Induced fit docking first performs a rigid docking step, then optimises the protein side chains and goes ahead with a second docking step.

AutoDock Vina and AutoDock4 include a certain amount of flexibility in the target. AutoDockFR (AutoDock for Flexible Receptors, ADFR) is a docking engine based on the AutoDock4 scoring function; it is specifically designed to handle receptor flexibility efficiently. It encodes molecular flexibility through a tree-like data structure called the Flexibility Tree. Each node of

the tree represents a set of atoms and can be assigned its motion information. The hierarchical structure of the tree allows for multi-resolution handling of flexibility [62].

Search algorithms

The goal of docking simulations is to identify the most favourable pose of a ligand towards a target. This is obtained by exploring many possible conformations of the ligand; these conformations can be automatically generated by *placement*, or *search algorithms/methods*.
Some placement methods are:

- *Matching algorithms* use probable binding sites on the protein surface to place ligand atoms until a complementarity of shape and electrostatic properties is found. Ligand atoms must be placed accordingly to the interatomic distance in the original structure of the ligand. Pharmacophore approaches can also be used to place the ligand.
- *Dynamics simulation-based methods* are better suited for post-docking minimisation and rescoring purposes.
- *Systematic search algorithms* divide into:
  - *Exhaustive search* of the conformational space of the ligand at regular intervals;
  - *Fragment-based search* splits the ligand into fragments and adds them incrementally by docking one at a time and linking it to the existing fragments;
  - *Ensemble-based methods* pre-generates poses, then perform rigid docking with them.
- *Stochastic algorithms* sample the conformational space of a ligand by generating random variations in the orientation of all rotatable bonds and sometimes random translations of the whole ligand within the binding site. This procedure can be applied to a single ligand or a population of conformations derived from the same molecular structure of the ligand. Each resultant confirmation is then evaluated according to a probability distribution or by estimating its binding affinity for the target. These methods usually implement a two-step iterative process: a random global change in the ligand coordinates followed by local optimisation. Among the search methods offered by AutoDock, the Lamarckian Genetic Algorithm follows this approach.

Scoring functions

Scoring functions quantify the energy of protein-ligand interactions such as hydrogen bonds, electrostatics, van der Waals, hydrophobic, $\pi - \pi$ *et cetera*.
The main types of scoring functions are:

- *Forcefield-based* functions are directly based on standard molecular dynamics forcefields. For non-covalent docking, only the non-bonded contribution to the potential is used. A drawback of this method is that the coefficients of the forcefield are optimised based on protein structures alone and not on protein-ligand complexes.
- *Empirical* scoring functions count the number of interactions and assing a score based on the number of occurrences.
- *Knowledge-based* methods consider a large dataset of known protein-ligand complex structures and compute the probability of the interaction between their atoms at a given distance, based on the observations in the dataset.
- *Machine learning* approaches exploit non-analytical functions obtained from regression models aimed to reproduce experimental binding energies.
- *Consensus docking* merges different approaches/scoring functions and assesses whether there is an agreement between them.

Consensus docking was introduced to overcome the limitations of docking in generating binding modes similar to the native ones and correctly ranking compounds. It consists in comparing the top-scoring poses generated by two or more docking programs. In cases where the poses are similar, usually within a given *Root-Mean Square Deviation* (RMSD) threshold, they are kept; otherwise, they are discarded. It was shown that poses passing some kind of *consensus filter* were more likely to correspond to the native binding mode of the ligand [35].
RMSD is calculated as the square root of the squared sum of pairwise differences between the positions of the particles belonging to the two structures *a* and *b*, divided by the number of particles.

$$RMSD = \sqrt{\frac{1}{N_i} \sum_i (r_{ai} - r_{bi})^2}$$

3.4 ADMET PROPERTIES

3.4.1 *Absorption*

Among the most common routes of administration of a drug, such as intravenous (IV) or intramuscular injection, and topical application, oral administration is the most used. When a drug is taken orally, it must cross the membranes of many cells to reach its target.

The *plasma membrane* of a cell is a lipid bilayer where phospholipids self-assembly with the hydrophobic tails inside and the polar heads towards the outside, in contact with either the extra- or intracellular aqueous solution. The drug molecule must be hydrophobic and rather small to diffuse across the membrane. Uncharged and non-polar drugs cross the plasma membrane with a mechanism called *passive transport*, i.e. diffusion along the concentration gradient: they move from a high concentration compartment to a low concentration one, thanks to their solubility in the lipid bilayer.

Lipophilic *and* hydrophobic *are not actually synonyms as some substances, namely silicones, are hydrophobic but not lipophilic. These substances, however, are rare and here, the two terms can be treated as synonyms.*

A means to calculate the lipophilicity/hydrophobicity is the *partition coefficient*. It is the ratio of concentrations of a compound in a mixture of two immiscible phases at equilibrium. Therefore, this ratio is a measure of the difference in solubility of the compound in these two phases. The following equation calculates it:

$$P = \frac{\left[drug\right]_{octanol}}{\left[drug\right]_{water}}$$

Octanol is a suitable solvent for the treatment of lipophilicity. ADMET predictors usually calculate the logarithm of P:

$$logP = log\left[drug\right]_{octanol} - log\left[drug\right]_{water}$$

The more positive the value of $logP$, the more hydrophobic the compound, and vice versa. Most drugs have a $logP$ value between 1 and 6. $logP$ can be predicted using computational models from a sum of fragmental, topological, knowledge- or atom-based terms plus various corrective terms.

A measure of the lipophilicity of a potential drug compound is of great importance because it correlates with many biological properties.

Another descriptor that correlates well with passive molecular transport through membranes is the molecular polar surface

area (PSA), i.e. the surface belonging to polar atoms. Calculating it with a "traditional" approach requires the generation of a 3D model and the actual calculation of the surface once polar atoms have been identified, making it rather time-consuming. Therefore a *fragment-based approach*, called TPSA (Topological Polar Surface Area), has been proposed. It is based on the summation of tabulated surface contributions of polar fragments. Its results are nearly identical to the "traditional" ones, but the computation speed is increased up to 3 orders of magnitude [29].

Charged or large polar molecules do not diffuse across the plasma membrane and are usually carried by specific proteins, which mediate another important transport mechanism called *active transport*. Unlike passive diffusion, this type of transport requires energy because the movement is opposed to a gradient. Active transport can also lead to unwanted effects. The *permeability glycoprotein* (P-glycoprotein or P-gp), for example, limits the oral absorption of some drugs. It is a membrane protein with a pump function that effluxes xenobiotics out of cells, thus reducing the effective concentration of the drug. It is extensively distributed in most tissues, with most abundance in the adrenal gland, and there is a distinct expression in endothelial cells of the central nervous system, contributing to the blood-brain barrier [7]. Suitable ADMET predictions indicate if a given compound is likely to be a substrate or an inhibitor of P-gp.

Other important parameters for drug absorption are water solubility, gastrointestinal absorption, and skin permeability:

- Water solubility is significant for orally administered drugs, especially enteral ones, as water-soluble compounds are better absorbed. It can be predicted with topological or fragment-based approaches, as well as with machine learning techniques.
  A common solubility unit is log S, corresponding to the 10-based logarithm of the solubility of a molecule measured in mol/L.
- Gastrointestinal (GI) absorption is important because the intestine is generally the primary absorption site for an orally administered drug.
- Skin permeability is of interest for the development of transdermal drug delivery.

Bioavailability

Bioavailability is the fraction of the administered drug dose that effectively reaches the systemic circulation upon any administration route. When a drug is administered via IV injection, the entire dose enters directly into the bloodstream, but for an oral drug, it can be reduced by metabolism when it passes through the liver; this is called the first-pass effect.

Many guidelines have been proposed to help develop orally bioavailable drugs. Among these, the most widely known and used is *Lipinski's Rule of Five*, a set of empirical rules proposed in the late 1990s by Chris Lipinski, a medicinal chemist working for Pfizer, and his research group. These rules are so-called because they are five and the number 5 recurs in each:

- Molecular Weight (MW) ≤ 500 Da
- Number of rotatable bonds ≤ 5: the molecule must be flexible, but not too much. This rule is about specificity (the probability of binding to the target) and selectivity (the probability of binding to the target and not to other targets).
- Number of hydrogen bond acceptors ≤ 10
- Number of hydrogen bond donors ≤ 5
- -2 < log P < 5: if the compound is too hydrophilic, it will bind to the surface of many proteins; if it is too hydrophobic, it can get stuck in the inner layer of membranes. In both cases, the drug would not reach its target.

A set of less stringent rules was later proposed, in which the accepted MW is 550 Da, and the number of rotatable bonds rises to 7. These rules are generally used in oncology.

Other rules are the *Ghose filter*, which takes into account the molar refractivity and the number of atoms of the compound; the *Veber filter*, developed by a GlaxoSmithKline group, that focuses on rotatable bonds and topological polar surface area; the *Egan filter*, and the *Muegge filter*.

3.4.2 *Distribution*

The circulatory system distributes the compound to its site of action after absorption (for an oral drug) or systemic administration (for IV drugs). A compound will likely exist in plasma at an equilibrium between an unbound state and bound to serum proteins. The unbound drug exerts the therapeutic effect, so

knowing which fraction of the dose is free is essential. The *Fraction Unbound* predicts the amount of a compound that would be unbound in plasma.

Many drugs accumulate in tissues, too. The *Volume of Distribution at steady state* (VDss) is the volume that the dose of the drug should have to be uniformly distributed to give the same concentration as in blood plasma. The higher this parameter is, the more drug is distributed in tissues rather than in plasma. This accumulation could lead to local toxicity phenomena.

ADMET distribution parameters also indicate the ability of the compound to reach some body districts. The brain is protected from exogenous compounds by the blood-brain barrier (BBB); the ability of a drug to cross it, called *BBB permeability*, is a vital parameter to help predict potential side effects or to improve the efficacy of a drug whose target is within the brain.

3.4.3 *Metabolism*

Since drugs are foreign substances to the body, they are metabolised by a large number of enzymes. These enzymes catalyse the chemical transformation of the compounds into more polar and water-soluble ones, called metabolites. This transformative process is called metabolism or biotransformation, and it can occur before or after the drug has reached its site of action; for example, an inactive prodrug can be converted into an active drug through a metabolic reaction.

Metabolic reactions are generally divided into two phases:

1. Phase I enzymes modify the functional groups of the compound through oxidation, reduction, hydrolysis, and carboxylation. The oxidation reactions are carried out by cytochrome P450 (CYP). CYP are a superfamily of enzymes, including approximately a hundred isoforms, involved in the oxidative reactions of many compounds. Twelve human CYPs are known to be important in the metabolism of xenobiotics, among which the most active are CYP2C, CYP2D, and CYP3A families. ADMET tools can predict whether a given compound is likely to be a substrate of CYPs. Higher lipophilicity increases the likeness of being metabolised by CYPs.

   It is also important to predict whether a drug candidate could be a CYPs inhibitor because, in the case of coadministration with other drugs, inhibiting a metabolising enzyme

causes another molecule that is usually metabolised by that enzyme to accumulate, potentially leading to adverse effects [63].

2. Phase II reactions are generally conjugation reactions producing a very water-soluble metabolite.

3.4.4 *Excretion*

Excretion, or elimination, indicates the irreversible removal of a drug from the body, which of course, reduces its concentration at the site of action. Total clearance mainly occurs as a combination of hepatic (i.e. metabolism in the liver and biliary clearance) and renal clearance, i.e. urinal excretion, and it is related to bioavailability. Having a parameter that predicts the *Total Clearance* is important for determining drug dosage.
Renal clearance is an important means of excretion. The Organic Cation Transporter 2 (OCT2) is a renal uptake transporter playing an essential role in the clearance of both endogenous and exogenous compounds. It is of interest to predict whether a potential drug is likely to be a *Renal OCT2 substrate* because it could lead to adverse effects if OCT2 inhibitors are coadministered.

3.4.5 *Toxicity*

A drug molecule is expected to be efficacious and to have a good toxicity profile as well. Several ADMET indicators have been developed to describe the toxicity of a drug. Some are derived from animal-based assays:

- *AMES toxicity* predicts whether a given compound is likely to be mutagenic. This prediction is based on the Ames test, which assesses the mutagenic potential using bacteria.
- *Minnow toxicity* indicator predicts the molecule concentration necessary to cause the death of 50% of the *Flathead Minnows* (LC50, *Lethal Concentration 50*). A low value of this concentration indicates high acute toxicity.
- *Rat acute toxicity* predicts the LD50 for rats, i.e. the amount of a compound given at once that causes the death of 50% of a group of test animals.
- *T. Pyriformis toxicity* predicts the negative logarithm of the concentration required to inhibit the 50% of the growth of *Tetrahymena Pyriformis* protozoa bacteria.

More important, though more difficult to predict, are toxicity predictors in humans:

- *Maximum Tolerated Dose*: the Maximum Recommended Tolerated Dose (MRTD) is an estimate of the toxic dose in humans; it is the highest dose of a drug or treatment that does not cause unacceptable side effects. The computational models are based on data from human clinical trials.

- *Maximum Therapeutic Dose*: estimates the upper limit beyond which a drug's efficacy is not increased and side effects begin to outweigh beneficial effects.

- *Hepatotoxicity* predicts if a given compound's effect could disrupt the normal function of the liver.

- *Skin Sensitisation* is a potential undesired effect of topically administered drugs.

- *Cardiotoxicity* is a major safety concern. This indicator predicts whether a given compound is likely to inhibit the potassium ion channels in cardiomyocytes. This channel is encoded by the human Ether-à-go-go-Related Gene or hERG, and thus commonly called hERG itself. Inhibition of hERG channels is the leading cause of developing a potentially fatal disorder called long QT syndrome: since the channel mediates repolarisation of the heart muscle, which helps coordinate the heart's beating, inhibiting its conductive ability leads to ventricular arrhythmia.

MODELLING OF HUMAN TUBULIN ISOTYPES

This section focuses on the methods used to obtain and prepare models of human tubulin structures.

4.1 ● γ tubulin

4.2 ● α-β tubulin dimer
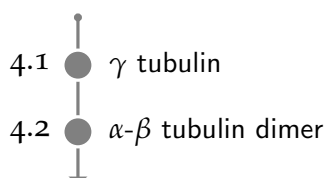
Despite bovine and porcine tubulin being the cheapest and most common choices when it comes to lab experiments, it is essential to investigate human tubulin if the focus is on clinical applications. While some experimentally resolved structures of human γ tubulin exist, human α-β dimers were obtained by comparative modelling.

4.1 γ TUBULIN

4.1.1 *Materials and Methods*

Template selection

Four PDB entries for human tubulin γ chain were taken into consideration:

- 1Z5V: Crystal structure of γ-tubulin bound to GTPγS
- 1Z5W: Crystal structure of γ-tubulin bound to GTP
- 3CB2: Crystal structure of human γ-tubulin bound to GDP
- 6V5V: Structure of γ-tubulin in the native human γ-tubulin ring complex

They all refer to human tubulin γI chain, encoded by the gene TUBG1. The structures were compared according to these three criteria, listed in order of importance:

1. Resolution: any resolution between 1.5 and 2.5 Å is considered to be excellent.
2. Number of missing residues (MRES): a structure with fewer missing residues is preferred. Moreover, it is best if these missing residues are not in or near known or probable active sites.

3. Publication date: compatibly with the other criteria, a more recent structure is preferable.

Table 4 shows these features for the four PDB structures. The entries are ordered from the least to the most recent.

Table 4: PDB entries comparison for γ tubulin.

PDB entry	Resolution (Å)	Publication date	Number of MRES
1Z5V	2.71	2005	39
1Z5W	3	2005	42
3CB2	2.3	2008	19 (chain A)
6V5V	3.8	2020	82

Supplementary figure A shows the missing residues locations in these proteins relative to the position of the residues forming the GDP binding site.

PDB entry 3CB2 was selected because it has the best characteristics in terms of resolution and number of missing residues, and it is fairly recent as well. The original PDB entry contains two γ chains; the A chain was selected to undergo the following steps because it had fewer missing residues than the B chain. The structure 6V5V was also taken into consideration as a means to compare the folding of the selected template since it is the structure of the protein in the native ring complex. Unlike the other three models, this structure was obtained by electron microscopy.

Structure inspection and optimization

The MOE Align/Superpose option in the SEQ panel was used to compare the high resolution γ tubulin structure 3CB2 with the native one in 6V5V. Since they both refer to the same protein, Sequence and Structural alignment was selected to proceed with the superposition. The RMSD (*Root Mean Squared Deviation*) value was used as a metric for structure similarity.

The 3CB2 structure was then prepared using the Structure preparation panel in MOE. The major fixes made during this process are listed below.

1. Alternates: residues with alternate locations were corrected by using the highest occupancy one.
2. Termini: missing backbone atoms in the protein chain C- or N- termini were deleted, and the terminus was capped.

3. Breaks: missing residues inside the chain were corrected by building a loop.
4. Library: inconsistencies between the residue name and its structure or missing atoms were corrected by using the sequence information over the structure.

The Protonate 3D application was then used to assign ionization states and position-optimized hydrogens. An energy minimization step was performed to relieve strains in the structure.

