Structural study of the therapeutic potential of protein-ligand interactions

Autor: Francisco Martínez Jiménez

TESI DOCTORAL UPF / ANY L'any de la tesi: 2016

DIRECTOR DE LA TESI

Director: Marc A. Marti-Renom Departament Departament: Biomedicine

Por la undecima

Acknowledgments

Abstract

Most of the cellular functions are driven small-molecules that selectively bind to their protein targets. Is such the importance that the pharmacological intervention of proteins by small molecule drugs is frequently used by the pharmaceutical industry to treat multiple conditions. Herein I present a thesis that leverages a three-dimensional study of small molecule protein interactions to improve their therapeutic relevance. More specifically, it introduces nAnnolyze, a method predicting structurally detailed protein-ligand interactions at proteome scale. The method exemplifies its applicability by predicting the human targets of all small molecule FDA-approved drugs. A second application of nAnnolyze in *Mycobacterium tuberculosis* identified the bacterial targets for two sets of compounds with known antitubercular activity. Finally, the thesis describes a computational model that predicts cancer associated mutations with the highest chances to confer resistance to a targeted cancer therapy. Additionally, for those mutations identified as responsible of resistance, the model also provided alternative non-resistant treatments.

Resumen

La mayoría de las funciones celulares están dirigidas por pequeñas moléculas que selectivamente se unen a sus proteínas diana. Es tal su importancia que la intervención farmacológica de proteínas mediante pequeñas moléculas es frecuentemente usada por la industria farmacéutica para tratar múltiples enfermedades. A continuación presento a una tesis que utiliza un estudio tridimensional de las interacciones entre pequeñas moléculas y proteínas para mejorar su relevancia terapéutica. Específicamente, presento nAnnolyze, un método que predice interacciones proteína-ligando estructuralmente detalladas y a nivel de proteoma. El método ejemplifica su aplicabilidad a través de la predicción de dianas terapéuticas humanas para todos las pequeñas moleculas usadas como fármacos aprobados por la FDA. Una segunda aplicación de nAnnolyze en *Mycobacterium tuberculosis* identificó las proteínas diana para dos conjuntos de compuestos con actividad en dicha bacteria. Finalmente, la tesis describe un modelo computacional que predice mutaciones asociadas a cáncer con alta probabilidad de conferir resistencia a una terapia dirigida. Además, para aquellas mutaciones identificadas como responsables de producir resistencia, el modelo también proporciona terapias alternativas predichas como no resistentes.

Preface

I never expected to become a computational biologist. To be honest when I was doing my computer science degree, I didn't even know such a thing existed. I remember the first time I heard the term *bioinformatics*. I was on a seminar, doing my degree's last year, and I though that it sounded like some fancy thing where crazy scientist were working with computers to analyze the DNA. But that was precisely what I always wanted to be: a crazy scientist who wears a lab white coat and writes on white boards.
Months later, after a quick chat with Dr. Luis Vazquez, I decided to enroll into the UCM Bioinformatics master in Madrid. I have to confess that, when the course started, I was completely lost with the biological part. At that point, I though it wasn't so important to know all the details of how biology works, at the end I was a computer scientist, and I always would be. Years later I realized that understanding the biology behind your problem makes the difference.
Over the course I became interested in proteins, and more in particular, in how proteins interact with small molecules. How famous drugs such as Ibuprofen or Viagra, which all of us are familiar with, really work in our body? That was amazing, I loved it! So I decided to do a three months internship at the Marc A. Marti-Renom's lab, working on protein ligand interactions.
I was 23 years old when I started in Marc's lab. At that time, we were four people in the lab: Marc, Davide, David and me. From the first day, I knew that was going to work for me. I liked the work, I liked the people, and I was doing the thing I like the most: learning. I was not only learning biology but also how research works. It was striking, outside of the research community, there is a oversimplification about how research works. There is a huge gap between research and society, a gap that we must bridge...
Soon after, Marc offered me the possibility of doing the PhD at his lab. In spite of there were different project options, I was committed to work on the very same topic: drug-protein interactions.

Four years later, I'm writing this thesis where I describe the work I've done over the course of my PhD. Things have significantly changed, we are more than ten in the lab, I often read biology stuff not related to my work, I become a decent cook, my hair is becoming white and I even speak Catalan (OK, that's not true but at least I try my best on it...). Sadly, I don't wear a lab coat and I don't write

on white boards neither... but I now consider myself a crazy scientist!! I hope you enjoy the scientific part of this story.

Contents

[1]In Subsection 3.1 we present nAnnolyze, a method for protein-ligand interaction prediction. In the introduction of the mentioned manuscript, there is a discussion of the current state-of-the-art methods in protein-ligand interaction prediction. Therefore, this section is focused in explaining the classification, underlying basics, advantages and

disadvantages of the different approaches.

List of Figures

List of Tables

List of publications

The list of publications is presented in reverse chronological order. Publications 1), 3), 4) and 5) compose the main body of the thesis.

1. **Martínez-Jiménez, F.**, Overington J. P., Al-Lazikani B., & Marti-Renom, M. a. (2016). **Rational design of non-resistant targeted cancer therapies**. Genome Medicine. (*in preparation*).

2. **Martínez-Jiménez, F.***, & Marti-Renom, M. a. (2016). **Should network biology be used for drug discovery?**. Expert Opinion on Drug Discovery (*under revision*).

3. Rebollo-Lopez, M. J., Lelièvre, J., Alvarez-Gomez, D., Castro-Pichel, J., **Martínez-Jiménez, F.**, Papadatos, G., ... Barros-Aguire, D. (2015). **Release of 50 new, drug-like compounds and their computational target predictions for open source anti-tubercular drug discovery**. PloS One, 10(12), e0142293. doi:10.1371/ journal.pone.0142293

4. **Martínez-Jiménez, F.**, & Marti-Renom, M. a. (2015). **Ligand-Target Prediction by Structural Network Biology Using nAnnoLyze**. PloS Computational Biology, 11(3), e1004157. doi:10.1371/ journal.pcbi.1004157

5. **Martínez-Jiménez, F.**, Papadatos, G., Yang, L., Wallace, I. M., Kumar, V., Pieper, U., ... Marti-Renom, M. a. (2013). **Target Prediction for an Open Access Set of Compounds Active against Mycobacterium tuberculosis**. PLoS Computational Biology, 9(10), e1003253. doi:10.1371/journal.pcbi.1003253

6. López-Pelegrín, M., Cerdà-Costa, N., **Martínez-Jiménez, F.**, Cintas-Pedrola, A., Canals, A., Peinado, J. R., ... Gomis-Rüth, F. X. (2013). **A novel family of soluble minimal scaffolds provides structural insight into the catalytic domains of integral membrane metallopeptidases**. The Journal of Biological Chemistry, 288(29), 21279–94. doi:10.1074/jbc.M113.476580

Introduction

1.1. Protein are essential molecules

The importance of proteins in biological chemistry is implicit in their name, derived from the Greek word *proteios*, and that means "of the first rank" [2]. Their presence is so essential that they constitute most of the cell dry mass [1]. They are not only the cell's building blocks, but also they perform nearly all the cell's functions. Some roles of proteins include serving as structural components of cells and tissues (e.g., keratin or collagen), transmission of information between cells by hormones such as the insulin or the oxytocin, facilitating the transport and storage of small molecules (e.g., the transport of oxygen by hemoglobin) or providing a defense against foreign invaders (e.g., antibodies). Other proteins such as the actin and the myosin are responsible of muscle contraction and therefore our movement. However, the most fundamental role of proteins is their ability to act as enzymes, which catalyzes most of the chemical reactions in biological systems. In summary, proteins are crucial macromolecules that are present in most of the processes carried out by the cell and, in spite of being extensively studied for many years, they still carry many unanswered questions.

1.1.1. Protein structure

A protein is a molecule made from a long chain of amino acids linked thorough a covalent peptide bond. Proteins are therefore also known as *polypeptides*. Attached to this repetitive chain are those portions of the amino acids that are not involved in the covalent bond, the side chains. Side chains confer the different

[2] The term protein was coined by Jons Jacob Berzelius in 1838. It was first used by Gerardus Johannes Mulder, advised by Berzelius, in its publication *Bulletin des Sciences Physiques et Naturelles en Néerlande (1838). pg 104. SUR LA COMPOSITION DE QUELQUES SUBSTANCES ANIMALES*, where he observed that all proteins seemed to have the same empirical formula and came out to the erroneous idea that they might be composed of a single type of very large molecule. Berzelius proposed the name because the material seemed to be the primitive substance of animal nutrition that plants prepare for herbivores.

physico-chemical properties of each of the 20 types of amino acids [2]. The composition of the amino acid sequence determines the function and the structure of a protein. That is a unique sequence creates a specific pattern of attractive and repulsive forces between amino acids along the polypeptide leading to a folding process that results in a specific three-dimensional (3D) structure. These forces are usually non-covalent interactions between the side chains of the amino acids. Non-covalent interactions are weaker than covalent, allowing the folded structure to certain degree of conformation mobility. This phenomenon is really important to enable the interaction with other molecules as we will explore further in 1.1.6.

Protein structures are complex conformation of atoms organized in a hierarchical manner 1.1. The first level of this hierarchy, referred to as the primary structure, is the ordered sequence of amino acids of the polypeptide. Certain segments of these chains, tend to form simple shapes such as helices, strands, turns or loops. These folding patterns are referred to as secondary elements and collectively constitute the secondary structure of the protein. The two most frequent type of secondary elements are the α-helixes and the β-sheets [3]. The overall chain tends to fold further into a 3D tertiary structure. Contrary to the secondary structure, the tertiary structure folding is driven by interactions from amino acids far apart in the primary sequence. The tertiary structure, is generally the most stable form of the protein, that is, the one that minimizes its free energy [4]. Furthermore, the tertiary structure is also the biologically active form of the protein, and its unfolding usually leads towards partial o total inactivation of the protein. Finally, some proteins are composed by multiple folded chains. In such cases, each folded subunit folds independently and then joins the others forming a biologically active complex. This type of organization is considered as the quaternary structure.

The traditional paradigm of protein structure has been challenged by some exceptions of proteins lacking of a fixed or ordered 3D structure. The intrinsically disordered proteins (IDPs) cover a wide spectrum of states from fully unstructured to partially structured including conformations such as *random coils* or *molten globules*. Moreover, some factors may lead to the permanent loss of structure of a protein, and when that occurs, they endanger the entire organism. How problematic protein misfolding can be for the organism is illustrated by examples such as cystic fibrosis, Alzheimer's, Parkinson's and Huntington's diseases.

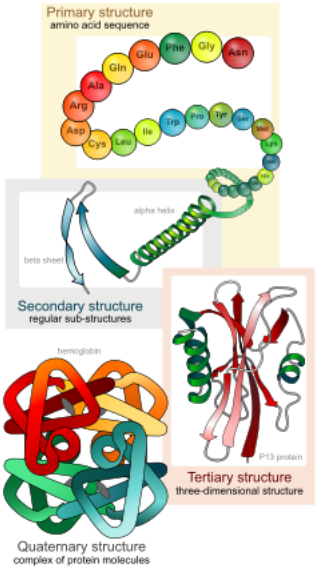


Figure 1.1: Hierarchical distribution of layers in protein structure. Photo Credit: Mariana Ruiz Villarreal (`https://commons.wikimedia.org/wiki/File:Main_protein_structure_levels_en.svg`)

Chothia and Lesk in the 80s [5] helped to set up the fundamentals of what is considered a central paradigm in protein evolution: protein structure is more conserved than sequence 1.2. However, not all the regions in a protein structure are equally conserved. It has been shown that functionally important amino acids, responsible of the interaction with other molecules, are more conserved than the rest of the protein structure [6]. Additionally, the structural core is more conserved than the surface [7]. The high degree of conservation of the protein core enables the protein to maintain the global shape, while the surface is free to change [8]. These evolutionary mechanisms are in accordance with the central *sequence* $\rightarrow$ *structure* $\rightarrow$ *function* paradigm that prevails in the protein evolution field.