4.1.2 *Results*

The RMSD between 3CB2 and 6V5V γ tubulin equals 1.185 Å. This result indicates a good grade of similarity between the structures. As shown in figure 8, some of the per-residue RMSD values are found near gaps due to missing residues. Figure 9 shows these results graphically: by selecting Ribbon → RMSD in the main MOE window upon superposition, the structures are represented according to a colour map in which a deeper shade of green indicates a lower RMSD value, while yellow, orange and red tones mean progressively higher values. Since 6V5V has more missing residues, some portions of 3CB2 that do not match any 3D structure in 6V5V are depicted in white.
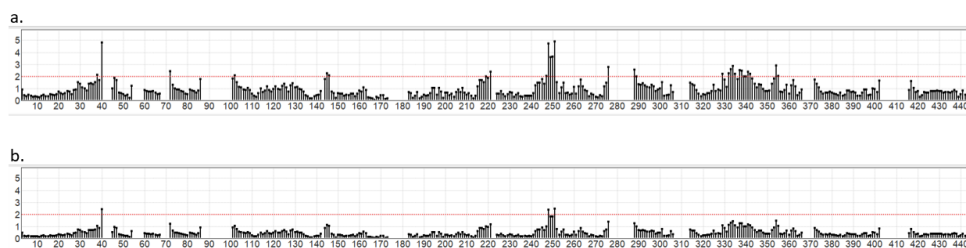
Figure 8: RMSD by residue. a. shows the RMSD real value; b. reports the averaged RMSD.

4.2 α-β TUBULIN DIMER

4.2.1 *Materials and Methods*

Template selection

Since gatastatin is a derivative of a colchicine site-binding compound, the search for templates of α-β tubulin for consensus
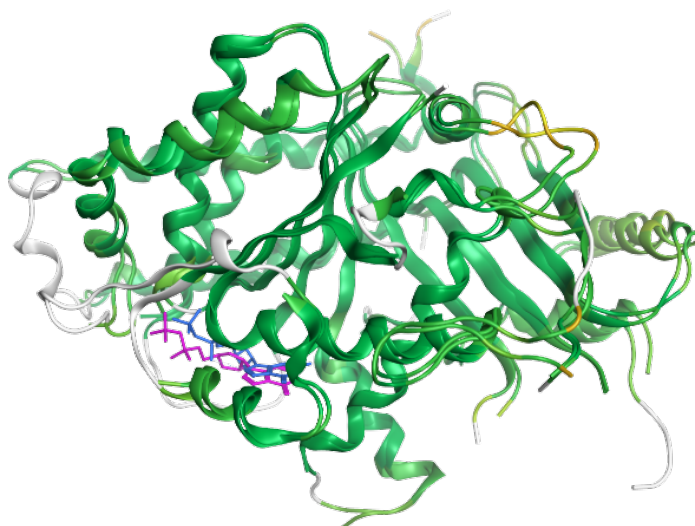
Figure 9: Structure superposition of 3CB2 and 6V5V. The complexed ligand GDP is shown in blue for 3CB2 and magenta for 6V5V.

docking simulations was oriented towards colchicine bound protein complexes. Many such structures have been deposited in the Protein Data Bank over the last two decades. As for γ tubulin, the PDB entries have been inspected in terms of resolution, number of missing residues, and publication date:

Table 5: PDB entries comparison for α-β tubulin.

PDB entry	Resolution (Å)	Publication date	Number of MRES
1SA0	3.58	2004	24/29 (chains A/C) 26 (chains B/D)
4O2B	2.30	2014	12 (chains A/C) 24 (chains B/D)
5EYP	1.90	2016	23 (chain A) 20 (chain B)
5NM5	2.05	2017	29 (chain A) 28 (chain B)

PDB entry 5EYP was selected, mainly for its best resolution. This structure is representative of free tubulin; 3J6F was selected as a second template to have a means of comparison with the tubulin dimer as it is inside microtubules. PDB entry 3J6F is an average structure of dynamic pig microtubules obtained by electron microscopy, with a resolution of 4.90 Å. The low

number of unmodeled residues (11 in α and just one in the β subunit) makes up for the sub-optimal resolution. Fewer missing residues mean that less computational modelling has to be performed on the initial template, leading to a lower probability of introducing errors in the structure.

Being a microtubule structure, 9 α-β heterodimers are present in it; chains A and B were used as the template structure.

Homology modelling

The selected templates were prepared following the procedure previously described in section 4.1.1.

The target sequences of human tubulin were downloaded as FASTA files from the UniProt KnowledgeBased (UniProtKB [8]) database. A homology model for each β tubulin isotype was built for both templates. Tubulin α1A chain was used for the α subunit in all the models. A library of human α-β tubulin structures was obtained; the models feature tubulin α-1A in complex with:

The FASTA file format is a nearly universal standard in the field of bioinformatics. It represents nucleotide or protein sequences in a text-based format in which nucleotides or amino acids are represented using single-letter codes.

- tubulin βI
- tubulin βIIa
- tubulin βIIb
- tubulin βIII
- tubulin βIVa
- tubulin βIVb
- tubulin βV
- tubulin βVI
- tubulin βVIII

for a total of nine structures for each of the selected templates.

Below is an outline of the steps required to obtain a homology model in MOE starting from the prepared template and the target sequences.

1. Open the template structure and the two target sequences - that is, tubulin α-1A and one of the β isotypes - in MOE.

2. Open the Sequence Editor panel using the SEQ button in MOE main window. This panel shows the sequences of all the items in MOE System.

3. Using the Set Up Chains button that comes with the Align/Superpose tool, it is possible to split the window

into two sub-figures, each dedicated to a protein subunit. Chains can be assigned to a subunit by either selecting the By Tag button that splits the sequences based on their letter tag (A and B, for example), or by selecting them and changing their destination with the Subunits buttons in the bottom-left corner of the window. These tools were used to assign the α subunit and the αIA sequence to one sub-window and the β subunit and FASTA sequence to the other.

Ligands were set to be ignored by the alignment tool using the dedicated *i* red button.

4. At this point, the sequences can be aligned by clicking on Align. It is possible to observe and adjust the sequence alignments, and Alignment → Similarity... can be selected to get a report of sequence identity and similarity.

Other than with the Align/Superpose application in MOE, the human isotypes sequences were compared with animal ones in order to inspect their differences in the colchicine binding site. Pig and ovine β tubulin, from 3J6F and 5EYP respectively, were used as means of comparison in this analysis. Just the β subunit was considered because α tubulin makes little contact with colchicine with respect to β.

The Clustal Omega application for multiple alignment was used to do so. It is provided by the *European Bioinformatics Institution* [45] and is freely available online [9].

5. By clicking Protein → Homology Model..., the panel dedicated to comparative models will show. Here, the sequences and corresponding template chains were set.

The ligands and the magnesium ions were set to be Used as Environment. This option includes the selected atoms in the energy tests and minimization stages of the process so that clashes and superimpositions with the co-complexed atoms are prevented.

Both experimentally derived structures had several (10 to around 20 depending on the isotype) missing residues at the carboxy terminus of the sequence. This outgap was not modelled as these regions are not involved in colchicine binding. Moreover, these regions are likely to be very mobile, which is probably why the crystallography could not resolve them in the first place, so trying to model them could result in errors that worsen the overall quality of the

model without adding useful information.

In Settings, the number of sidechain samples was increased to 5. Since human and animal tubulins share above 80% similarity in all cases, they do not differ in the main chain, so just one model would have been created if the sidechain sample number option was left to the default value of 1. This way, five models were created and ranked for each template-target combination.

Optimization of ionization states and proton placement with the Protonate 3D application and the subsequent refinement step were applied to all intermediates.

6. After settings and options in the Homology Model panel were adjusted, calculations were started by clicking OK.

A database containing five models resulted from these steps. No changes were made to the default Model Scoring option, so homology models were ranked according to the GB/VI score, which is a generalized Born model for estimating the free energy of hydration developed in 2008 by Labute *et al.* [40].

GB/VI stands for Generalized Born/Volume Integral; it estimates the free energy of hydration as a classical electrostatic energy contribution plus a cavitation energy contribution based upon a VI London dispersion energy instead of the atomic surface area (SA) used in GB/SA hydration models.

The best model was selected based on the GB/VI score and geometric parameters such as the number of atom clashes, rotamer, and bond length outliers. It was then imported into the MOE main window. An energy minimization step was performed to relieve strains and atom clashes in the structure while tethering the backbone heavy atoms to avoid losing the complexed protein folding.

Validation

Each homology model was carefully inspected using the tools described in section 3.2. The structures were saved as .pdb files and uploaded on the SAVES server [5] to obtain ERRAT, Verify3D, WHATCHECK and PROCHECK insights, and on the QMEAN website [6].

4.2.2 Results

Differences between human and animal β tubulin in the colchicine binding site

Figure 10 shows the results of the sequence analysis made with Clustal Omega. The residues making up the colchicine binding site on β tubulin are reported. Dots indicate perfect identity, while mismatches are colour-coded as follows:

- Green highlighted residues indicate conservative replacements that scored above 0.5 in the PAM250 matrix, indicating strong similarity.
- Orange means weak similarity: these replacements have a PAM250 score of 0.5 or less and are called semi-conservative.
- Red residues are not conserved.

		·237			·240			·246			·249			·252	
TBB-PIG	G	V	T	T	C	L	L	N	A	D	L	R	K	L	A
TBB-SHEEP	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.
Human βI	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.
Human βIIa	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.
Human βIIb	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.
Human βIII	.	.	.	.	s	.	.	.	.	.	.	.	.	.	.
Human βIVa	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.
Human βIVb	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.
Human βV	.	.	.	.	s	.	.	.	.	.	.	.	.	.	.
Human βVI	.	i	.	.	s	.	.	.	.	.	.	.	.	.	.
Human βVIII	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.

	·255			·312			·315			·348			·351	
TBB-PIG	V	N	M	T	V	A	A	V	N	N	V	K	T	A
TBB-SHEEP	.	.	.	.	.	.	.	.	.	.	.	.	.	.
Human βI	.	.	.	.	.	.	.	.	.	.	.	.	.	.
Human βIIa	.	.	.	.	.	.	.	i	.	.	.	.	.	.
Human βIIb	.	.	.	.	.	.	.	i	.	.	.	.	.	.
Human βIII	.	.	.	.	.	.	t	.	.	.	.	.	v	.
Human βIVa	.	.	.	.	.	.	.	.	.	.	.	.	.	.
Human βIVb	.	.	.	.	.	.	.	.	.	.	.	.	.	.
Human βV	.	.	.	.	.	.	t	.	.	.	.	.	v	.
Human βVI	.	.	.	.	.	.	c	i	.	.	.	.	v	.
Human βVIII	.	.	.	.	a	.	.	i	.	.	.	.	.	.

Figure 10: Comparison between animal β tubulins and the human isotypes over the residues forming the colchicine binding site.

For the most part, mismatches are conservative or semi-conservative, except for the one in position 315. Non-conservative replacements could cause a change in the cleft's charge or polarity that, in turn, would affect the ligands' modes of binding and binding affinities.

Validation

For the sake of brevity, in this section, homology models will be indicated just with the β tubulin isotype they feature, so for example, the model of the heterodimer α-1A and β-1 will be called β-1.

All models obtained from 5EYP got an ERRAT score of around 94%, and those obtained from 3J6F scored above 96%, indicating that the percentage of errors in the computational models is reasonably low in both cases.

All models passed the Verify3D test, i.e. at least 80% of each structure's amino acids scored 0.2 or above in the 3D-1D profile. All models from 3J6F scored above 97% and the ones from 5EYP scored around 99%.

All homology models had around 91% of their amino acids in the most favoured (core) regions according to the analysis with PROCHECK, and none showed residues in unfavourable regions.

These scores, along with QMEAN4 and QMEAN6 normalized scores, are reported in more detail in tables 6 and 7.

Table 6: Scores of the homology models obtained from 5EYP.

	βI	βIIa	βIIb	βIII	βIVa	βIVb	βV	βVI	βVIII
ERRAT	94.38	94.01	93.65	94.26	94.25	94.26	94.37	94.21	94.26
Verify3D	99.42%	99.31%	99.42%	99.19%	99.42%	99.54%	98.27%	99.65%	99.54%
Ramachandran	92.2%	91.1%	91.1%	91.4%	91.8%	91.7%	91.0%	91.1%	91.5%
QMEAN4	0.73	0.73	0.73	0.73	0.73	0.73	0.73	0.72	0.73
QMEAN6	0.71	0.71	0.71	0.70	0.70	0.71	0.71	0.71	0.71

Table 7: Scores of the homology models obtained from 3J6F.

	βI	βIIa	βIIb	βIII	βIVa	βIVb	βV	βVI	βVIII
ERRAT	97.40	96.69	96.81	96.81	96.57	97.16	96.81	96.46	96.69
Verify3D	97.69%	98.04%	98.50%	97.81%	97.92%	98.15%	98.04%	98.61%	98.96%
Ramachandran	91.1%	91.1%	91.1%	90.9%	91.1%	91.2%	91.0%	91.1%	91.1%
QMEAN4	0.74	0.74	0.74	0.74	0.74	0.74	0.73	0.73	0.74
QMEAN6	0.71	0.70	0.71	0.71	0.71	0.71	0.70	0.71	0.70

MOLECULAR DOCKING

This section describes the molecular docking procedure, with particular attention to the adopted consensus docking protocol.

5.1 ● Input files preparation

5.2 ● Methods

5.3 ● Results
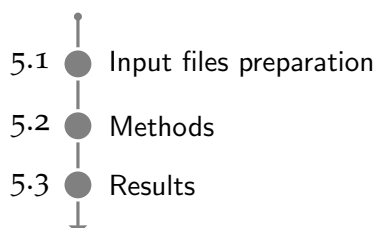
5.1 INPUT FILES PREPARATION

5.1.1 *Receptors*

The prepared 3CB2 structure of γ tubulin was used for both consensus docking and blind docking simulations.
As gatastatin is derived from a colchicine site-binding compound, consensus docking on the α-β dimer focused on the colchicine binding site. Homology models obtained from 5EYP, representing unpolymerised tubulin heterodimers, were used for this purpose. Blind docking was performed on homology models from 3J6F to investigate the compounds' binding to microtubules.

Regarding MOE, for γ tubulin the outputs of the optimization step described in section 4.1.1 were used; input files for α-β tubulin were the energy minimized homology models. These structures were converted from the proprietary .moe format to .pdb files so that the program AutoDockTools could handle them and produce .pdbqt format input files for AutoDock and AutoDock Vina. The PDBQT file format is very similar to PDB, but it includes additional information on partial charges and AutoDock 4 atom types. The prepare_receptor4.py script, that is part of the AutoDockTools (ADT) package, was used to obtain them. For docking simulations in the colchicine binding site of α-β tubulin, options were set to maintain the native GTP ligand at the interface between the subunits.

5.1.2 *Ligands*

The co-complexed ligands - colchicine for the models obtained by 5EYP and GDP for γ tubulin, shown in figure 11 - were extracted from the original PDB files.
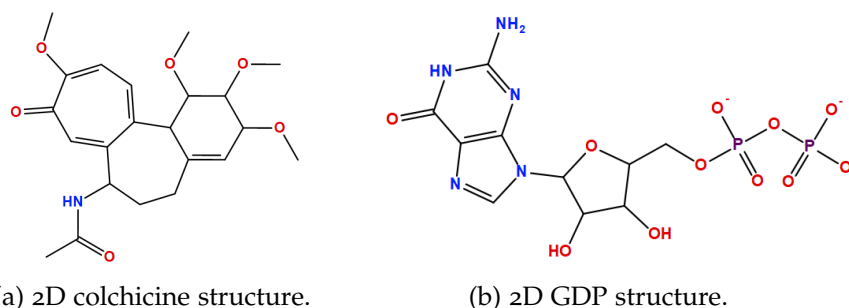

(a) 2D colchicine structure.   (b) 2D GDP structure.

Figure 11: Chemical structures of co-complexed ligands for α-β and γ tubulin.

PubChem is an open chemistry database at the National Institutes of Health (NIH), containing mostly small molecules.

The 3D structure of gatastatin, as first identified by Chinen *et al.* in 2015 [21], was downloaded from PubChem [10]. This structure was then used as the base to build all its derivatives, developed by Shintani *et al.* and described in their 2020 article [52]; the Builder application in MOE was used to obtain them. The derivatives can be divided into three main groups based on the modifications made to the gatastatin structure:

- B-ring substituted derivatives;
- O^6 modified derivatives;
- O^6 modified, 3'-4' acetonide derivatives.

Their 2D structures are shown in tables 8, 9 and 10, respectively. Figure 12 shows the substitution points: the ring, shaded in blue, and O^6, highlighted in red.

To proceed with docking simulations in MOE, once gatastatin derivatives' structures were obtained, all ligands were imported into a MOE database. Following MOE's guidelines for database preparation, three steps were applied:

1. Database *washing* to correct possible structural errors in the ligands imported from external sources;
2. *Partial Charges calculation* using AM1-BCC forcefield;
3. *Energy Minimization*.

Ligand input files for AutoDock and Vina must be in the .pdbqt format as well. The ADT prepare_ligand4.py script was used to convert PDB ligand files into PDBQT.
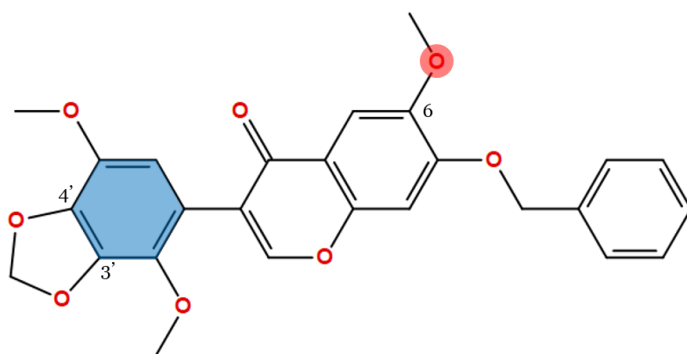
Figure 12: 2D structure of gatastatin with highlighted points of derivation.