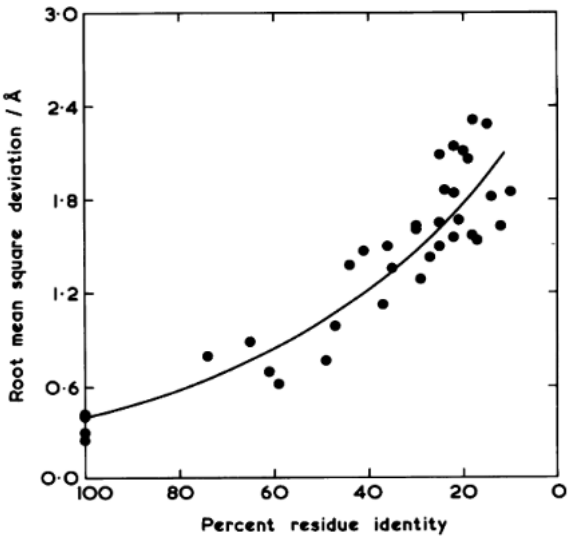


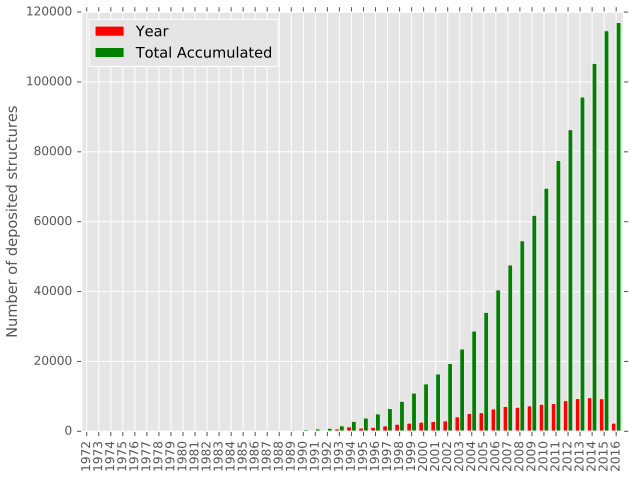
Figure 1.2: The original plot of the relation of residue identity and the RMSD deviation of the backbone atoms of the common cores of 32 pairs of homologous proteins. Figure extracted from [5]

1.1.2. Protein Structure Determination

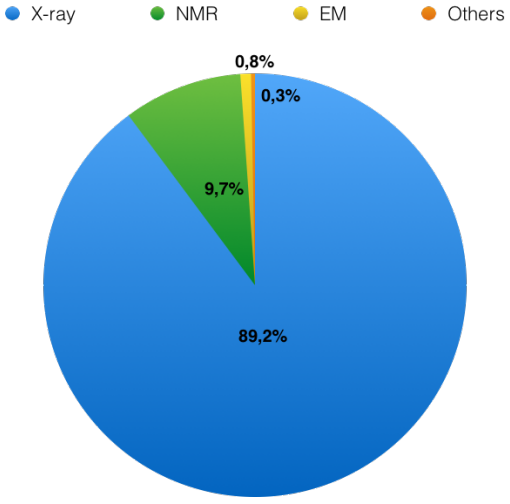
Since 1960, when the Brithis biochemist John Kendrew determined the myoblobin structure [9], more than 37,000 different protein structures have been deposited in the Protein Data Bank (PDB) [10]. The PDB is a repository created in the 1970s with the aim of storing all the 3D protein structures and unifying their format. Figure 1.3a shows the variation of the number of deposited structures over the time. The number of PDB structures has significantly increased over the last years thanks to initiatives such as the Protein Structure Initiative (PSI) [11] or the Structural Genomics Consortium [12]. The later, was born with the aim of determining the structure of all human proteins. However, soon after, they realized that the goal was unrealistic. Fortunately, the number of folds which represent the complete *fold space* observed in nature is much smaller that the number of proteins. Therefore, the current goal is to determine the structure of a representative set of proteins, that is, at least one protein per fold class. Subsequently, using the structure of representative proteins as templates, and thanks to the *homology modeling* techniques (Subsection 1.1.4), it is usually possible to infer the structure of other proteins belonging to the same fold class as we will explore further in the next Subsection 1.1.3.

Several methods are currently used to experimentally determine the 3D structure of a protein. More than 99% of structures deposited in the PDB have been determine by the three main methods: X-ray crystallography, nuclear magnetic resonance spectroscopy (NMR) and electron microscopy (EM) (Figure 1.3b). These methods provide experimental data that helps the scientist to elucidate the final structure of a protein. However, in most cases, the experimental data is not sufficient by itself to build an atomic model. Additional knowledge about the molecular structure most be added. For example, the preferred geometry of atoms in a standard protein, the patterns of repulsion and attraction of amino acids, etc. All this information allows the building of the final model that is consistent with both the experimental data and the prior knowledge of the 3D geometry of the molecules. I next briefly explain the three aforementioned methods:

(a) **X-ray crystallography**. Currently, it is the most widely used method in protein structure determination. Almost 90% of the structures deposited in PDB come from X-ray crystallization (Figure 1.3b). In this method, X-



(a)



(b)

Figure 1.3: a) Growth of released structures per year. Data extracted from PDB. b) Pie chart with the percetange of structures determined by the different methods. Data extracted from PDB.

rays fired at a crystal of the molecule are diffracted by the electron clouds of the protein atoms, forming an unique pattern that is printed as a picture of the atomic density map. Subsequently, the diffraction pattern is combined with other physio-chemical knowledge of the protein, such as composition or atomic geometrical restrictions, in order to build the final 3D model [13].

Before the X-ray exposition, it is then necessary a prior step of crystallization of the molecule. Unfortunately, the crystallization step introduces itself a great number of limitations. The flexibility of proteins is one of the these limitations. The flexible nature of proteins makes really difficult the creation of an accurate and homogeneous alignment of multiple molecules used to create the crystal. Another important limitation is the different conditions required for crystallizing each different molecule. These limitation are especially noteworthy in membrane proteins. Despite of nearly 30% of eukaryotic proteins are membrane proteins, only 604 unique membrane protein structures have been solved to date (data extracted from `http://blanco.biomol.uci.edu/mpstruc/`, March 2016). Therefore, alternative innovative techniques are needed to overcome the numerous obstacles associated with X-ray structure determination of membrane proteins [14].

The accuracy of the final atomic structure relies on the quality of the generated crystals. Two important measures of the accuracy of a crystallographic structure are the *atomic resolution*, which refers to the smallest separation between crystal lattice planes that is resolved in the diffraction pattern [15], and the *R-factor*, which measures how well the refined structure predicts the observed data [16].

(b) **NMR spectroscopy**. The NMR spectroscopy technique has been used for years to determine the structure of proteins. Currently, almost 10% of the structures deposited in PDB have been determined by this method (Figure 1.3b). In NMR spectroscopy, the protein is purified, placed in a strong magnetic field, and eventually probed with radio waves. The observed set of atomic resonances is then analyzed to retrieve a list of atomic nuclei that are close in the space. Similarly to X-ray crystallography, this set of restraints is subsequently used to build the structural model of the protein that contains the 3D conformation of each atom in the space [17].

NMR spectroscopy has a major advantage over X-ray crystallography: it provides information on proteins in solution. Therefore, this method is the main method for studying the atomic structure of highly flexible proteins. A standard NMR structure includes an ensemble of protein structures, all of them being consistent with the experimentally observed set of restraints. The ensemble of structures are very similar in those regions with strong restrains, less constrained regions of the proteins, on the other hand, show less agreement in the generated models. These lack of restriction areas are presumably the flexible regions of the protein since they do not provide a strong signal in the experiment.

A limitation in comparison with X-ray crystallography is its applicability, this technique is usually limited to proteins smaller than 35 kDa. Moreover, NMR can only be applied to soluble proteins that do not aggregate and are stable during the NMR experiment. NMR is also inherently insensitive and milligram amount of proteins are required [17].

(c) **Electron microscopy (EM)** methods. EM methods are emerging as a very versatile tool for determining the structure of large macromolecular complexes. To date, less than 1% of proteins in PDB have been determined by EM methods (Figure 1.3b). However, in recent years there has been dramatic increase in the number of complexes determined by EM technologies. The revolution in the structural biology field is perfectly manifested by the cryo-electron microscopy (cryoEM) method: in 2015 alone, cryoEM was used to map the structures of more than 100 different molecules [18]. In cryoEM a beam of electrons is fired at a frozen protein solution. The emerging scattered electrons pass through a lens to create a magnified image on the detector, and the structure can then be deduced afterwards.
The utility of cryoEM and others EM tools lies on the fact that it allows the observation of molecules that have not been fixed in any way, showing them in their native environment. This is the opposite of the crystallization step in X-ray crystallography, which many times hampers the success of the whole procedure. CryoEM have been traditionally used in large molecules such as ribosomes [19] or the V-ATPase [20], but they have also shown their potential in small membrane proteins [21] and medically relevant proteins [22].
However, there is a big a room for improvement for EM methods. Despite of recent advances in the resolution, most of the cryoEM structures have

lower resolution than the X-ray ones. Furthermore, there are numerous unsolved technical problems that need to be addressed to make easier its standardization and systematic application.

1.1.3. Protein Structure Prediction

Despite of the advances in methods for protein structure determination, most of the known proteins lack of deposited structure in the PDB. There are more than 65 billion protein sequences in UniProtKB (`http://www.uniprot.org`, August 2016), including 551,705 manually annotated and reviewed. However, only about 4% of the later group (*i.e.* 23,195 different protein sequences) have a link to a PDB structure. Therefore, there is a gap between the number of known protein sequences and the number of determined structures, the so-called *sequence-structure gap* [23]. Computational methods for structure determination are helping to bridge this gap. The prediction of the 3D structure of a protein from its amino acid sequence has always been one of the most desirable goals in computational biology. It would save a lot of resources, and it would set a milestone in the structural biology field. Unfortunately, we are still far from being able to predict the structure of many proteins from their primary sequence. Overall, four different approaches are commonly used. The first, and most extensively used, is the *homology* or *comparative modeling*, that uses similar experimentally determined structures to model the structure of the protein of interest Subsection 1.1.4. Second, *fold recognition* and *threading* methods are used to model protein structures with low similarity to known protein structures [24, 25]. Third, *de novo* or *ab initio* methods make their predictions by combining the principles of physics that rule protein folding, with information derived from known structures but without relying in any type of similarity or evolutionary relationship to known folds [26]. Finally, the *integrative* or *hybrid* methods combine different computational and/or experimental sources to perform the structure prediction [27].

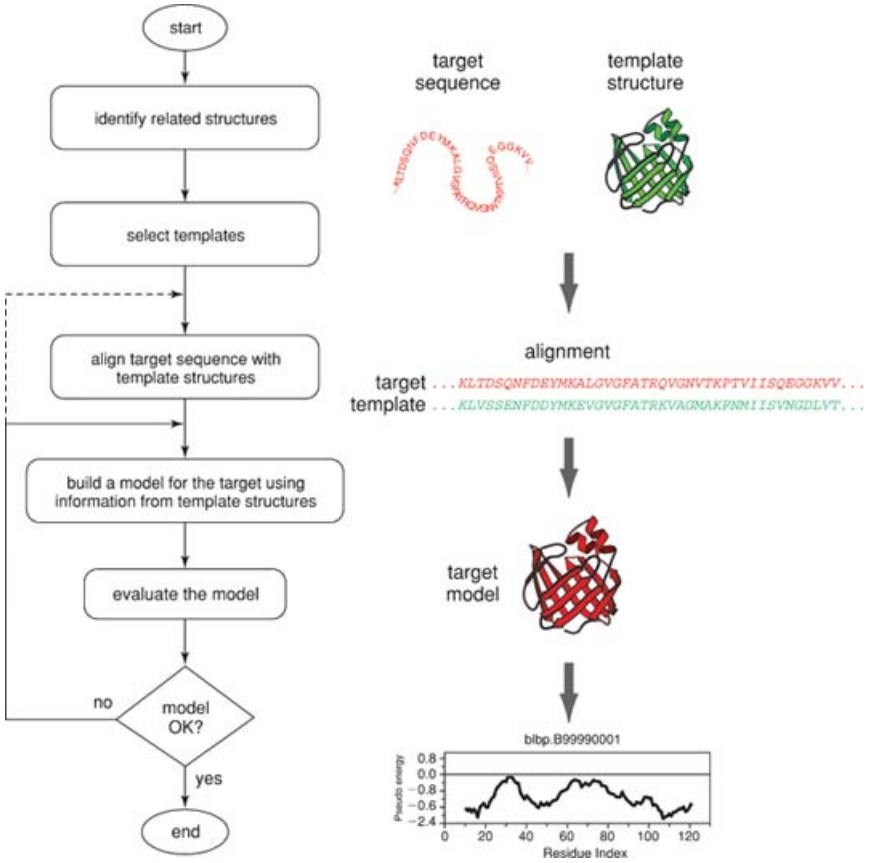


Figure 1.4: Workflow in comparative protein structure modeling. The figure has been extracted from [28]

1.1.4. Homology modeling

This type of protein structure prediction methods exploits the evolutionary relationship between the *target* protein (i.e., the protein being modeled) and the template(s) with known experimental structure. They are based on the biological premise that evolutionary related sequences tend to have similar 3D structures (Subsection 1.1.1 and 1.1.2). Figure 1.4 shows the regular steps in comparative protein structure modeling:

1. **Identification of suitable template(s) structure(s) similar to the target protein**. This step consist on a search for similar sequences with known 3D structure. This task is easy when the 3D structure of a close homologue of the target protein has been experimentally determined. Initiatives such as the PSI [11] are helping to this issue by increasing the number of modellable proteins. However, there are still many proteins with lack of homologous proteins in PDB. In these cases, alternative methods such as *ab initio* modeling might be used.