Among the O^6 modified structures, the authors identified O^6-propargyl gatastatin, i.e. S9, as a more potent γ tubulin specific inhibitor than gatastatin; therefore, it was renamed gatastatin G2, and it will be called G2 in the following, while the original gatastatin molecule will be referred to as G1.

Table 8: 2D structures of B-ring modified gatastatin derivatives.

Derivative	Structure
S3	
S4	
S5	

Table 9: 2D structures of O^6 modified gatastatin derivatives.
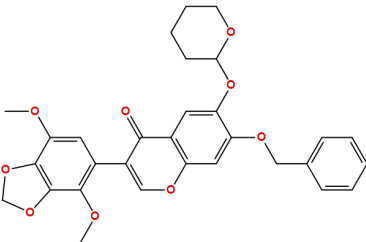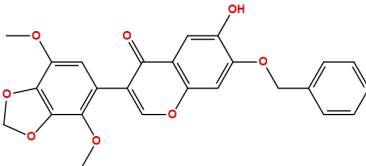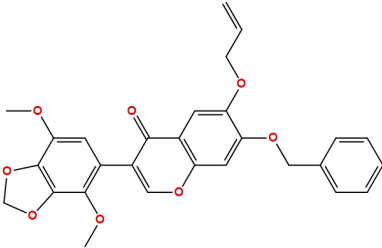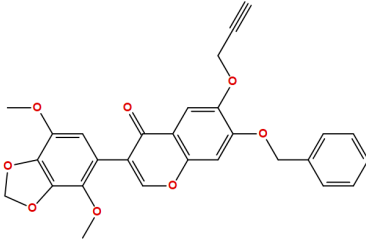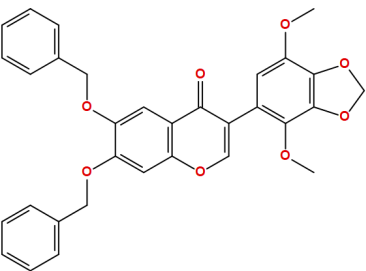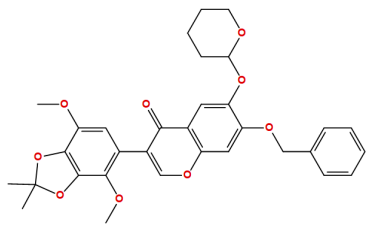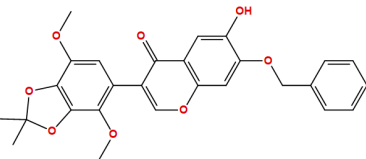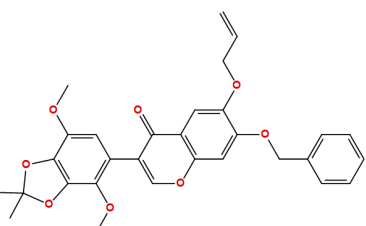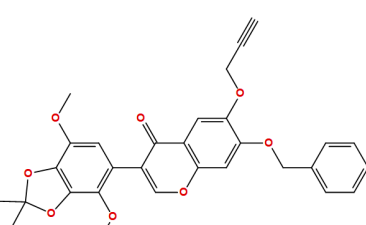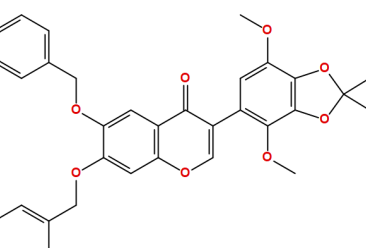
Derivative	Structure
S6	
S7	
S8	
S9	
S10	

Table 10: 2D structures of O^6 modified, 3'-4' acetonide gatastatin derivatives.

Derivative	Structure
S11	
S12	
S13	
S14	
S15	

5.2 METHODS

5.2.1 *Consensus Docking*

The outputs from the three docking engines were compared to look for a possible consensus to seek more reliable results.

Search space definition

Most of the consensus docking simulations were applied as re-docking, i.e. the compounds were docked in the cleft occupied by a co-complexed ligand in the PDB structure. In the case of γ tubulin, this ligand was the native GDP, while it was colchicine for the α-β dimer. The co-complexed ligands from PDB were also redocked, and the results were used as a benchmark to assess the docking process reliability. Using MOE, it is sufficient to select the co-complexed ligand in the Dock panel to define the binding site. AutoDock and Vina, on the other hand, require the definition of a 3D box in which to focus the search. The program AMDock, freely available online [11], was used to calculate optimal box size and placement for each receptor-ligand pair though the Center on Hetero option. The AMDock application implements a script developed by W. P. Feinstein to optimise the box size [30, 56].

Other possible binding sites than the GDP one were explored on γ tubulin: to do so, the results from the Site Finder application in MOE were used. It is a geometric method - meaning it does not use energy models - to identify possible active sites in a receptor. It is based upon α-shapes, which are piecewise linear curves families in the Euclidean plane, associated with the shape of a finite set of points [26]. The method identifies regions of tight atomic packing and proceeds to filter out unlikely sites such as protrusions, inaccessible regions or too solvent-exposed ones. The potential sites are then ranked according to their *Propensity for Ligand Binding* (PLB) score, based on their amino acid composition, that was implemented from [53].

Among the 30 Site Finder results for γ tubulin, six have a positive PLB score and of these the first three scored above 1, as summarised in table 11. Site 1 overlaps with the GTP-GDP binding site identified by the co-complexed GDP molecule, so binding of gatastatin G1 and G2 in sites 2 and 3 was analysed: for docking simulations in MOE, the option Dummies... in the Site Finder panel was used to create dummy atoms corresponding to the α spheres; these dummies were subsequently used to guide ligand placement during the docking simulation.

For AutoDock and Vina, the search space was prepared using the Center on Residue(s) option in AMDock: this application obtains the list of residues forming the cleft as input and centres the box on their geometric centre.

Table 11: Top-scoring Site Finder results for γ tubulin.

Site	PLB	Residues
1	4.56	10-17; 68-70; 98-102; 139-147; 171-185; 207-216; 221-225; 228-229; 232; 394; 397-398; 404
2	1.15	270; 298; 300; 303; 305; 307-317; 345-347; 377; 381; 383; 386
3	1.02	24; 27-28; 217-218; 227; 320-231; 234; 237-238; 273; 275-280; 283; 325; 367-371; 374

Lastly, the putative colchicine site on γ tubulin predicted by Friesen *et al.* was also investigated for gatastatin G1 and G2 binding. Colchicine was docked in this cleft, too, in order to compare the predicted binding affinity for γ and α-β tubulin.

Consensus Docking Protocol

The adopted consensus scheme is similar, at least in the early stages, to the one described in [35]: the top-scoring poses from different docking engines were compared in terms of RMSD and just the similar ones - that is, the ones that differ less than a given RMSD threshold - were kept.
Below is a more detailed description of the first steps of the process, which are also graphically summarised in figure 13.

1. The five highest-scoring poses were extracted from each program's run, converted to mol2 format and stored in a folder. The poses are named as follows: x⟨program⟩y, where x is the number of the docking run from which the pose was extracted, ⟨program⟩ is autodock, moe or vina, and y indicates the number of the pose inside the single docking simulation (1-5 for MOE and Vina, random for AutoDock since the output poses are not sorted according to their score).

mol2 format files are plain text (ASCII) tabular files storing atomic coordinates, chemical bond information, and metadata about a single or multiple chemical compounds.

2. All the stored poses were compared to one another by computing the RMSD between them in a pairwise fashion. The program DockRMSD, freely available online [12], was used to calculate RMSD values. It is particularly suitable for comparing poses generated by different programs thanks to the atomic correspondence search it implements. Moreover, it features a symmetry correction that allows fair comparison for molecules with symmetric functional groups [17].

Different poses from the same docking program were also compared to assess the consistency between intra- and inter-programs results.

3. For each comparison between two docking results, if the calculated RMSD value was below 2 Å, then the names of the two poses were stored in a row of a text file. The resulting list of pose pairs constituted the input for the subsequent steps.
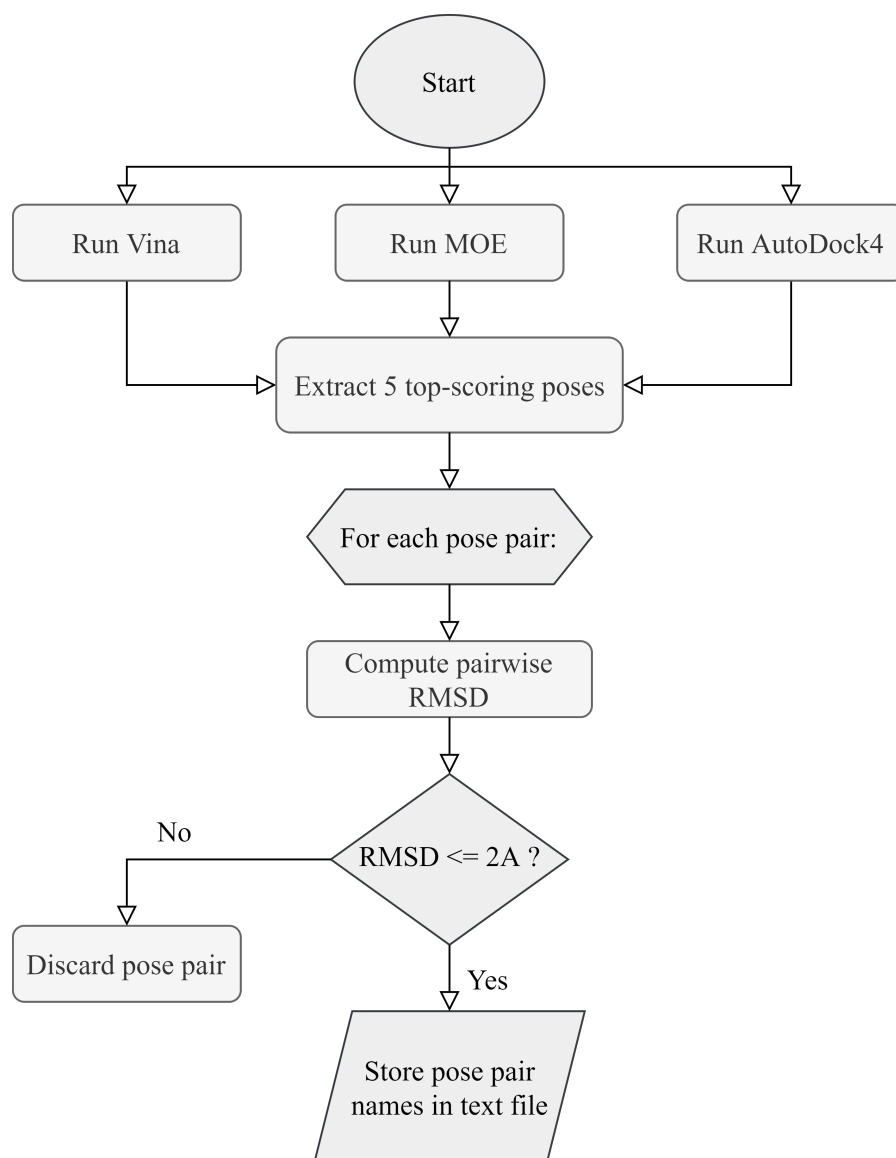
Figure 13: Preliminary steps of the adopted consensus protocol.

Since three docking programs were used, a clustering approach was adopted to find a consensus. The *clique problem* theory was exploited to do so: in computer science, the clique

problem is the computational problem of finding cliques in a graph. *Cliques*, also called *complete subgraphs*, are subsets of vertices in the graph that are all *adjacent*, i.e. connected to one another by means of so-called *edges*. Despite originating to discover groups of mutual acquaintances in social networks, the clique problem has found many applications in bioinformatics and computational chemistry.

In this particular problem, the set of poses is assimilated to a graph, and the list of pose pairs represents its edges, so the concept of adjacency is represented by similarity in terms of RMSD: two vertices (i.e. poses) that differ less than 2 Å are considered to be adjacent.

The second part of the process was performed as follows:

4. A Python script was used to identify all the cliques of size three or more. The code was adapted from an algorithm to find cliques of a given size that was published online on the opengenus website [13]. The combinations of pose names forming the cliques were stored as rows of a text file that could be quickly processed in the following filtering step.

5. Another Python script was used to filter the obtained cliques: only clusters meeting the simple criteria of including at least a pose from each docking program are retained and thus represent a consensus.

If no cliques survived the filtering process, steps 1-5 were repeated by adding the newly generated poses to the previously obtained ones. This way, the pool of possible docking results was widened with each iteration.

If after eight runs (meaning that 40 poses from each docking program were compared) no filtered clusters were found, steps 2 to 5 were repeated while relaxing the RMSD threshold from 2 to 2.5 Å. If no clusters resulted from this additional analysis, it was considered very unlikely to obtain a consensus for that particular ligand in the analysed binding site.

Results analysis

The poses forming each cluster were imported in MOE as a database in order to inspect ligand-protein interactions through *Protein-Ligand Interactions Fingerprints*, or PLIF. This tool clas-

sifies interactions such as hydrogen bonds and ionic ones and builds them into a fingerprint scheme.

Given the binding energies (ΔG) predicted by the docking engines and expressed in kcal/mol, the equivalent K_d value was calculated using the following relation:

$$\Delta G = RT \ln K_d \;\; \rightarrow \;\; K_d = \exp \frac{\Delta G}{RT}$$

with a temperature of 310 K. Since differently sized clusters were obtained, the average score values were calculated considering the best-scoring pose of each program to obtain a non-biased mean. The equivalent K_d values were calculated and averaged accordingly.

5.2.2 *Blind Docking*

Given the larger receptor size, MOE was used to perform all blind docking simulations on α-β tubulin as it provides the best trade-off between accuracy and computation time. Each run generated one hundred poses, and the top-scoring 30 poses were retained for further analysis.

Blind docking simulations on γ tubulin were also performed in AutoDock and Vina.

In order to verify the actual preference for the GDP binding site, both on γ and on β tubulin, a blind docking run was performed upon removing the co-complexed GDP molecule from the structure.

Search space definition

As the MOE guidelines for binding site identification recommend to indicate as the docking site a subset, rather than all, of the receptor, the Site Finder application was used; dummy atoms were generated in all the resulting sites to define a search space that samples the entire accessible surface of the receptor(s).

For AutoDock and Vina, a search box large enough to contain the whole receptor was created using the AutoDockTools graphical interface. To get a further mean of comparison, the Automatic mode on AMDock was also used. Its definition of the search space is similar to the one applied in MOE: the AutoLigand application is used to predict possible binding sites, and search boxes are placed at the centre of each.

5.3 RESULTS

5.3.1 γ Tubulin

Consensus docking: GDP binding site

GDP redocking produced a consensus both among the programs and with the co-complexed pose, meaning that all three docking engines could reproduce the ligand's native binding conformation. On average, the predicted binding energy is -9.61 ± 1.14 kcal/mol and the corresponding predicted K_d value is 0.49 ± 0.74 μM.

Regarding colchicine, a consensus was reached within 3 runs; the average binding energy value is -6.48 ± 0.16 kcal/mol, and the mean K_d is 27.72 ± 6.52 μM.

For gatastatin G1 and G2, 1 and 3 runs were needed to get a consensus, respectively; table 12 reports their mean scores and K_d values.

Table 12: G1 and G2 binding affinities to the GDP site on γ tubulin.

Compound	Score (kcal/mol)	Equivalent K_d (μM)
Gatastatin G1	-8.42 ± 0.57	1.64 ± 1.10
Gatastatin G2	-8.11 ± 0.44	2.43 ± 1.56

All gatastatin derivatives were docked against γ tubulin in the GDP binding site; all yielded a consensus, except for the compound S15. For each of them, table 13 lists the number of runs required to get a consensus, the average score on the cluster, and the mean predicted K_d value. Compound S11 has the highest predicted affinity, but it took four runs to obtain a consensus, possibly indicating some variability in the binding poses of this ligand. Among the compounds that yielded a consensus within the first run, S5 has the lowest K_d value. Figure 14 shows the PLIF of the MOE poses of gatastatin and its derivatives, except S15. The compound forming the highest number of interactions is S4, but as its binding affinity is among the lowest predicted ones, these interactions may not be high energy.

Consensus Docking: Site Finder *results*

Except for gatastatin G2 in site 2, which required eight runs, no clear consensus was obtained. A high degree of variability

was noticeable through visual inspection of the docked poses in MOE, suggesting that binding to these sites may not be very specific. For gatastatin G1, this hypothesis is supported by the docking scores, which were lower than the binding energy in the GDP binding site. Table 14 summarises this analysis; for the compounds that did not yield a consensus cluster, the scores and equivalent K_d values of the top-scoring MOE pose are reported. It should be noted that, since no consensus was obtained, these results are even more purely indicative.

Table 13: Gatastatin derivatives consensus results.