2. **Alignment between the target and the template(s) sequence(s)**. The sequence identity of the target-template alignment is the most frequently used measure for similarity. Consequently, the sequence identity is also a good predictor of the final 3D model quality. The overall accuracy of models calculated from alignments with sequence identity higher than 40% is usually good (i.e., RMSD [3] lower than 2.0Å). In the 30%-40% identity range, errors can be more severe and are often locate in loops and highly flexible regions. Below the 30% of sequence similarity, often referred to as *twilight region*, serious errors occurs including the basic fold being mispredicted [29, 30]. Figure 1.5 shows an empirical threshold for homology modeling extracted from [31]. The region below the curve gathers those cases where the alignment does not carry enough information to model the 3D structure, while area above the threshold curve, include those cases where homology modeling is applicable.

[3] Root Mean Square Deviation is the measure of the average distance between the atoms of two superimposed proteins. Equation $RMSD = \sqrt{\frac{1}{N}\sum_{i=1}^{N}\delta_i^2}$ where δ_i is the distance between the N_i pair of equivalent atoms (usually the Cα).

3. **Modeling and refinement of the structurally conserved regions and prediction of the structurally variable regions.** There are different algorithms to assign the spatial coordinates of the target protein using the template(s)-target alignment information. Highly conserved regions are generally well modeled, while those regions with insertions or gaps are more prone to include errors and suboptimal atomic orientations. Next, the model is refined to idealize bond geometry and to remove errors that may have been introduced in the modeling step. The refinement process pursues the free energy minimization of the generated 3D protein model. Many different algorithms have been applied to perform the minimization step, including molecular mechanics force fields [32], molecular dynamics [33], Monte Carlo simulations [34] or genetic algorithms [35].

4. **Evaluation of the model(s).** Model evaluation seeks for the identification of the different errors that might have occurred during the modeling process. Multiple methods have been developed to assess the quality of a 3D model. In fact, 3D model assessment was included from the seventh edition of the CASP experiment [36]. The general question of how accurate is a model can be reformulated in several specific questions:

   I *Is the selected fold correct?* The fold assessment consist of deciding whether a given protein model has the right fold. Residue-based combined accessible surface and distance-dependent scoring functions have shown the best performance in this task [37].

   II *How do we select the best model among the set of decoys or alternative solutions?* Several models can be generated by making changes in the template-target alignment, by selecting different template(s) structure(s) or by using different seeds in the refinement non-deterministic algorithms. Atom-based distance-depend scoring functions have proved to be useful for this particular task in some cases [38]. However, there is not a gold standard for ranking the generated 3D models. Thus, the model selection eventually relies on the expertise of the person running the experiment.

   III *Which is the overall accuracy of the model? Which is the accuracy of the model in a particular region of the model?* These questions can be addressed by defining scores that correlate with the RMSD after superimposing a model and its native structure. Numerous

scoring functions have been implemented to address this issue. The physics-based scoring functions attempt to approximately calculate the atomic interaction energies in the system. These scoring functions usually encode a set of parameters that describes the energy of a system of particles. Examples of these scoring functions are AMBER [39], CHARMM [40] or MM-PDBSA [41]. Differently, the knowledge-based potentials or potentials of mean force, are scoring functions derived from an analysis of empirical information. The physical meaning of potentials of mean force has been widely disputed since their introduction [42]. Nevertheless, since they frequently correlate with the actual free energy differences, they have been broadly used with significant success [43, 44, 45].

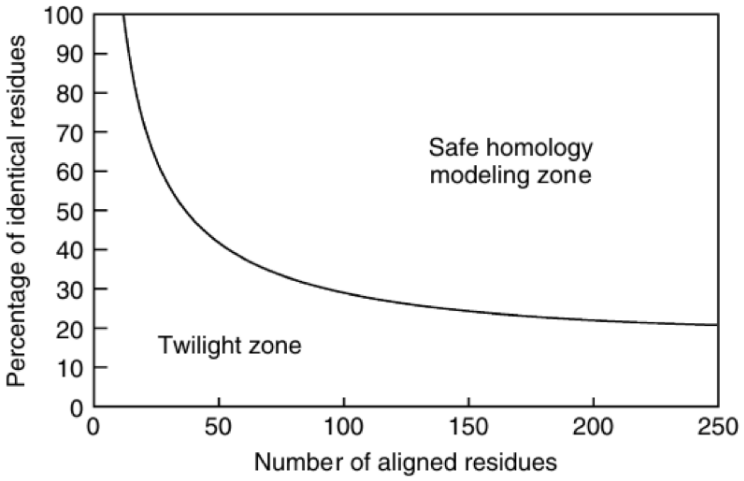


Figure 1.5: Homology threshold curve as a function of alignment length. Data extracted from [31]

The application of comparative modeling is limited by several aspects. First, is the availability of a suitable template. Despite of the efforts made to determine at least one structure per known fold [11], divergences between the template and the target hampers the modeling of a correct 3D structure. In fact, models based on alignments with sequence identity below 30% may be unsatisfactory

(Figure 1.5). The lack of template problem is even more noticeable in membrane proteins. The limited number of membrane proteins with 3D experimentally determined structure available makes its modeling an extremely difficult task. However, the high value of these proteins in diverse therapeutic areas [46, 47] is fostering the development of specific membrane protein modeling methods [48]. Another aspect that restricts the success of homology modeling is the innate flexibility of proteins. Highly flexible regions are more difficult to model and consequently are more prone to errors than more rigid parts of the structure. Despite of these limitations, homology modeling has been successfully applied to many proteins and its currently the main approach to computationally model the 3D structure of proteins[4].