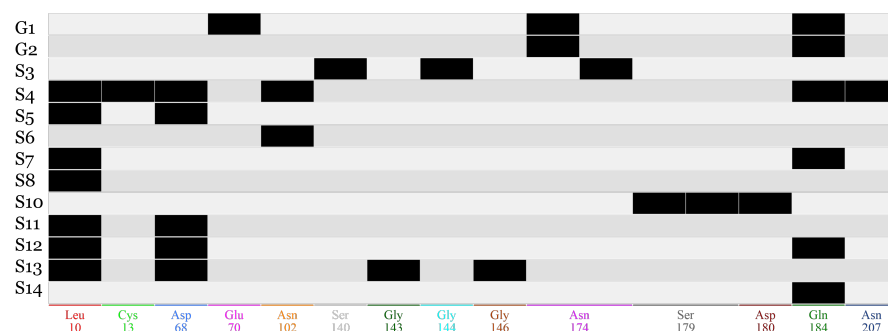| Compound	Number of runs	Score (kcal/mol)	Equivalent K_d (μM)
S3	1	-8.23 ± 0.48	2.11 ± 1.51
S4	4	-7.58 ± 0.50	5.85 ± 3.20
S5	1	-9.42 ± 0.63	0.39 ± 0.39
S6	2	-8.96 ± 0.74	0.87 ± 0.80
S7	2	-8.88 ± 0.83	1.27 ± 1.42
S8	3	-8.29 ± 0.58	2.02 ± 1.24
S10	5	-8.59 ± 0.42	1.10 ± 0.68
S11	4	-9.29 ± 0.72	0.45 ± 0.29
S12	1	-9.04 ± 0.82	0.75 ± 0.49
S13	2	-8.68 ± 0.51	0.99 ± 0.54
S14	3	-7.94 ± 0.41	3.04 ± 1.47

Figure 14: PLIF of gatastatin and its derivatives in the GDP binding site.

Consensus Docking: putative colchicine site

Three runs were required to obtain a consensus for colchicine. The average score value on the cluster was -6.01 ± 0.05 kcal/mol, which resulted in a mean K_d of 58.16 ± 4.99 μM.

Table 14: Docking scores in the binding sites from Site Finder.

Compound	Site 2		Site 3	
	Score (kcal/mol)	K_d (μM)	Score (kcal/mol)	K_d (μM)
G1	−7.46	5.50	−7.43	5.78
G2	−7.14 ± 0.11	9.46 ± 1.69	−8.35	1.30

Regarding gatastatin G1 and G2, a single run was enough to identify a consensus among the three programs. Their mean scores and K_d values are listed in table 15. Overall, they showed similar affinities, and both seem to have a higher affinity to γ than colchicine in this site.

Table 15: G1 and G2 binding affinities to the putative colchicine site on γ tubulin.

Compound	Score (kcal/mol)	Equivalent K_d (μM)
Gatastatin G1	−7.31 ± 0.24	7.48 ± 2.79
Gatastatin G2	−7.26 ± 0.20	8.08 ± 2.91

Blind Docking

As shown in figure 15, which shows MOE results, the vast majority of the poses resulting from docking simulations performed in the absence of the co-complexed GDP were located within its binding site, which is shaded in red. A gatastatin G2 pose was also arranged near the putative colchicine site. Few gatastatin G2 poses are located in the Site Finder site 2, shown in green, while a single G1 pose is docked into site 3 (blue). Vina and AutoDock agreed with MOE in identifying the GDP cleft as the primary interaction site, as the best scoring poses are located there. However, some more variability was observed as more poses arranged in Site Finder sites 2 and 3 and identified the putative colchicine site, too.

5.3.2 *α-β Tubulin*

Consensus Docking

As stated previously, consensus docking simulations on the *α-β* dimer were entirely focused on the colchicine binding site.
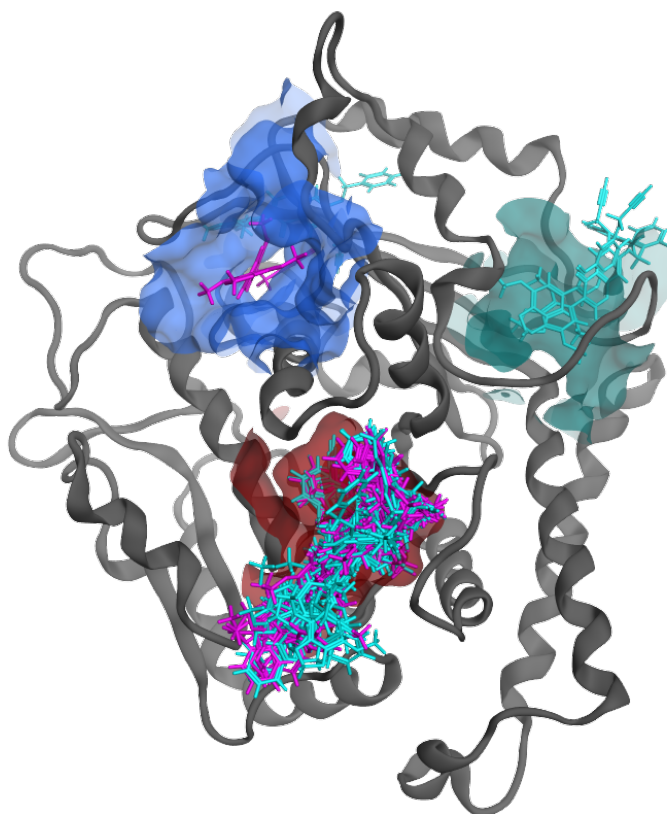
Figure 15: G1 and G2 blind docking poses on γ tubulin.

Redocking of colchicine resulted in a clear inter-program consensus, and all redocked poses were highly similar to the co-complexed one (RMSD < 1 Å). Mean binding energy values were around −9.9 kcal/mol (K_d around 0.14 μM) for all receptors.

Table 16 provides an overview of the number of runs required to obtain a consensus for each ligand in the nine human tubulin models. All gatastatin derivatives were investigated. Such compounds as gatastatin G1, S3, S7, and S8 gave a clear consensus within the first five runs in all receptors while other ligands, especially S13-15, did not yield a consensus for most α-β structures.

As they are considered important in glioblastoma multiforme (ref. chapter 2), consensus docking results for tubulin α-βIIa, III, and IVa are discussed more in-depth. Their average scores and K_d values are reported in table 17. A consensus for compounds S11, 13, 14, and 15 was not frequently found. The other ligands have similar predicted affinities, with S4 on the lowest end of the range of affinities and S10 on the highest one. It should also

be noted that a consensus for gatastatin G2 was reached only by relaxing the RMSD threshold for both α-βIIa and III.

Results for all other α-β structures are reported in the Supplementary Material (A) in tables 33-38.

Blind Docking

Similarly to what was observed for γ tubulin, gatastatin docking on the α-β polymerised dimer in the absence of the GDP molecule in the exchangeable site on β resulted in the majority of the docked poses located in this same cleft. Figure 16 illustrates these results for α-βIII; similar results were obtained for the other models. Gatastatin G1 is depicted in magenta, G2 in cyan. The GTP molecule at the non-exchangeable site is depicted in red.
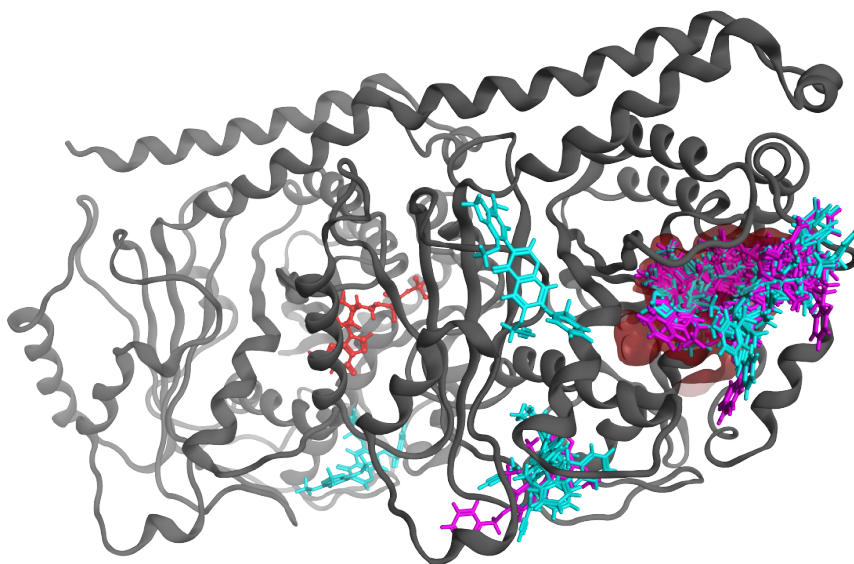

Figure 16: Blind docking results on α-βIII.

Table 16: Number of runs needed to obtain a consensus for docking on the α-β dimer.

Compound	$\alpha\beta$I	$\alpha\beta$IIa	$\alpha\beta$IIb	$\alpha\beta$III	$\alpha\beta$IVa	$\alpha\beta$IVb	$\alpha\beta$V	$\alpha\beta$VI	$\alpha\beta$VIII
G1	2	2	3	1	3	3	2	5	1
G2	2	4+	4+	5+	1	2	5	✖	4+
S3	3	2	1	2	3	3	1	2+	3
S4	5	6	1	4	4	1	✖	✖	2
S5	7+	✖	✖	2	3	4+	3	✖	✖
S6	✖	6	✖	1	4+	1	8	2	3+
S7	3	2	1	3	1	1	1	2	1
S8	3	4	2	2	1	2	1	4	2
S10	6	7	3	3	5+	3	4	8	3
S11	✖	5	7+	✖	✖	✖	4	✖	✖
S12	5	3	6	2	1	1	2	5+	2
S13	✖	✖	✖	✖	✖	✖	✖	✖	2+
S14	✖	✖	✖	✖	✖	3	✖	✖	✖
S15	6	✖	✖	8	✖	8	4+	✖	6

Legend
- ● $1 \leq$ #runs < 4
- ● $4 \leq$ #runs < 7
- ● $7 \leq$ #runs ≤ 8 or relaxed threshold (+)
- ✖ no consensus found

Table 17: Results of consensus docking on tubulins α-βIIa, III, and IVa.

Compound	$\alpha\beta$IIa Score (kcal/mol)	$\alpha\beta$IIa K_d (μM)	$\alpha\beta$III Score (kcal/mol)	$\alpha\beta$III K_d (μM)	$\alpha\beta$IVa Score (kcal/mol)	$\alpha\beta$IVa K_d (μM)
G1	-8.97 ± 0.58	0.74 ± 0.65	-9.02 ± 0.65	0.79 ± 0.82	-8.99 ± 0.60	0.73 ± 0.66
G2	-9.07 ± 0.29	0.45 ± 0.21	-9.56 ± 0.11	0.18 ± 0.035	-9.31 ± 0.43	0.35 ± 0.27
S3	-9.03 ± 0.44	0.53 ± 0.27	-8.90 ± 0.44	0.65 ± 0.35	-8.83 ± 0.60	0.84 ± 0.52
S4	-8.20 ± 1.39	5.51 ± 4.07	-7.77 ± 1.28	9.99 ± 7.67	-8.25 ± 1.42	5.58 ± 4.73
S5	-	-	-8.37 ± 0.20	1.31 ± 0.38	-8.90 ± 0.06	0.53 ± 0.05
S6	-9.93 ± 0.45	0.13 ± 0.10	-8.98 ± 0.52	0.66 ± 0.53	-9.63 ± 0.68	0.28 ± 0.25
S7	-9.05 ± 0.36	0.48 ± 0.21	-8.73 ± 0.29	0.78 ± 0.35	-9.03 ± 0.43	0.53 ± 0.26
S8	-9.49 ± 0.70	0.40 ± 0.44	-9.36 ± 0.80	0.55 ± 0.61	-9.03 ± 0.61	0.63 ± 0.44
S10	-10.26 ± 0.54	0.09 ± 0.08	-10.00 ± 0.57	0.14 ± 0.13	-9.2 ± 0.40	0.40 ± 0.25
S11	-10.37 ± 0.45	0.06 ± 0.05	-	-	-	-
S12	-9.20 ± 0.24	0.35 ± 0.15	-8.15 ± 0.38	2.15 ± 1.20	-8.9 ± 0.59	0.76 ± 0.52
S13	-	-	-	-	-	-
S14	-	-	-	-	-	-
S15	-	-	-9.41 ± 0.79	0.53 ± 0.62	-	-

ADMET PROPERTIES ANALYSIS

This section focuses on the analysis of the pharmacokinetic properties of gatastatin and its derivatives.

6.1 MATERIALS AND METHODS

Physicochemical descriptors, pharmacokinetic properties and drug-likeness of gatastatin and its derivatives were predicted using the commercial software ADMET Predictor 10.2 by Simulations Plus and the online resources SwissADME [14] and pkCSM [15].
The results from the three tools were then compared.

6.1.1 *SwissADME*

SwissADME is an online tool developed by the Swiss Institute of Bioinformatics to compute physicochemical descriptors and pharmacokinetics of one or multiple small molecules.
Inputs can be provided either in a list of SMILES or by directly drawing the molecule in a specific window.
Upon prediction completion, a panel summing up its properties will appear for each compound. The properties are divided as follows:

GRAPHICAL OVERVIEW: 2D depiction of the compound, canonical SMILES and *Bioavailability Radar*, shown in figure 17. This first section gives a graphical description of the molecule and its drug-likeness.
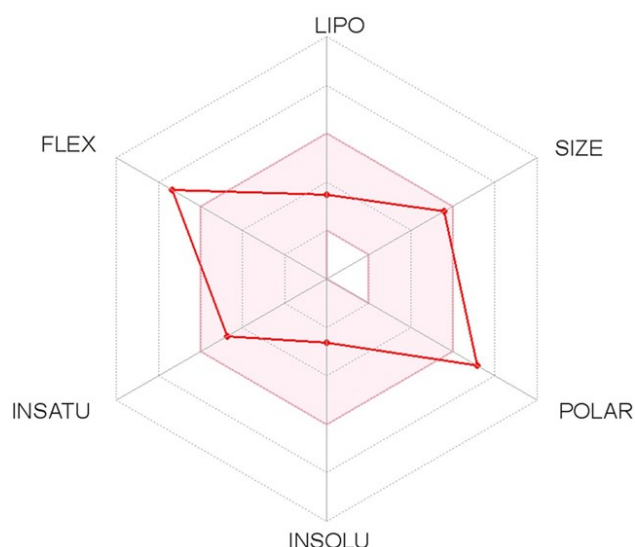
Figure 17: Example of SwissADME Bioavailability Radar.

The pink area represents the optimal range for each of the listed properties:

- Lipophilicity: XLOGP3 between -0.7 and +5.0;
- Size: molecular weight between 150 and 500 g/mol;
- Polarity: Topological Polar Surface Area (TPSA) between 20 and 130 Å^2;
- Solubility: logS lower than 6;
- Saturation: fraction of carbons in the sp^3 hybridization not less than 0.25;
- Flexibility: no more than 9 rotatable bonds.

The compound in the example of figure 17 is predicted not to be orally available because it is too flexible and too polar [24].

PHYSICOCHEMICAL PROPERTIES:

- Formula
- Molecular weight
- Number of heavy atoms
- Number of heavy aromatic atoms
- Fraction of carbons in the sp^3 hybridisation
- 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: results from five different partition coefficient predictors are reported.

- iLOGP: implicit logP method developed by the Swiss Institute of Bioinformatics. It is a physics-based multiple linear model based on two GB/SA (Generalized Born/Solvent Accessible Surface Area, among the most commonly used implicit solvent model combinations) parameters [23].
- XLOGP3: atomistic and knowledge-based approach calculated by the XLOGP program, version 3.2.2, provided by the Shanghai Institute of Organic Chemistry.
- WLOGP: atomistic method based on the atom type classification system developed in 1999 by S. A. Wildman and G. M. Crippen [59].
- MLOGP: predictive structure-log P model obtained by multiple regression analysis. The topological method provided by SwissADME was implemented from [48, 47, 43].
- SILICOS-IT: hybrid fragmental/topological method calculated by FILTER-IT program, version 1.0.2, provided by SILICOS-IT.

A consensus is calculated by averaging the five predictors.

WATER SOLUBILITY: results from three different approaches are provided.

- ESOL: it stands for *E*stimated *SOL*ubility. It is a method for estimating the aqueous solubility of a molecule directly from its structure. The model was implemented based on the work of J. S. Delaney [25].
- Ali: topological method implemented from the study of Ali *et al.* [16].
- SILICOS-IT: a fragmental method calculated by FILTER-IT program, version 1.0.2, provided by SILICOS-IT.

Along with the log S value, a *Solubility Class* is given for each predictor. In log S scale, it is calculated as follows:

Insoluble < − 10 < *Poorly* < −6 < *Moderately* < −4 <
< *Soluble* < − 2 < *Very* < 0 < *Highly*

PHARMACOKINETICS: this section lists such ADME properties as gastrointestinal (GI) absorption, skin permeation, interaction with cytochromes P450 and permeability glycoprotein (P-gp), and blood-brain barrier (BBB) permeation.

Interactions with P-gp and cytochromes are predicted through an SVM (Support Vector Machine) model, i.e. a machine learning classifier, while skin permeation is provided as $\log K_p$, where K_p indicates the skin permeability, i.e. the rate of a chemical penetrating the *stratum corneum*. It is calculated with a Quantitative Structure-Property Relationship (QSPR) model implemented from [51].

DRUGLIKENESS ESTIMATIONS: qualitative assessments of the chance for a molecule to become an oral drug with respect to bioavailability, based on several filters:

Lipinski (Pfizer) filter, implemented from [43]:
- Molecular weight ≤ 500
- MLOGP ≤ 4.15
- Number of hydrogen bond acceptors (expressed as the sum of nitrogen and oxygen atoms) ≤ 10
- Number of hydrogen bond donors (expressed as the sum of NH and OH groups) ≤ 5

Ghose filter, implemented from [32]:
- 160 ≤ molecular weight ≤ 480
- -0.4 ≤ WLOGP ≤ 5.6
- 40 ≤ molar refractivity ≤ 130
- 20 ≤ number of atoms ≤ 70

Veber (GSK) filter, implemented from [57]:
- Number of rotatable bonds ≤ 10
- TPSA ≤ 140

Egan (Pharmacia) filter, implemented from [27]:
- WLOGP ≤ 5.88
- TPSA ≤ 131.6

Muegge (Bayer) filter, implemented from [49]:
- 200 ≤ molecular weight ≤ 600
- -2 ≤ XLOGP ≤ 5
- TPSA ≤ 150
- Number of rings ≤ 7
- Number of carbon atoms > 4
- Number of heteroatoms > 1
- Number of rotatable bonds ≤ 15
- Number of hydrogen bond acceptors ≤ 10
- Number of hydrogen bond donors ≤ 5

Along with the Abbott bioavailability score, which predicts a compound's probability of having at least 10% oral bioavailability in rats [46].