There are many computational methods to predict the 3D structure of proteins (Table 1.1). Each of these methods have their own strengths and weakness and none of them clearly outperforms the others for all cases.

Modeling Tool	**Website**	**Reference(s)**
Modeller	`https://salilab.org/modeller/`	[28, 51]
SwissModel	`http://swissmodel.expasy.org`	[52]
HHPred	`http://toolkit.tuebingen.mpg.de/hhpred`	[53]
I-Tasser	`http://zhanglab.ccmb.med.umich.edu/I-TASSER/`	[54, 55, 56]
Rosetta	`http://robetta.bakerlab.org/`	[57]
RaptorX	`http://raptorx.uchicago.edu/`	[58]
3DJIGSAW	`http://bmm.crick.ac.uk/~3djigsaw`	[59]
WhatIf	`http://swift.cmbi.ru.nl/whatif/`	[60]

Table 1.1: Examples of public protein modeling tools alongside their website and original references.

1.1.5. Protein function

One of the main questions in the protein biology field is to understand the protein sequence-structure-function relationship. It is known the structure of a protein

[4]For a comprehensive review of homology modeling methods, applications and limitations please consider [49, 50]

determines its biological function. However, different regions of the structure can perform semi-independent functions from each other. These regions are referred to as protein domains. A domain is a substructure produced by any part of the polypeptide chain, which folds independently into a compact and stable structure [61, 62, 63]. Domains on average contain 80-250 residues [64]. Estimates of the number of domains per protein say that more than 70% of procaryotik proteins and 80% of eukaryotic proteins include more than one domain [65, 66]. Among this multi-domain proteins, 95% of them contains only two to five protein domains [65]. Domains are not only the basic functional units of proteins but also their evolutionary units. As proteins have evolved, domains have been modified and combined to build new proteins [67, 68]. Such is the importance of domains in protein evolution, that they have been included in current protein classification methods as one of the major classification parameters. Some of these domain classification methods such as SCOP [69] or CATH [70] are purely based on the structure, while others such as Pfam [71] or INTERPRO [72] include information about the function in their classification.

Domains, and consequently proteins, perform its biological activity by interacting with other molecules. Proteins can interact with other proteins, constructing a protein-protein complex; with ions or with small-molecules. The substance that is bound to the target protein is called the ligand, while the region of the protein where the ligand is binding is called ligand's *binding site* [5]. The next section is focused on protein-compound interaction presenting the basis for all the work developed during the thesis.

1.1.6. Protein-Ligand Interactions

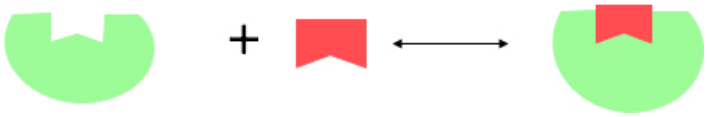
The roles played by the protein ligands are diverse. Catalysis of enzyme substrates, regulation of the protein activity, cellular communication or defense from external attackers are just few examples of the multiple functions that small-molecule ligands perform in living organisms. All these functions are performed by small-molecules that selectively bind to their target proteins. However, given

[5]For simplicity, in this thesis, unless otherwise indicated, the term ligand will only refer to small molecules ligands, while proteins ligands will be explicitly named as protein-protein interactions

the vast amount of proteins and small molecule ligands in the cytoplasm, how do the small molecule ligands select their protein targets? There have been several protein-ligand binding theories. In the *Lock and Key theory* [73], Emil Fischer proposed a system where the binding sites of enzymes are rigid and pre-adjusted geometrically to the natural substrate Figure 1.6a. This theory became widely accepted for years. Nevertheless, during subsequent years, evidence started to accumulate suggesting that the binding sites of proteins do not match perfectly their ligands, but rather the binding process triggers some conformational changes in the enzyme. Therefore the obsolete Lock and Key model was replaced by the *Induced fit theory* [74]. The induced fit theory proposes that initially enzymes do not perfectly match their substrate geometrically. Rather, the binding process triggers a set of conformational changes in the protein binding site that improves the match Figure 1.6b. More recently, another theory called the *Monod-Wyman-Changeux model or MWC model* came up [75]. This theory contends that proteins are able to shift spontaneously between multiple conformations called *substates* [76, 77]. This model could also explain *allostery*, a phenomenon in which the binding of the molecule to the catalytic site is affected the binding of other ligand to a different site. This theory has undergone some changes and the current accepted theory posits that ligands bind preferentially to one of the conformation sampled spontaneously by the protein, and therefore stabilizes it. It means that, by changing the protein's energy landscape, ligands change a less favorable conformation into the most favorable one. This model does not necessarily refute the induce fit theory since in many cases, the restrains applied by the ligand on the binding site is expected to induce some conformational changes that will further stabilize the interaction [78, 79].

1.1.7. Protein-ligand binding energetics

The high variety of functions that ligands perform by binding to proteins is also reflected in the diversity of their binding affinity. Binding energies usually range from -2.5kcal/mol to -22kcal/mol [80]. The binding strength displayed by proteins matches the biological goal of the binding. For instance, ligands involved in protein communication tend to bind weakly to enable a quick state switch. Cofactors binding to enzymes, on the other hand, tend to bind strongly to their targets. The negative sign of the values reflects that is a favorable binding that



(a) Fischer’s *Lock and Key model*. The protein is represented in green and the ligand in red. The ligand’s binding site of the protein matches the ligand perfecltly.

(b) Koshland’s *induced fit model*. The protein is represented in green and the ligand in red. The overall shape of the ligand matches the binding site. The ligand bindings causes some conformational changes that improves the interaction.

(c) *MWC model’s* representation. The protein changes its conformation constantly (one color per conformation), with at least one these conformation matching the ligand. Its binding triggers some conformational changes that improves the protein’s energy landscape.

Figure 1.6: Schematic representation of the three classic protein-ligand binding theories.

releases energy: *the binding free energy*. This energy can be measured experimentally, thorough the equilibrium constant of the binding, or it can be calculated computationally. Formula 1.1 and 1.2 shows, under thermodynamic equilibrium conditions, the relationship between the Gibbs free energy or binding affinity and the equilibrium constant of the binding. R represents the ideal gas constant, T is the temperature, $[C]$ the complex concentration and $[P]$ and $[L]$ the protein and ligand concentration respectively.

$$\triangle G = -RTlnK_{bind} \tag{1.1}$$

$$Kbind = \frac{[C]}{[P] * [L]} \tag{1.2}$$

These equations show that the binding free energy can be measured experimentally. However, in many cases the experimental measurement are unfeasible or very difficult due to technical problems. Additionally, the expenses associated with these experiments often restricts its broader application. In such cases, computational methods to calculate the free binding energy are needed. The calculation of binding free energy have acquired a remarkable importance in the drug discovery field where the calculation of ligand-target affinity is crucial for preclinical phases (Subsection 1.2). Unfortunately, calculation of the ligand-target binding affinity is a extremely challenging task. The main points that should be addressed to accurately calculate the binding free energy are:

1. **The free energy of binding (Formula 1.1) is the difference of two large energies**. The energy of the complex (E_{pl}) and the energy of the unbound partners ($E_p + E_l$) (Formula 1.3):

$$\triangle G_bind = E_pl - (E_p + E_l) \tag{1.3}$$

2. **There are two opposite and complex energies driving the process**. The binding enthalpy ($\triangle H_{bind}$) and the loss of entropy of both the ligand and the protein ($\triangle S_{bind}$):

$$\triangle G_bind = \triangle H_{bind} - T \triangle S_{bind} \tag{1.4}$$

3. **Non explicit representation of the energetic interactions of the system**. Small molecule binding events on a protein cavity implies the displacement of solvent molecules (i.e., usually water molecules). The energy generated by the this exclusion of water molecules is the main driving force in ligand-protein binding [81]. Unfortunately, explicit representation and simulation of all the forces involved this event is computationally very expensive. A popular approach to model is to use implicit solvent force fields [82, 83, 84], where the water molecules are represented as a continuous medium instead of individual explicit molecules. The implicit solvation model is justified in liquids, where the potential of mean force are applied to approximate the behavior of many highly dynamic solvent molecules. However, there could be other medias with specific solvation or dielectric properties that are continuous, but not necessarily uniform, since their properties can be described by different analytic functions [85]. Among the most famous implicit models we can find those based on the Poisson-Boltzmann theory (PB) [86] and those based on the Generalized-Born (GB) approximation [87].

   Hydrogen bonds and salt bridges between the ligand and the protein can also be a source of free energy gain upon ligand binding. This energy gain comes from the displacement of water molecules bound to the protein. The net gain of energy upon hydrogen bond is around 1-2 kcal/mol. Some scoring functions treat all hydrogen bonds equally, while others, distinguish between neutral and charged ones. Other energies that could be modeled and that contribute to the binding affinity calculations are those generated by interactions with metal ions [88]. However, because there may be a covalent component in this type of interactions, its overall binding energy contribution is difficult to model. Finally, nonspecific van der Waals and hydrophobic interactions are also included in some methods as additional energy contributors to the overall free energy of binding [89].

One of the main applications of binding free energy calculation is predicting whether a ligand is binding a particular protein target. In other words, given the predicted binding free energy determine whether a specific compound targets a specific protein binding site. In the next section we will explore further these and other approaches aiming at protein-compound interaction prediction.

1.1.8. Protein-ligand prediction[6]

The importance of ligand-protein interaction prediction is reflected by the large number of available methods that use multiple different approaches [90, 91]. We can distinguish between *free structure* methods (i.e., methods that do not rely on the protein structure to perform its predictions) and *structure-based* methods.
Free structure methods do not require the protein structure to perform their predictions. They usually use prior knowledge on protein compound interactions, to further extend the interactions to new and unseen compounds. The development of these methods can be split in two phases. The first step consist on the creation of a predictive model that uses collection(s) of protein-compound interactions to learn hidden relationships between compound and their protein targets. In the second step, these predictive models are used to extrapolate this knowledge to new and unseen compounds (or targets). The extrapolation relies on different measures of compound or protein similarity. Knowledge-based free structure methods have been assisted by the emergence of new *high-thorughpout screening methods* (HTS) that enabled the creation of large computational compound-protein databases such as ChEMBL [92], Therapeutic Target Database (TTD) [93], Binding MOAD [94], BindingDB [95], PubChem [96, 97] or ZINC among others [98]. The recent growth of these collections is accordingly improving their accuracy and coverage. Moreover, since they do not rely on protein structure they can be theoretically applied to any protein or to any compound. Nevertheless, free structure methods do not provide detailed information about the ligand-protein relationship. Information such as binding localization, type of interaction (e.g., allosteric, on-target or off-target) or predicted free energy of binding; that is absolutely essential in the drug discovery process (Figure 1.7). Consequently, free structure methods are mostly employed in early stages of the drug discovery pipeline.

Structure based target prediction methods leverage protein's 3D structure to determine whether a small-molecule interact with a protein target. Docking meth-

[6]In Subsection 3.1 we present nAnnolyze, a method for protein-ligand interaction prediction. In the introduction of the mentioned manuscript, there is a discussion of the current state-of-the-art methods in protein-ligand interaction prediction. Therefore, this section is focused in explaining the classification, underlying basics, advantages and disadvantages of the different approaches.

ods have traditionally dominated the structure-based target prediction field. Virtual docking consist on predicting the preferred orientation of one molecule (i.e., the ligand) to a second (i.e., the protein). The process of finding the best orientation of molecule, the so-called binding *pose*, to the protein target is not simple, since several entropic, enthalpic and environmental factors have an impact on the interactions between them (Subsection 1.1.7). The underlying idea of this approach is to generate a comprehensive set of ligand-protein conformations, and then to rank them accordingly to a specific scoring function [99]. The importance of virtual docking methods is not only reflected by the large number of published methods, which include AutoDock [100, 101], DOCK [102], FLEXX [103], GOLD [104] or GLIDE [88], among others; but also by their success in drug discovery applications [105, 106, 107] [7]. However, virtual docking methods also have some limitations. The most apparent one is that they rely on protein structure. As mentioned above (Subsection 1.1.3), the coverage of the human structural proteome is below 40%. Thus, some of the most interesting targets in drug discovery lack of experimentally determined 3D structure. In addition to these structurally inherent problems and despite of some massive applications [109], virtual docking methods are still computationally very expensive (Figure 1.7). Additionally, they need the prior knowledge of the binding localization on the protein surface, which many times is unknown before the screening.