MEDICINAL CHEMISTRY: parameters that help to understand whether a molecule is suitable for optimisation.

- PAINS: Pan Assay Interference Structures are chemical compounds that often give false-positive results in high-throughput screens.
- Brenk: Structural Alert consists in a list of 105 fragments identified by Brenk *et al.* [20] to bear properties responsible for poor pharmacokinetics such as toxicity, chemical reactivity, and metabolic instability.
- Leadlikeness filter: a compound passes this test if it has a molecular weight between 250 and 350 g/mol, XLOGP ≤ 3.5, and 7 or less rotatable bonds. [55]
  The concept of leadlikeness is similar to drug-likeness yet focusing on physicochemical boundaries defining a good lead, i.e. a molecular entity suitable for optimisation. Leads are subjected to chemical modifications that will most likely increase size and lipophilicity; therefore, they are required to be smaller and less hydrophobic than drug-like molecules.
- Synthetic accessibility: this calculation provides a score from 1 (very easy) to 10 (very difficult).

When all inputs have been processed, a Show BOILED-Egg (acronym for Brain Or IntestinaL EstimateD permeation) button appears. The resulting image shows the graphical output of two important ADME parameters: the passive gastrointestinal absorption and the BBB permeation. It is an egg-shaped classification plot in which the yolk represents the physicochemical space for highly probable BBB permeation and the white is for GI absorption. The Egg is based on just two descriptors, namely WLOGP for lipophilicity and TPSA for apparent polarity. Moreover, colour-coding gives information about active efflux from the brain or the GI lumen: blue for P-gp substrates and red for non-substrates.

6.1.2 *pkCSM*

pkCSM is another tool available online. Its calculations rely on graph-based signatures: chemical entities can be mathematically

represented through a graph model having atoms as nodes and covalent bonds as edges. The graph can then be decorated with labels indicating physicochemical properties of said atoms and bonds, and structural patterns can be identified. Several descriptors can thus be extracted from the graph model of a compound; these descriptors, along with experimental data on ADMET properties, can be used to train predictive models via machine learning methods. pkCSM calculates physicochemical properties (molecular weight, partition coefficient, *et cetera*) and 30 such predictors, divided into five major classes:

ABSORPTION

Caco-2 cells are a human epithelial colorectal adenocarcinoma cell line. A monolayer of these cells is used as an in vitro model of the human intestinal mucosa to predict oral drugs absorption.

- Caco-2 permeability: predicts the logarithm of the apparent permeability coefficient (log P_{app}, where P_{app} is expressed in cm/s); it is considered high if the predicted values are greater than 0.90.
- Intestinal absorption: the predicted proportion of the compound absorbed through the small intestine is given. A compound with absorbance less than 30% is predicted to be poorly absorbed.
- Water solubility: the log S value is given, i.e. the 10-based logarithm of the molar concentration.
- P-glycoprotein substrate and P-glycoprotein I and II inhibitor: a categorical (yes/no) value is provided.
- Skin permeability: the log K_p value is provided. A log K_p value greater than 2.5 indicates a relatively low skin permeability.

DISTRIBUTION

- VDss (Volume of Distribution at a steady-state): log VDss is provided. The predicted value is considered to be low if below -0.15 (VDss < 0.71 L/kg) and high if above 0.45 (VDss > 2.81 L/kg).
- Fraction unbound (Fu): the fraction that would be unbound in plasma is predicted.
- BBB permeability: this parameter is provided as log BB, i.e. the logarithmic ratio of brain-to-plasma drug concentrations. Compounds having log BB < -1 are predicted to be poorly distributed to the brain, while log BB > 0.3 indicates a rapid crossing of the barrier.
- CNS permeability: the log PS, i.e. blood-brain permeability-surface area product, is calculated. A compound having

log PS < -3 is predicted to be unable to penetrate the central nervous system, while log PS values above -2 indicate CNS permeation.

METABOLISM: a categorical value (yes/no) predicts whether the compound is likely to be an inhibitor of the isoforms CYP1A2, CYP2C19, CYP2C9, CYP2D6, CYP3A4 of the cytochrome P450, or a substrate of the CYP2D6 and CYP3A4 isoforms. A compound is considered a CYP450 inhibitor if a concentration under 10 μM results in 50% inhibition.

EXCRETION

- Renal organic cation transporter 2 (OCT2) substrate: a categorical value indicating whether the molecule is likely to be an OCT2 substrate.
- Total clearance: the prediction is given as the logarithm of the total clearance CL_{tot}, which is expressed in mL/min/kg.

TOXICITY

- Rat LD50: given in mol/kg
- AMES toxicity: mutagenicity prediction (categorical value)
- *T. Pyriformis* toxicity: the prediction is given in log μg/L; a value above -0.5 is considered toxic.
- Minnow toxicity: the model calculates the logarithm of LC50. Values below -0.3 (LC50 < 0.5 mM) indicate high acute toxicity.
- Maximum Recommended Tolerated Dose (MRTD): in logarithmic scale, MRTD ≤ 0.477 log(mg/kg/day) is regarded as low, while a MRTD above 0.477 is considered high.
- Oral rat chronic toxicity, given in log(mg/kg$_{bw}$/day)
- Hepatotoxicity: categorical value
- Skin sensitization: categorical value
- Cardiotoxicity: hERG I and II inhibitors (categorical value)

Inputs can be provided either in a SMILES file or by simply entering a single SMILES string. The user is free to decide which of the ADMET properties he wants to obtain as output.

6.1.3 *ADMET Predictor*

Simulations Plus ADMET Predictor is a state-of-the-art ADMET property prediction software. It is a machine learning software

The experimental derivation of PS is a more direct measurement of brain penetration: since it is obtained from in situ brain perfusions by injecting the compound directly into the carotid artery, it lacks the systemic distribution effects that may distort brain penetration measurements.

tool that quickly and accurately predicts over 175 pharmacokinetic properties starting simply from the 2D structure of the compound.

Loading a SMILES file containing the ligands of interest is enough to get started. Upon selecting the ADMET button in the main window, the software will prompt the selection of the desired features, and the Calculate button will start the calculations. The resulting table can be exported as an Excel spreadsheet.

Along with most of the properties also provided by the other two tools, Simulations Plus offers some interesting cumulative scores linked to the compound's risk of failing as a drug due to ADMET issues and the related codes explaining which these issues are. These scores are:

- Absn_Risk: a score between 0 and 8, indicating the number of potential oral absorption problems the compound is likely to have.
- RuleOf5: a score indicating the number of potential problems a compound is expected to have with passive oral absorption, according to Lipinski's Rule of Five.
- CYP_Risk: risk connected with cytochromes P450 oxidation. The score is in the range 0-6 and indicates the number of potential problems the compound might have due to metabolism by one or more of the five major cytochrome P450 isoforms.
- MUT_Risk: a score in the 0-5.4 range indicating the risk of mutagenicity.
- TOX_Risk: a score in the 0-6 range indicating the number of predicted toxicity problems.

The ADMET_Risk prediction aggregates all these results in a full ADMET risk score. This score is in the range 0-22 and indicates the number of potential ADMET problems the investigated compound may have.

6.2 RESULTS

This section presents the most relevant ADMET properties for the analysis of gatastatin and its derivatives. Where applicable, the outputs of the three predictors are compared. Given its reliability, Simulations Plus is regarded as a golden standard and is used as a means of comparison for the other two tools.

6.2.1 *Absorption*

Overall, the compounds were predicted moderately to poorly soluble, with some variability among the results of the different tools. As shown in table 18, which reports the log S values, most results from SwissADME seem to be in line with the Simulations Plus predictor, while pkCSM differs some more. However, it must be noted that a consensus is not achieved even among the three predictors provided by SwissADME (i.e. ESOL, Ali, and SILICOS-IT), with the results from the SILICOS-IT tool being the ones that differ the most.

Table 18: Log S values.

Compound	Simulations Plus	pkCSM	ESOL	Ali	SILICOS-IT
G1	−6.50	−5.54	−5.48	−5.88	−8.79
G2	−6.66	−4.32	−5.61	−6.01	−8.86
S3	−6.72	−3.99	−5.34	−5.56	−8.59
S4	−5.68	−4.25	−5.51	−6.02	−6.53
S5	−6.79	−3.93	−6.02	−6.53	−9.52
S6	−6.77	−3.65	−6.33	−6.97	−9.22
S7	−6.20	−3.68	−5.27	−5.78	−8.10
S8	−6.59	−4.37	−5.89	−6.56	−9.22
S10	−7.64	−3.60	−6.80	−7.44	−11.24
S11	−7.04	−3.80	−6.87	−7.61	−9.95
S12	−6.47	−3.90	−5.80	−6.42	−8.84
S13	−6.90	−4.65	−6.42	−7.19	−9.96
S14	−6.97	−4.60	−6.14	−6.64	−9.60
S15	−7.88	−3.77	−7.33	−8.07	−11.96

According to pkCSM, all compounds are well absorbed by the human intestine since their predicted absorbed percentage is well above 30%, as shown in figure 18.

Results of the prediction of the interaction with P-glycoprotein are reported in table 19, which reports Yes if the compound is likely to be a P-glycoprotein substrate, and No otherwise. There is a reasonable degree of agreement between Simulations Plus and SwissADME, while the results from pkCSM are almost entirely unrelated.

Regarding the prediction of P-glycoprotein inhibitors, on the other hand, there is a complete consensus between Simulations Plus and pkCSM, that both predict all compounds to be P-gp inhibitors, while SwissADME does not provide this parameter.

To sum up, according to Simulations Plus ADMET predictor, all investigated compounds are likely to be both substrates and inhibitors of P-gp.
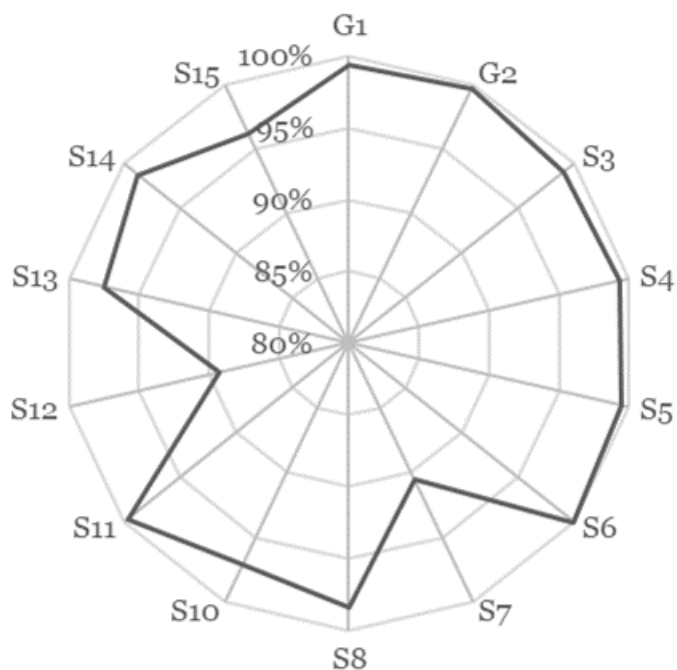

Figure 18: Intestinal absorption as calculated by pkCSM.

Table 19: P-glycoprotein substrate prediction.

Compound	Simulations Plus	SwissADME	pkCSM
G1	Yes	Yes	No
G2	Yes	Yes	No
S3	Yes	Yes	No
S4	Yes	Yes	No
S5	Yes	Yes	No
S6	Yes	Yes	No
S7	Yes	Yes	Yes
S8	Yes	Yes	No
S10	Yes	No	Yes
S11	Yes	Yes	No
S12	Yes	Yes	Yes
S13	Yes	Yes	No
S14	Yes	Yes	No
S15	Yes	Yes	No

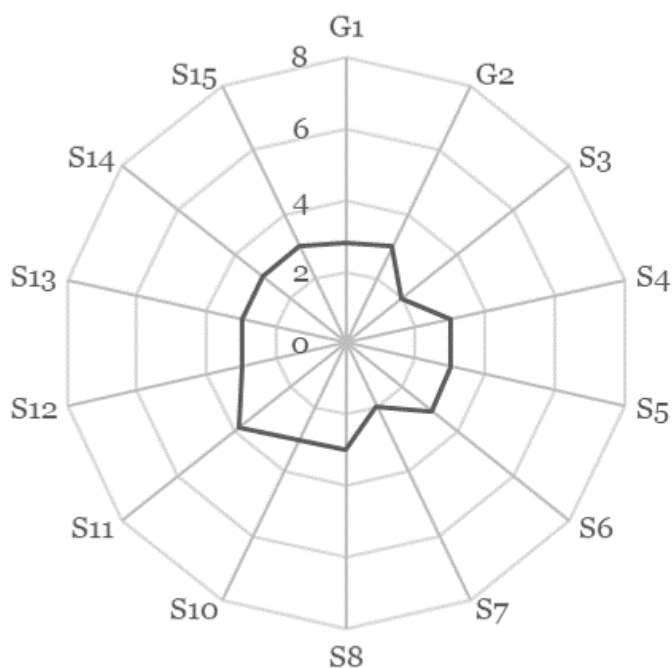Lastly, Absn_Risk and RuleOf5 scores by Simulations Plus are reported in figure 19 and tables 20 and 21.

Figure 19: Absn_Risk scores.

Table 20: Absn_Risk codes.

Compound	Absn_Risk code
G1	Size, Lipophilicity, Water solubility
G2	Size, Lipophilicity, Water solubility
S3	Lipophilicity, Water solubility
S4	Size, Lipophilicity, Water solubility
S5	Size, Lipophilicity, Water solubility
S6	Size, Lipophilicity, Water solubility, H-bond acceptors
S7	Size, Lipophilicity, Water solubility
S8	Size, Lipophilicity, Water solubility
S10	Size, Lipophilicity, Water solubility
S11	Size, Lipophilicity, Water solubility, H-bond acceptors
S12	Size, Lipophilicity, Water solubility
S13	Size, Lipophilicity, Water solubility
S14	Size, Lipophilicity, Water solubility
S15	Size, Lipophilicity, Water solubility

Table 21: Lipiski's Rule of Five violations.

Compound	RuleOf5_Risk score	RuleOf5_Risk code
G1	0	
G2	0	
S3	0	
S4	0	
S5	0	
S6	1	Molecular weight
S7	0	
S8	0	
S10	1	Molecular weight
S11	1	Molecular weight
S12	0	
S13	1	Molecular weight
S14	1	Molecular weight
S15	1	Molecular weight

6.2.2 *Distribution*

Volume of distribution (VDss) and fraction unbound (Fu) values from Simulations Plus and pkCSM are reported in tables 22 and 23. SwissADME does not provide information about these parameters.

Again, the results of pkCSM highly differ from the values provided by Simulations Plus.

As stated in section 6.1.2, pkCSM provides some guidelines to interpret numeric values: log VDss is low if below -0.15 and high if above 0.45. According to these rules, all VDss values as calculated by pkCSM would be low; those provided by Simulations Plus would be much higher, but not so much as to be considered high by the pkCSM guidelines.

Blood-brain barrier permeation is among the most critical parameters for this application. Table 24 presents a comparison between the values of log BB provided by Simulations Plus and pkCSM. As previously noted for other predictors, there is some variability between the two tools, both in pairwise comparisons and in the overall range of the values.

According to the pkCSM guidelines, a log BB lower than -1 indicates poor distribution to the brain, while compounds having log BB > 0.3 are predicted to cross the BBB readily. None of the analysed compounds shows a high capability to cross the

blood-brain barrier, as most of them have a log BB value smaller than -1; the highest log BB value, -0.111 of the derivative S3, is still quite far from the 0.3 threshold.

Table 22: Log VDss predictions.

Compound	Simulations Plus	pkCSM
G1	0.139	−0.470
G2	0.164	−0.555
S3	0.182	−0.335
S4	0.175	−0.606
S5	0.208	−0.451
S6	0.120	−0.601
S7	0.234	−0.542
S8	0.211	−0.506
S10	0.129	−1.023
S11	0.191	−0.526
S12	0.322	−0.484
S13	0.279	−0.411
S14	0.235	−0.464
S15	0.203	−0.950

Table 23: Fraction unbound predictions.

Compound	Simulations Plus	pkCSM
G1	5.1%	25.4 %
G2	4.0%	22.1 %
S3	3.6%	26.3 %
S4	5.3%	14.8 %
S5	5.1%	20.5 %
S6	4.4%	26.6 %
S7	4.8%	22.5 %
S8	4.7%	22.3 %
S10	3.1%	35.6 %
S11	4.4%	24.2 %
S12	4.8%	19.9 %
S13	4.7%	19.7 %
S14	4.1%	19.5 %
S15	3.1%	34.2 %

Table 24: Comparison between the Simulations Plus and pkCSM logBB predicted values.

Compound	Simulations Plus	pkCSM
G1	−0.589	−1.244
G2	−0.716	−1.182
S3	−0.515	−0.111
S4	−0.715	−1.234
S5	−0.558	−1.195
S6	−0.607	−1.399
S7	−0.652	−0.863
S8	−0.659	−1.226
S10	−0.374	−1.232
S11	−0.584	−1.389
S12	−0.641	−0.820
S13	−0.627	−1.217
S14	−0.692	−1.173
S15	−0.353	−1.222

Table 25: Log PS values calculated by pkCSM, indicating CNS penetration.

Compound	Log PS
G1	−3.010
G2	−3.070
S3	−2.052
S4	−3.197
S5	−3.082
S6	−3.339
S7	−3.234
S8	−3.047
S10	−2.787
S11	−3.212
S12	−3.107
S13	−2.929
S14	−2.952
S15	−2.669

In addition to the log BB value, pkCSM provides a log PS value that may hold a more truthful measure of the capability to penetrate the brain. Results from this predictor are shown

in table 25. Most of the compounds have a log PS value under -3, indicating that they cannot penetrate the brain. Again, the highest value is the log PS of derivative S3, closer to the -2 threshold, indicating good CNS permeability.

SwissADME summarises GI absorption, BBB permeation, and P-glycoprotein interaction predicted values in the Boiled-EGG shown in figure 20.

The results of the BBB permeability predictions from SwissADME are in line with the calculations by pkCSM in the fact that the only derivative that is predicted to be able to cross the barrier is the one having the highest log BB value in pkCSM. Again, no compound is predicted to be able to cross the blood-brain barrier readily.

Regarding intestinal absorption, there is a consensus between SwissADME and pkCSM on all compounds except S11 and S15; SwissADME predicts these two molecules to be poorly absorbed, while according to pkCSM (ref. to figure 18), their absorbed percentage is high.
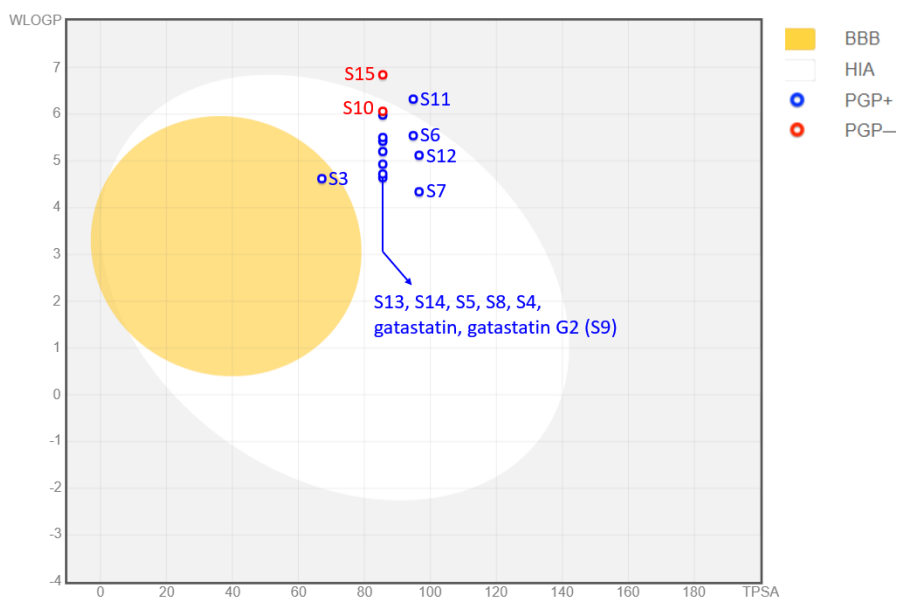

Figure 20: Boiled-EGG calculated by SwissADME.

6.2.3 *Metabolism*

All three tools provide a prediction on whether the investigated compounds are inhibitors of the five major cytochromes P450 through a categorical value. The results are compared in table 26.

For most compounds, there is some degree of agreement among the predictions by the three tools on CYP2C19 and CYP3A4. Regarding CYP1A2 and CYP2C9, SwissADME and pkCSM agree on most compounds but provide a result opposite to Simulations Plus. Lastly, results for cytochrome CYP2D6 are a combination of the cases described previously.

Table 26: Cytochromes P450 inhibitors.

Compound	CYP1A2	CYP2C19	CYP2C9	CYP2D6	CYP3A4
G1	▲🟥●	▲■●	▲■●	▲■●	▲■●
G2	▲🟥●	▲■●	▲■●	▲■●	▲■●
S3	▲■●	▲■●	▲■●	▲■●	▲■●
S4	▲🟥●	▲■●	▲■●	▲■●	▲■●
S5	▲🟥●	▲■●	▲■●	▲■●	▲■●
S6	▲🟥●	▲■●	▲■●	▲■●	▲■●
S7	▲🟥●	▲■●	▲■●	▲■●	▲■●
S8	▲🟥●	▲■●	▲■●	▲🟥●	▲■●
S10	▲🟥●	▲■●	▲■●	▲■●	▲■●
S11	▲🟥●	▲🟥●	▲🟥●	▲■●	▲🟥●
S12	▲🟥●	▲■●	▲■●	▲■●	▲🟥●
S13	▲🟥●	▲■●	▲■●	▲■●	▲🟥●
S14	▲🟥●	▲■●	▲■●	▲■●	▲🟥●
S15	▲🟥●	▲■●	▲■●	▲■●	▲🟥●

Legend:	▲ Simulations Plus	■ SwissADME	● pkCSM
Colour coding:	Yes / No		

pkCSM provides a prediction on whether the compounds are CYP2D6 or CYP3A4 substrates. The tool predicted all compounds to be substrates of CYP3A4 and not of CYP2D6. These results agree with the Simulations Plus software except for the S3 derivative, which is predicted to be a CYP2D6 substrate by Simulations Plus.

Figure 21 reports the results of the CYP_Risk score prediction. All compounds have an alert about high microsomal clearance, and most of them are predicted to have a high CYP1A2 clearance. Derivative S3, which has the highest CYP_Risk score, has also an alert about CYP2D6 clearance.

Figure 21: Simulations Plus CYP_Risk.

6.2.4 *Excretion*

Table 27: OCT2 substrates.

Compound	Simulations Plus	pkCSM
G1	No	No
G2	Yes	No
S3	No	No
S4	Yes	No
S5	No	No
S6	No	No
S7	No	No
S8	No	No
S10	No	No
S11	No	No
S12	No	No
S13	No	No
S14	Yes	No
S15	No	No

Simulations Plus predicts the clearance mechanism to be primarily due to metabolism and not to renal or hepatic uptake, for all compounds.

Predictions on whether the compounds are likely to be Organic Cation Transporter 2 (OCT2) are compared in table 27, while table 28 reports the total clearance as calculated by pkCSM. Only results from Simulations Plus and pkCSM are reported, as SwissADME does not provide information about excretion.

Table 28: Total clearance.

Compound	Total Clearance log(mL/min/kg)
G1	0.246
G2	0.407
S3	0.249
S4	0.533
S5	0.523
S6	0.417
S7	0.307
S8	0.380
S10	0.337
S11	−0.199
S12	0.462
S13	0.536
S14	0.563
S15	0.493

6.2.5 *Toxicity*

Only Simulations Plus and pkCSM predictions are reported for toxicity, as SwissADME does not provide such information.

Animal-based assays derived indicators

Figure 22 summarises the results of Minnow toxicity predictions graphically. Once again, it is noticeable how often pkCSM's results differ from Simulations Plus. All compounds other than S7 and S12 are predicted by pkCSM to cause high acute toxicity, while results from Simulations Plus are much lower in all cases.

Regarding Ames toxicity, according to pkCSM, no compound is expected to be mutagenic, while Simulations Plus predicts gatastatin G2 and derivatives S7 and S14 to be Ames positive.
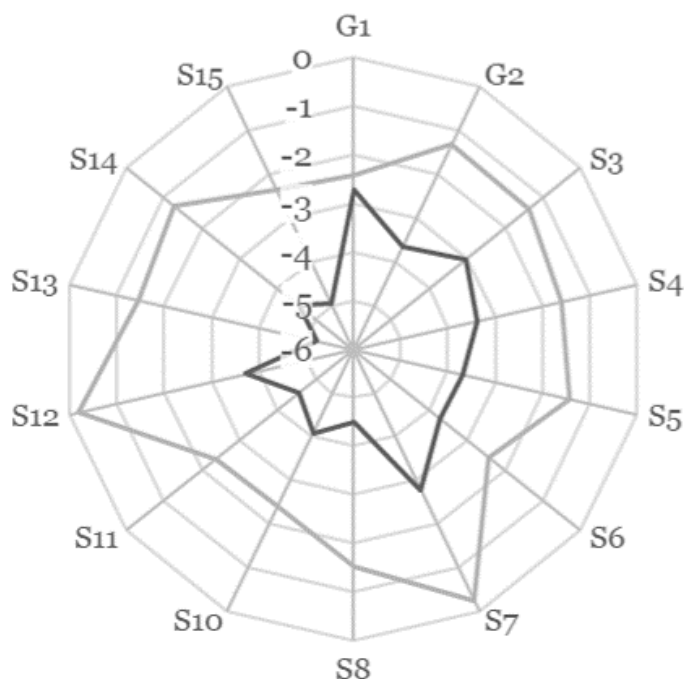
Figure 22: Comparison of Minnow toxicity predictions. Simulations Plus results are depicted in dark grey, while pkCSM is in light grey.

Toxicity predictors in humans

pkCSM provides predictions of the Maximum Recommended Tolerated Dose on a logarithmic scale. Simulations Plus reports the Maximum Recommended Therapeutic Dose qualitatively to indicate whether the MRTD is higher or lower than 3.16 mg/kg/day (circa 0.5 in logarithmic scale). These results are presented in table 29.

The two tools agree that no compound is considered likely to cause skin sensitisation or hepatotoxicity, except for the S10 derivative, which pkCSM predicts to be hepatotoxic.

Regarding cardiotoxicity, both tools provide a categorical value indicating potential hERG inhibition. pkCSM distinguishes between hERG I and II channels, while Simulations Plus refers simply to hERG. According to the Human Protein Atlas, hERG I, encoded by gene KCNH2, is almost 4-fold more expressed in cardiomyocytes than hERG II, encoded by KCNH6. Results of the cardiotoxicity evaluation are compared in table 30.

Lastly, table 31 report the Simulations Plus TOX_Risk and figure 23 shows the ADMET_Risk codes, summarising all Risk predictions.

Table 29: Maximum Recommended Tolerated and Therapeutic Dose predictions by pkCSM and Simulations Plus, respectively.

Compound	pkCSM log(mg/kg/day)	Simulations Plus above/below
G1	0.577	↑
G2	0.767	↓
S3	0.739	↑
S4	0.673	↑
S5	0.700	↑
S6	0.705	↑
S7	0.558	↑
S8	0.788	↓
S10	0.613	↑
S11	0.682	↓
S12	0.511	↑
S13	0.750	↓
S14	0.726	↓
S15	0.592	↓

Table 30: Cardiotoxicity prediction.

Compound	Simulations Plus	pkCSM	
G1	●	I	II
G2	●	I	II
S3	●	I	II
S4	●	I	II
S5	●	I	II
S6	●	I	II
S7	●	I	II
S8	●	I	II
S10	●	I	II
S11	●	I	II
S12	●	I	II
S13	●	I	II
S14	●	I	II
S15	●	I	II

Colour coding: Not inhibitor   Inhibitor

Table 31: Simulations Plus TOX_Risk predictions.

| Compound	TOX_Risk score	TOX_Risk code
G1	0.500	Carcinogenicity in rat
G2	2.065	Carcinogenicity in rat, Acute rat toxicity, Ames positive, hERG inhibition
S3	0.500	Carcinogenicity in rat
S4	0.00	-
S5	1.236	Carcinogenicity in rat, Acute rat toxicity, hERG inhibition
S6	1.900	Carcinogenicity in rat, Acute rat toxicity, hERG inhibition
S7	1.502	Carcinogenicity in rat, Acute rat toxicity, Ames positive
S8	1.500	Carcinogenicity in rat, Acute rat toxicity
S10	1.725	Carcinogenicity in rat, Acute rat toxicity, hERG inhibition
S11	2.000	Carcinogenicity in rat, Acute rat toxicity, hERG inhibition
S12	0.876	Carcinogenicity in rat, Acute rat toxicity
S13	1.500	Carcinogenicity in rat, Acute rat toxicity
S14	2.854	Carcinogenicity in rat, Acute rat toxicity, Ames positive, hERG inhibition
S15	2.000	Carcinogenicity in rat, Acute rat toxicity, hERG inhibition
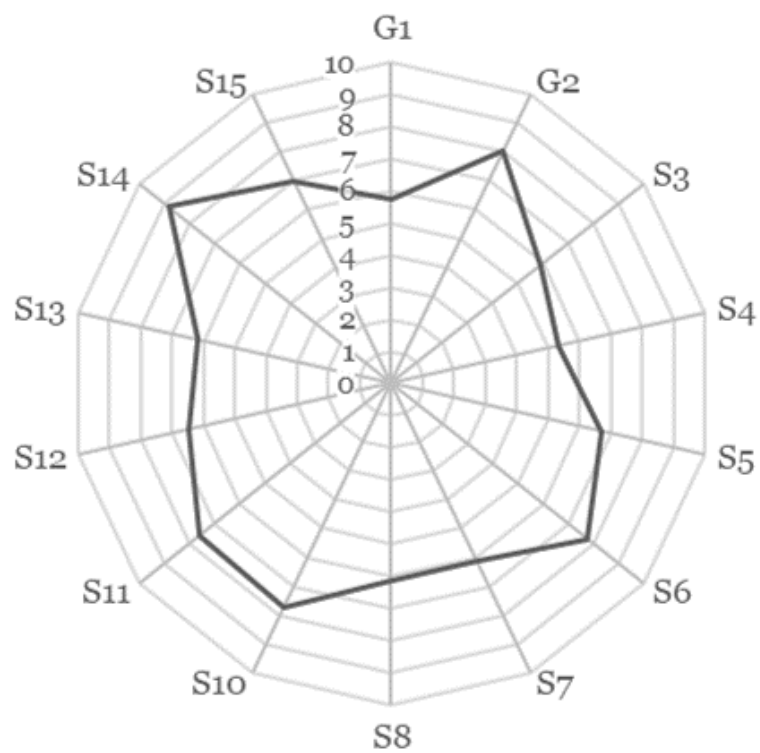
Figure 23: Full ADMET Risk scores. Original range: 0-22, truncated to 10 for clarity.

DISCUSSION

This concluding section summarises and integrates docking and AD-MET properties prediction results.

While not as accurate as other prediction methods, blind docking suggests gatastatin's preference towards the GTP binding site on γ tubulin. The K_d values calculated from consensus docking results support this hypothesis: as shown in table 32, the lowest values are those obtained in the cleft. The gatastatin G2 score in site 3 is the absolute least, but since it was not obtained from a consensus, it was considered less reliable than the average scores.

Table 32: Comparison of predicted K_d values of gatastatin G1 and G2 on γ tubulin identified sites.

Compound	K_d (μM)			
	GDP site	Putative colch. site	Site 2	Site 3
G1	1.64 ± 1.10	7.48 ± 2.79	5.50	5.78
G2	2.43 ± 1.56	8.08 ± 2.91	9.46 ± 1.69	1.30

Chinen *et al.* attributed gatastatin's specificity for γ tubulin to the O^7 benzyl group. The PLIF analysis performed on the consensus poses in the GDP binding site on γ tubulin gave similar results: figure 24 shows a G1 and a G2 pose, and the O^7 benzyl ring is highlighted in both, meaning it is involved in the interaction with the protein. As shown in figure 25 for gatastatin G2, the ring forms an arene-H interaction. The blue-shaded regions in figure 25 are solvent-exposed areas. Shintani *et al.* considered gatastatin G2 the most promising derivative based on a cytotoxicity analysis on HeLa cells [52] and identified the O^6 as an important position for γ tubulin inhibition. However, the propargyl group in position O^6 does not seem to interact with the receptor in any of the analysed docking poses.
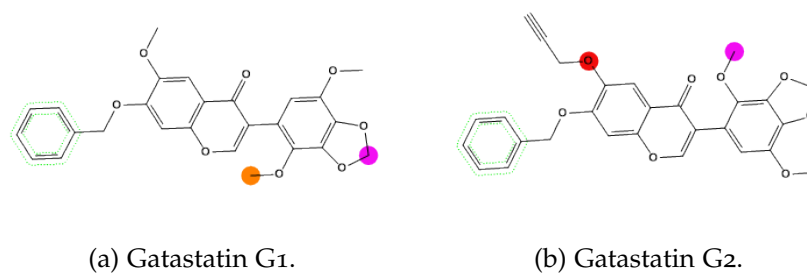
(a) Gatastatin G1.

(b) Gatastatin G2.

Figure 24: Substructures of G1 and G2 interacting with γ tubulin in the GDP site.
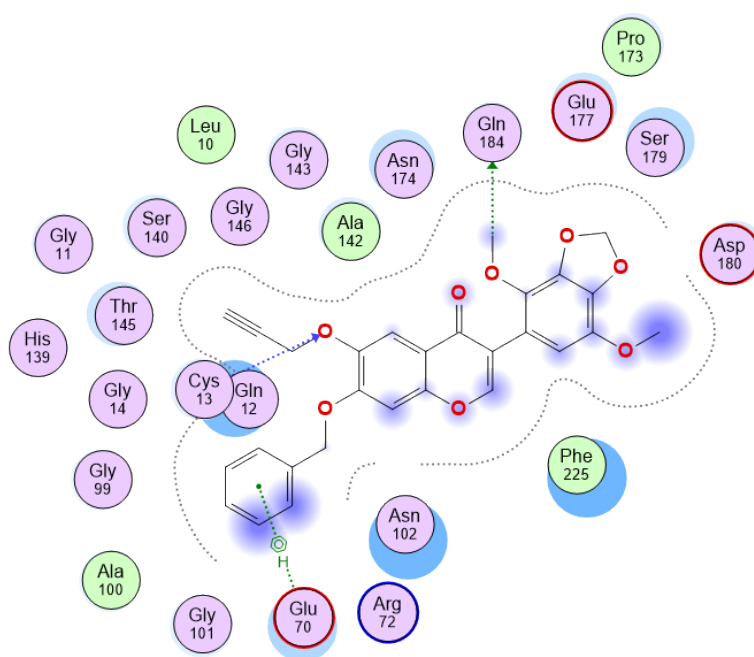

Figure 25: G2 interactions summary

According to the consensus docking results, gatastatin does not show a clear preference towards γ tubulin: the predicted binding affinities for α-β are similar, when not higher, than on γ tubulin. These results do not agree with the binding assay results by Chinen *et al.*, which showed gatastatin G1's 12-fold higher affinity for γ than to α-β.

Moreover, some results of the pharmacokinetic analysis showed that neither G1 nor G2 seem to be able to cross the blood-brain barrier, which would make them unsuitable candidates for a possible therapeutic application for glioblastoma multiforme. Also, Simulations Plus predicted G2 to be an hERG inhibitor.

Of all the analysed compounds, the only one predicted to cross the BBB is S3, which also has a better toxicity profile than G2. Its predicted binding affinity to γ in the GDP binding site is an intermediate value between gatastatin G1 and G2: S3 scored -8.23 ± 0.48 kcal/mol and the mean equivalent K_d is 2.11 ± 1.51 μM. Its interaction profile is shown in figure 26. Similarly to G2, the O^7 benzyl group makes a strong arene-H interaction, but it does not interact with the same residue as the molecule has a different orientation inside the cleft. As shown in figure 27, the O^7 benzyl rings of the two docked poses do not overlap. In agreement with Shintani *et al.* observations, as they do not establish interactions with the receptor, the O^6 position and the dioxolane seem suitable sites for derivatisation.

As shown in figure 28, S3 is the smallest of all gatastatin derivatives, leaving room for further derivatisation to improve its binding, but the results of BBB permeability (figure 20) suggest there is a limited scope to do so. In order to readily cross the blood-brain barrier, in fact, a molecule ought to be relatively small.
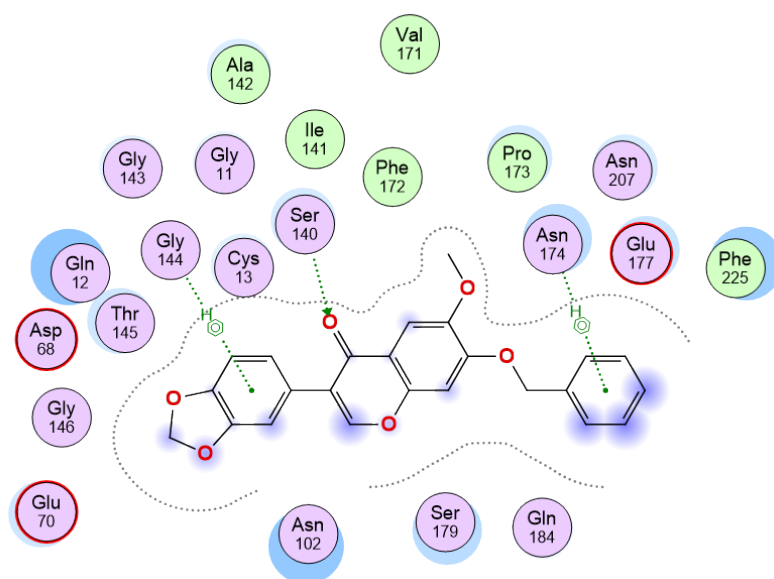

Figure 26: S3 interaction profile.

Comparison with available experimental data

Comparison of predicted colchicine and gatastatin binding affinities for α-β tubulin are closer to Q. Wang's results than to what was reported by Chinen *et al.*
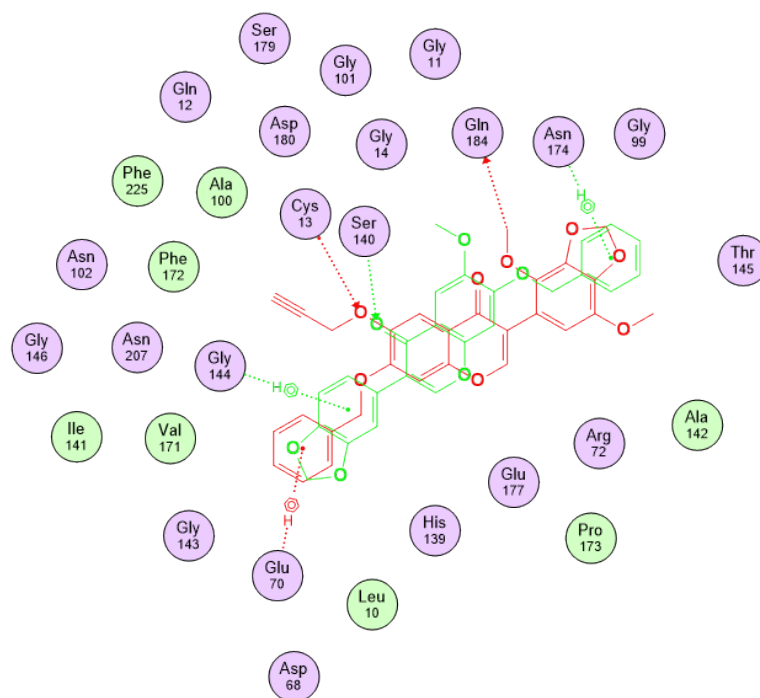
Figure 27: S3-G2 comparison in the GDP binding site.

On the other hand, colchicine docking results on α-β led to much higher predicted binding affinities than those on γ tubulin, similarly to the binding assay results by Chinen *et al.*

As previously stated, no indication of the specificity of the compound for γ tubulin emerged from the comparison of gatastatin binding to γ and α-β. Again, this result is in contrast with what was found by Chinen *et al.*

In addition to this, it must be noted that α-β tubulin expression is constitutive to all cells, while γ makes less than 1% of the total tubulin content of the cell. Even when it is overexpressed, as it was found for glioblastoma multiforme cancer cells, high specificity for γ tubulin would be needed to target it *in vivo*. This is all the more true if one considers that the proposed inhibitory mode of action of gatastatin is to prevent GTP binding to γ tubulin: the most probable way it could succeed in doing so within tolerable doses is by having a higher binding affinity to γ tubulin than GTP itself. As GDP redocking predicted it to have a higher affinity than gatastatin, and γ has a preference to bind to GTP than GDP [31], this situation may be considered unlikely. Figure 29 shows that GDP has a higher number of interaction points with the receptor.
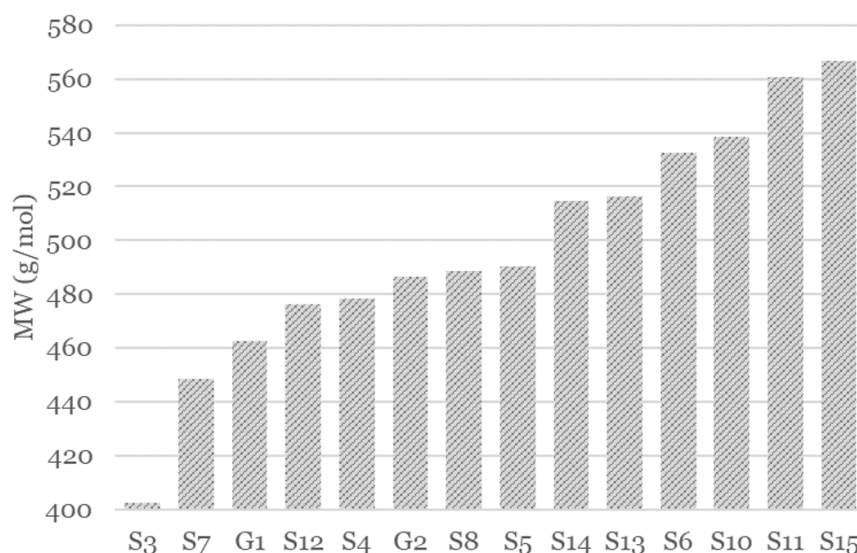
Figure 28: Molecular weights of gatastatin and its derivatives, in ascending order.
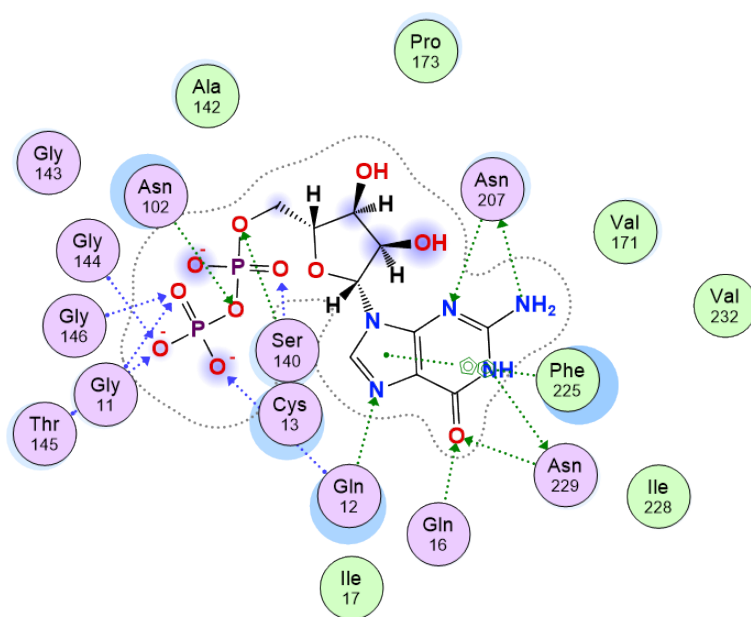

Figure 29: Interactions made by GDP in its native binding site on γ.

No conclusive interpretation can be given based on just computational results, and given their varying degree of agreement with past experiments, further experimental evaluation is needed to assess them.

Table 33: Consensus docking results for α-βI.

Compound	Score (kcal/mol)	Equivalent K_d (μM)
G1	-8.03 ± 0.89	4.16 ± 2.87
G2	-9.22 ± 0.89	0.85 ± 1.02
S3	-8.48 ± 0.20	1.10 ± 0.32
S4	-7.87 ± 1.26	8.77 ± 7.73
S5	-8.48 ± 0.15	1.08 ± 0.26
S6	-	-
S7	-8.60 ± 0.35	0.99 ± 0.47
S8	-8.74 ± 0.79	1.26 ± 1.07
S10	-10.39 ± 0.60	0.07 ± 0.07
S11	-	-
S12	-8.95 ± 0.29	0.54 ± 0.24
S13	-	-
S14	-	-
S15	-9.19 ± 0.65	0.58 ± 0.58

Table 34: Consensus docking results for α-βIIb.

Compound	Score (kcal/mol)	Equivalent K_d (μM)
G1	-8.93 ± 0.58	0.73 ± 0.52
G2	-9.00 ± 0.66	0.68 ± 0.45
S3	-8.89 ± 0.52	0.74 ± 0.51
S4	-7.91 ± 0.43	3.26 ± 1.72
S5	-	-
S6	-	-
S7	-8.68 ± 0.11	0.78 ± 0.13
S8	-9.00 ± 0.64	0.68 ± 0.41
S10	-9.22 ± 0.63	0.47 ± 0.32
S11	-9.04 ± 0.60	0.59 ± 0.34
S12	-9.04 ± 0.26	0.46 ± 0.17
S13	-	-
S14	-	-
S15	-	-

Table 35: Consensus docking results for α-βIVb.

Compound	Score (kcal/mol)	Equivalent K_d (μM)
G1	-9.27 ± 0.38	0.62 ± 0.41
G2	-9.23 ± 0.40	0.38 ± 0.25
S3	-8.85 ± 0.58	0.81 ± 0.51
S4	-7.90 ± 1.00	5.95 ± 4.63
S5	-8.96 ± 0.12	0.49 ± 0.10
S6	-9.99 ± 0.96	0.27 ± 0.33
S7	-8.98 ± 0.40	0.55 ± 0.26
S8	-9.02 ± 0.66	0.69 ± 0.54
S10	-10.37 ± 0.68	0.09 ± 0.10
S11	-	-
S12	-9.03 ± 0.27	0.47 ± 0.17
S13	-	-
S14	-9.66 ± 0.51	0.22 ± 0.19
S15	-9.67 ± 0.90	0.44 ± 0.54

Table 36: Consensus docking results for α-βV.

Compound	Score (kcal/mol)	Equivalent K_d (μM)
G1	-8.25 ± 0.38	1.82 ± 1.02
G2	-9.07 ± 0.23	0.42 ± 0.15
S3	-8.91 ± 0.59	0.78 ± 0.64
S4	-	-
S5	-8.61 ± 0.15	0.87 ± 0.23
S6	-9.69 ± 0.78	0.33 ± 0.38
S7	-9.11 ± 0.38	0.44 ± 0.20
S8	-9.47 ± 0.76	0.42 ± 0.43
S10	-9.68 ± 0.53	0.20 ± 0.11
S11	-9.26 ± 0.62	0.50 ± 0.49
S12	-8.75 ± 0.32	0.76 ± 0.31
S13	-	-
S14	-	-
S15	-9.15 ± 0.68	0.61 ± 0.57

Table 37: Consensus docking results for α-βVI.

Compound	Score (kcal/mol)	Equivalent K_d (μM)
G1	-8.71 ± 0.82	1.33 ± 1.06
G2	-	-
S3	-8.71 ± 0.33	0.82 ± 0.35
S4	-	-
S5	-	-
S6	-8.94 ± 1.27	2.20 ± 2.74
S7	-9.02 ± 0.24	0.47 ± 0.19
S8	-8.66 ± 1.36	2.53 ± 1.83
S10	-9.02 ± 1.16	1.32 ± 1.34
S11	-	-
S12	-8.42 ± 0.13	1.18 ± 0.22
S13	-	-
S14	-	-
S15	-	-

Table 38: Consensus docking results for α-βVIII.

Compound	Score (kcal/mol)	Equivalent K_d (μM)
G1	-8.52 ± 0.51	1.43 ± 1.23
G2	-9.73 ± 0.20	0.14 ± 0.04
S3	-8.81 ± 0.52	0.82 ± 0.50
S4	-7.87 ± 1.02	6.58 ± 5.49
S5	-	-
S6	-8.96 ± 0.99	1.26 ± 1.36
S7	-8.78 ± 0.13	0.66 ± 0.14
S8	-9.32 ± 1.01	0.96 ± 1.23
S10	-10.22 ± 0.41	0.08 ± 0.05
S11	-	-
S12	-8.54 ± 0.32	1.07 ± 0.44
S13	-8.22 ± 0.91	4.33 ± 5.16
S14	-	-
S15	-9.06 ± 1.03	1.55 ± 2.01

Figure 30: Missing residues locations in the γ tubulin PDB structures. Missing residues are shaded in black; the green rectangles frame residues that form the GDP binding site.

```
                    5        10        15              40        45        50
1Z5V          M P R E I I T L Q L G Q C G N Q I — — I V E E F A T E G T D R K D V F F
1Z5W          M P R E I I T L Q L G Q C G N Q I — — I V E E F A T E G T D R K D V F F
3CB2 chain A  M P R E I I T L Q L G Q C G N Q I   I V E E F A T E G T D R K D V F F
3CB2 chain B  M P R E I I T L Q L G Q C G N Q I   I V E E F A T E G T D R K D V F F
6V5V          M P R E I I T L Q L G Q C G N Q I   I V E E F A T E G T D R K D V F F

                   55        60        65        70        90        95
1Z5V          Y Q A D D E H Y I P R A V L L D L E P R — Y N P E N I Y L S E H G G G
1Z5W          Y Q A D D E H Y I P R A V L L D L E P R   Y N P E N I Y L S E H G G G
3CB2 chain A  Y Q A D D E H Y I P R A V L L D L E P R   Y N P E N I Y L S E H G G G
3CB2 chain B  Y Q A D D E H Y I P R A V L L D L E P R   Y N P E N I Y L S E H G G G
6V5V          Y Q A D D E H Y I P R A V L L D L E P R — Y N P E N I Y L S E H G G G

             100     140     145        175        180
1Z5V          A G N N — H S I A G G T G S P N Q D E M S D V V V — D N T Q N P S
1Z5W          A G N N — H S I A G G T G S P N Q D E M S D V V V   D N T Q N P S
3CB2 chain A  A G N N   H S I A G G T G S P N Q D E M S D V V V   D N T Q N P S
3CB2 chain B  A G N N   H S I A G G T G S P N Q D E M S D V V V   D N T Q N P S
6V5V          A G N N   H S I A G G T G S P N Q D E M S D V V V   D N T Q N P S
```

96

```
           225       230                 280       285                 310             335
1Z5V         F S Q I N Q   L T T D Q S V A S V R K T T   S T G R D R Q T N   V H K
1Z5W         F S Q I N Q   L T T D Q S V A S V R K T T   S T G R D R Q T N   V H K
3CB2 chain A F S Q I N Q   L T T D Q S V A S V R K T T   S T G R D R Q T N   V H K
3CB2 chain B F S Q I N Q   L T T D Q S V A S V R K T T   S T G R D R Q T N   V H K
6V5V         F S Q I N Q   L T T D Q S V A S V R K T T   S T G R D R Q T N   V H K

           340       345                 370                 405       410       415
1Z5V         S L Q R I R E R K L   Y L P S A H R   E A F L E Q F R K E D M F K D
1Z5W         S L Q R I R E R K L   Y L P S A H R   E A F L E Q F R K E D M F K D
3CB2 chain A S L Q R I R E R K L   Y L P S A H R   E A F L E Q F R K E D M F K D
3CB2 chain B S L Q R I R E R K L   Y L P S A H R   E A F L E Q F R K E D M F K D
6V5V         S L Q R I R E R K L   Y L P S A H R   E A F L E Q F R K E D M F K D

           440       445       450
1Z5V         R P D Y I S W G T Q E Q
1Z5W         R P D Y I S W G T Q E Q
3CB2 chain A R P D Y I S W G T Q E Q
3CB2 chain B R P D Y I S W G T Q E Q
6V5V         R P D Y I S W G T Q E Q
```

BIBLIOGRAPHY

[1] URL: https://www.rcsb.org/structure/6V6S.

[2] URL: https://www.ncbi.nlm.nih.gov/books/NBK9932/.

[3] URL: https://www.ncbi.nlm.nih.gov/books/NBK21537/.

[4] URL: https://www.proteinatlas.org/search/TUBB.

[5] URL: https://saves.mbi.ucla.edu/.

[6] URL: https://swissmodel.expasy.org/qmean/.

[7] URL: https://www.proteinatlas.org/ENSG00000085563-ABCB1.

[8] URL: https://www.uniprot.org/help/uniprotkb.

[9] URL: https://www.ebi.ac.uk/Tools/msa/clustalo/.

[10] URL: https://pubchem.ncbi.nlm.nih.gov/.

[11] URL: https://github.com/Valdes-Tresanco-MS/AMDock.

[12] URL: https://zhanggroup.org/DockRMSD/.

[13] URL: https://iq.opengenus.org/algorithm-to-find-cliques-of-a-given-size-k/.

[14] URL: http://www.swissadme.ch/index.php.

[15] URL: http://biosig.unimelb.edu.au/pkcsm/prediction.

[16] Jogoth Ali et al. "Revisiting the General Solubility Equation: In Silico Prediction of Aqueous Solubility Incorporating the Effect of Topographical Polar Surface Area". In: *Journal of Chemical Information and Modeling* 52.2 (Jan. 2012), pp. 420–428. DOI: 10.1021/ci200387c. URL: https://doi.org/10.1021/ci200387c.

[17] Eric W. Bell and Yang Zhang. "DockRMSD: an open-source tool for atom mapping and RMSD calculation of symmetric molecules through graph isomorphism". In: *Journal of Cheminformatics* 11.1 (June 2019). DOI: 10.1186/s13321-019-0362-7. URL: https://doi.org/10.1186/s13321-019-0362-7.

[18] Pascal Benkert, Marco Biasini, and Torsten Schwedel. "Toward the estimation of the absolute quality of individual protein structure models". In: *Bioinformatics* (2011). DOI: 10.1093/bioinformatics/btq662.

[19] Pascal Benkert, Silvio Tosatto, and Dietmar Schomburg. "QMEAN: A comprehensive scoring function for model quality assessment". In: *Proteins* 71 (Apr. 2008), pp. 261–77. DOI: 10.1002/prot.21715.

[20] Ruth Brenk et al. "Lessons Learnt from Assembling Screening Libraries for Drug Discovery for Neglected Diseases". In: *ChemMedChem* 3.3 (Mar. 2008), pp. 435–444. DOI: 10.1002/cmdc.200700139. URL: https://doi.org/10.1002/cmdc.200700139.

[21] Takumi Chinen et al. "The γ-tubulin-specific inhibitor gatastatin reveals temporal requirements of microtubule nucleation during the cell cycle". In: *Nature Communications* 6.1 (Oct. 2015). DOI: 10.1038/ncomms9722. URL: https://doi.org/10.1038/ncomms9722.

[22] Chris Colovos and Todd O. Yeates. "Verification of protein structures: Patterns of nonbonded atomic interactions". In: *Protein Science* 2 (1993). DOI: 10.1002/pro.5560020916.

[23] Antoine Daina, Olivier Michielin, and Vincent Zoete. "iLOGP: A Simple, Robust, and Efficient Description of n-Octanol/Water Partition Coefficient for Drug Design Using the GB/SA Approach". In: *Journal of Chemical Information and Modeling* 54.12 (Nov. 2014), pp. 3284–3301. DOI: 10.1021/ci500467k. URL: https://doi.org/10.1021/ci500467k.

[24] Antoine Daina, Olivier Michielin, and Vincent Zoete. "SwissADME: a free web tool to evaluate pharmacokinetics, drug-likeness and medicinal chemistry friendliness of small molecules". In: *Scientific Reports* 7 (2017). DOI: 10.1038/srep42717. URL: https://www.nature.com/articles/srep42717.pdf.

[25] John S. Delaney. "ESOL: Estimating Aqueous Solubility Directly from Molecular Structure". In: *Journal of Chemical Information and Computer Sciences* 44.3 (Mar. 2004), pp. 1000–1005. DOI: 10.1021/ci034243x. URL: https://doi.org/10.1021/ci034243x.

[26] H. Edelsbrunner, D. Kirkpatrick, and R. Seidel. "On the shape of a set of points in the plane". In: 29.4 (July 1983), pp. 551–559. DOI: 10.1109/tit.1983.1056714. URL: https://doi.org/10.1109/tit.1983.1056714.

[27] William J. Egan, Kenneth M. Merz, and John J. Baldwin. "Prediction of Drug Absorption Using Multivariate Statistics". In: *Journal of Medicinal Chemistry* 43.21 (Sept. 2000), pp. 3867–3877. DOI: 10.1021/jm000292e. URL: https://doi.org/10.1021/jm000292e.

[28] David Eisenberg, Roland Lüthy, and James U. Bowie. "VERIFY3D: Assessment of protein models with three-dimensional profiles". In: *Macromolecular Crystallography Part B*. Vol. 277. Methods in Enzymology. Academic Press, 1997, pp. 396–404. DOI: https://doi.org/10.1016/S0076-6879(97)77022-8.

[29] Peter Ertl, Bernhard Rohde, and Paul Selzer. "Fast Calculation of Molecular Polar Surface Area as a Sum of Fragment-Based Contributions and Its Application to the Prediction of Drug Transport Properties". In: *Journal of Medicinal Chemistry* 43.20 (Sept. 2000), pp. 3714–3717. DOI: 10.1021/jm000942e. URL: https://doi.org/10.1021/jm000942e.

[30] Wei P. Feinstein and Michal Brylinski. "Calculating an optimal box size for ligand docking and virtual screening against experimental and predicted binding pockets". In: 7.1 (May 2015). DOI: 10.1186/s13321-015-0067-5. URL: https://doi.org/10.1186/s13321-015-0067-5.

[31] Douglas E. Friesen et al. "Discovery of Small Molecule Inhibitors that Interact with γ-Tubulin". In: *Chemical Biology & Drug Design* 79.5 (Feb. 2012), pp. 639–652. DOI: 10.1111/j.1747-0285.2012.01340.x. URL: https://doi.org/10.1111/j.1747-0285.2012.01340.x.

[32] Arup K. Ghose, Vellarkad N. Viswanadhan, and John J. Wendoloski. "A Knowledge-Based Approach in Designing Combinatorial or Medicinal Chemistry Libraries for Drug Discovery. 1. A Qualitative and Quantitative Characterization of Known Drug Databases". In: *Journal of Combinatorial Chemistry* 1.1 (Dec. 1998), pp. 55–68. DOI: 10.1021/cc9800071. URL: https://doi.org/10.1021/cc9800071.

[33] Linda Gombos et al. "GTP regulates the microtubule nucleation activity of γ-tubulin". In: 15.11 (Oct. 2013), pp. 1317–1327. DOI: 10.1038/ncb2863. URL: https://doi.org/10.1038/ncb2863.

[34] Rob W. W. Hooft et al. "Errors in protein structures". In: *Nature* (1996). DOI: 10.1038/381272a0.

[35] Douglas R. Houston and Malcolm D. Walkinshaw. "Consensus Docking: Improving the Reliability of Docking in a Virtual Screening Context". In: *Journal of Chemical Information and Modeling* 53.2 (Feb. 2013), pp. 384–390. DOI: [10.1021/ci300399w](https://doi.org/10.1021/ci300399w). URL: https://doi.org/10.1021/ci300399w.

[36] Taskeen Iqbal Janjua et al. "Frontiers in the treatment of glioblastoma: Past, present and emerging". In: 171 (Apr. 2021), pp. 108–138. DOI: [10.1016/j.addr.2021.01.012](https://doi.org/10.1016/j.addr.2021.01.012). URL: https://doi.org/10.1016/j.addr.2021.01.012.

[37] Christos D. Katsetos et al. "Altered Cellular Distribution and Subcellular Sorting of γ-Tubulin in Diffuse Astrocytic Gliomas and Human Glioblastoma Cell Lines". In: 65.5 (May 2006), pp. 465–477. DOI: [10.1097/01.jnen.0000229235.20995.6e](https://doi.org/10.1097/01.jnen.0000229235.20995.6e). URL: https://doi.org/10.1097/01.jnen.0000229235.20995.6e.

[38] Christos D. Katsetos et al. "Class III β-Tubulin and γ-Tubulin are Co-expressed and Form Complexes in Human Glioblastoma Cells". In: 32.8 (Apr. 2007), pp. 1387–1398. DOI: [10.1007/s11064-007-9321-1](https://doi.org/10.1007/s11064-007-9321-1). URL: https://doi.org/10.1007/s11064-007-9321-1.

[39] Christos D. Katsetos et al. "Tubulin targets in the pathobiology and therapy of glioblastoma multiforme. II. γ-tubulin". In: 221.3 (July 2009), pp. 514–520. DOI: [10.1002/jcp.21884](https://doi.org/10.1002/jcp.21884). URL: https://doi.org/10.1002/jcp.21884.

[40] Paul Labute. "The generalized Born/volume integral implicit solvent model: Estimation of the free energy of hydration using London dispersion instead of atomic surface area". In: *Journal of Computational Chemistry* 29.10 (Feb. 2008), pp. 1693–1698. DOI: [10.1002/jcc.20933](https://doi.org/10.1002/jcc.20933). URL: https://doi.org/10.1002/jcc.20933.

[41] Roman Laskowski et al. "PROCHECK: A program to check the stereochemical quality of protein structures". In: *Journal of Applied Crystallography* 26 (Apr. 1993), pp. 283–291. DOI: [10.1107/S0021889892009944](https://doi.org/10.1107/S0021889892009944).

[42] Gilbert Lederman et al. "Recurrent Glioblastoma multiforme: Potential Benefits Using Fractionated Stereotactic Radiotherapy and Concurrent Taxol". In: 69.1-4 (1997), pp. 162–174. DOI: [10.1159/000099870](https://doi.org/10.1159/000099870). URL: https://doi.org/10.1159/000099870.

[43] Christopher A Lipinski et al. "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". In: *Advanced Drug Delivery Reviews* 46.1-3 (Mar. 2001), pp. 3–26. DOI: 10.1016/s0169-409x(00)00129-0. URL: https://doi.org/10.1016/s0169-409x(00)00129-0.

[44] Richard F. Ludueña and Asok Banerjee. "The Tubulin Superfamily". In: Humana Press, pp. 177–191. DOI: 10.1007/978-1-59745-336-3_7. URL: https://doi.org/10.1007/978-1-59745-336-3_7.

[45] Fábio Madeira et al. "The EMBL-EBI search and sequence analysis tools APIs in 2019". In: *Nucleic acids research* 47.W1 (July 2019), W636–W641. ISSN: 0305-1048. DOI: 10.1093/nar/gkz268. URL: https://europepmc.org/articles/PMC6602479.

[46] Yvonne C. Martin. "A Bioavailability Score". In: *Journal of Medicinal Chemistry* 48.9 (Apr. 2005), pp. 3164–3170. DOI: 10.1021/jm0492002. URL: https://doi.org/10.1021/jm0492002.

[47] Ikuo MORIGUCHI et al. "Comparison of Reliability of log P Values for Drugs Calculated by Several Methods." In: *Chemical and Pharmaceutical Bulletin* 42.4 (1994), pp. 976–978. DOI: 10.1248/cpb.42.976. URL: https://doi.org/10.1248/cpb.42.976.

[48] Ikuo MORIGUCHI et al. "Simple Method of Calculating Octanol/Water Partition Coefficient." In: *Chemical and Pharmaceutical Bulletin* 40.1 (1992), pp. 127–130. DOI: 10.1248/cpb.40.127. URL: https://doi.org/10.1248/cpb.40.127.

[49] Ingo Muegge, Sarah L. Heald, and David Brittelli. "Simple Selection Criteria for Drug-like Chemical Matter". In: *Journal of Medicinal Chemistry* 44.12 (May 2001), pp. 1841–1846. DOI: 10.1021/jm015507e. URL: https://doi.org/10.1021/jm015507e.

[50] L.Burt Nabors, Birgit Surboeck, and Wolfgang Grisold. "Complications from pharmacotherapy". In: Elsevier, 2016, pp. 235–250. DOI: 10.1016/b978-0-12-802997-8.00014-1. URL: https://doi.org/10.1016/b978-0-12-802997-8.00014-1.

[51] Russell O. Potts and Richard H. Guy. In: *Pharmaceutical Research* 09.5 (1992), pp. 663–669. DOI: 10.1023/a:1015810312465. URL: https://doi.org/10.1023/a:1015810312465.

[52] Kana Shintani et al. "Structure Optimization of Gatastatin for the Development of γ-Tubulin-Specific Inhibitor". In: *ACS Medicinal Chemistry Letters* 11.6 (Mar. 2020), pp. 1125–1129. DOI: 10.1021/acsmedchemlett.9b00526. URL: https://doi.org/10.1021/acsmedchemlett.9b00526.

[53] Shinji Soga et al. "Use of Amino Acid Composition to Predict Ligand-Binding Sites". In: 47.2 (Jan. 2007), pp. 400–406. DOI: 10.1021/ci6002202. URL: https://doi.org/10.1021/ci6002202.

[54] Glen H.J. Stevens and Lizbeth Robles. "Neuromuscular complications". In: Elsevier, 2011, pp. 283–291. DOI: 10.1016/b978-1-4377-1015-1.00027-8. URL: https://doi.org/10.1016/b978-1-4377-1015-1.00027-8.

[55] Simon J. Teague et al. "The Design of Leadlike Combinatorial Libraries". In: *Angewandte Chemie International Edition* 38.24 (Dec. 1999), pp. 3743–3748. DOI: 10.1002/(sici)1521-3773(19991216)38:24⟨3743::aid-anie3743⟩3.0.co;2-u. URL: https://doi.org/10.1002/(sici)1521-3773(19991216)38:24%3C3743::aid-anie3743%3E3.0.co;2-u.

[56] Mario S. Valdés-Tresanco et al. "AMDock: a versatile graphical tool for assisting molecular docking with Autodock Vina and Autodock4". In: 15.1 (Sept. 2020). DOI: 10.1186/s13062-020-00267-2. URL: https://doi.org/10.1186/s13062-020-00267-2.

[57] Daniel F. Veber et al. "Molecular Properties That Influence the Oral Bioavailability of Drug Candidates". In: *Journal of Medicinal Chemistry* 45.12 (May 2002), pp. 2615–2623. DOI: 10.1021/jm020017n. URL: https://doi.org/10.1021/jm020017n.

[58] Anping Wang and Guibin Zhang. "Differential gene expression analysis in glioblastoma cells and normal human brain cells based on GEO database". In: (Sept. 2017). DOI: 10.3892/ol.2017.6922. URL: https://doi.org/10.3892/ol.2017.6922.

[59] Scott A. Wildman and Gordon M. Crippen. "Prediction of Physicochemical Parameters by Atomic Contributions". In: *Journal of Chemical Information and Computer Sciences* 39.5 (Aug. 1999), pp. 868–873. DOI: 10.1021/ci990307l. URL: https://doi.org/10.1021/ci990307l.

[60] Dawnne O'Neal Wise, Ralf Krahe, and Berl R. Oakley. "The γ-Tubulin Gene Family in Humans". In: 67.2 (July 2000), pp. 164–170. DOI: 10.1006/geno.2000.6247. URL: https://doi.org/10.1006/geno.2000.6247.

[61] Akiko Yuba-Kubo et al. "Gene knockout analysis of two γ-tubulin isoforms in mice". In: 282.2 (June 2005), pp. 361–373. DOI: 10.1016/j.ydbio.2005.03.031. URL: https://doi.org/10.1016/j.ydbio.2005.03.031.

[62] Yong Zhao, Daniel Stoffler, and Michel Sanner. "Hierarchical and multi-resolution representation of protein flexibility". In: *Bioinformatics* 22.22 (Sept. 2006), pp. 2768–2774. ISSN: 1367-4803. DOI: 10.1093/bioinformatics/btl481. eprint: https://academic.oup.com/bioinformatics/article-pdf/22/22/2768/499526/btl481.pdf. URL: https://doi.org/10.1093/bioinformatics/btl481.

[63] Haizhen A. Zhong. "ADMET Properties: Overview and Current Topics". In: *Drug Design: Principles and Applications*. Springer Singapore, 2017, pp. 113–133. DOI: 10.1007/978-981-10-5187-6_8. URL: https://doi.org/10.1007/978-981-10-5187-6_8.