1.1.9. Comparative docking approach

To overcome the computational limitations of virtual docking approaches, some structure-based methods use the so-called *comparative docking* approach, which solely relies on structural comparisons, both of compounds and protein targets, to infer new interactions. Comparative docking methods are based on the biological premise that structurally conserved proteins tend to conserve the biological function [110, 111, 112, 113]. In other words, structurally similar protein binding sites tend to bind similar ligands. Unlike virtual docking methods, comparative docking approaches do not require the computationally expensive calculations needed to obtain the structural orientation (i.e., the binding pose) of the compound at the protein binding site. Rather, they provide a more sim-

[7] For a comprehensive review of virtual docking methods an applications please consider [108]

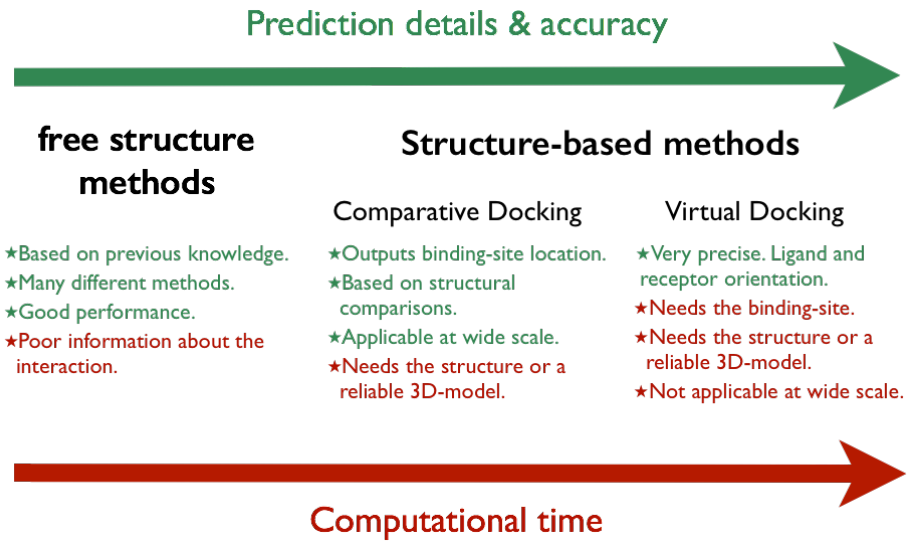


Figure 1.7: Classification of the methods for ligand-target interaction prediction alongside their advantages and disadvantages. The red arrow represent the increase in the computational time of the calculus needed for each prediction. The green arrow represents the level of detail of the given output.

plistic representation, where the output is usually limited to the binding location on the protein surface, omitting information about the exact binding orientation. Consequently, comparative docking approaches are generally faster and more suitable for large scale virtual screenings than virtual docking methods (Figure 1.7). Several ligand-target interaction prediction methods leverage comparative docking approaches to perform their predictions [114, 115, 116, 117, 110]. Subsection 3.1 presents nAnnolyze, a network-based version of Annolyze [110], which is focused on predicting ligand-target interactions at proteome scale. The nAnnolyze chapter further discusses the applications, advantages, disadvantages and limitations of comparative docking approaches in general, and nAnnolyze, in particular.

1.2. Drug discovery

Drug discovery is the process by which potential new medications are discovered. It involves a wide range of scientific disciplines, including biology, chemistry, pharmacology and recently also the computational branches of these fields. Historically, drugs were discovered through the identification of the active ingredient from traditional remedies or by serendipitous discovery. Later, the development of synthetic methods allowed the generation of purely synthetic structures that were not found in nature and that were investigated as potential therapeutic agents. More recently, the advent of new genomics, proteomics and HTS techniques, resulted in the identification of large number of novel targets for future drug discovery research. In addition to this *technological revolution*, the advances in bioinformatics and system biology field has prompted the change in drug discovery paradigm towards a more target-centric approach. This modern drug discovery paradigm usually implies the screening of thousands of molecules to identify those that have the desirable therapeutic effect in the previously validated protein target [118, 119]. Figure 1.8 shows the current drug discovery pipeline alongside the estimated cost and time of each of the phases. Most modern drug discovery programs begin with the identification of a bio-molecular target which pharmacological intervention is theoretically beneficial for the treating disease. A target is a broad term that can be applied to a range of biological entities including proteins, DNA and RNA. The target needs to be accessible to the putative drug molecule(s), this property is referred to as *target druggabil-*

ity. Wrong selection of the target (i.e., weak association between the target and the treating disease) implies lack of the expected efficacy, which is the most important cause of project failure in clinical trials [120, 121]. During the *hit-identification* stage, the target is screened against a set of candidate molecules seeking for the identification of those which able to perform the desired therapeutic activity. Alternatively, in some cases the first step of the discovery process is based on a *Phenotypic screening* of a collection of molecules. This screening pursues the identification of those molecules that perform a predefined function in a biological model. In any case, prior knowledge of the bio-molecular target of the therapeutic activity is generally associated with better outcomes in clinical trials [122]. However, there are various drugs in the market with unknown *mechanism of action* (i.e., the drug target remains unknown) [123], most of them coming from the traditional drug discovery paradigm. After hit(s)[8] identification, at the *hit-to-lead* stage, molecular hits are evaluated and undergo limited optimization to identify promising lead compounds for further stages. The optimization to convert a hit to a lead molecule, implies several properties, including the potency, the selectivity and the pharmacokinetics (PK) properties. These lead compounds undergo more extensive optimization in a subsequent step of drug discovery called lead optimization (LO). The main goal of this stage is to maintain favorable properties of lead compound(s) while improving on deficiencies in the lead structure(s). Finally, the selected lead(s) enters into the preclinical stage where the main goals are to determine the safe dose for *First-in-man study* and the first assessment of the product's safety profile. Estimates say that, on average, of every 5,000 to 10,000 compounds that begins the pre-clinical stage, only one becomes an approved drug [124].

According to the The Tufts Center for the Study of Drug Development (`http://csdd.tufts.edu`), the development and marketing approval for a New Molecular Entity (NME) takes more than 13 years and around $2.6 billion (Figure 1.8). In fact, the cost of developing a new drug has dramatically increased since the 1970s (Figure 1.9). Currently, the cost of developing a NME is more than two times the 1990s cost, and more than ten times of the cost of the 1970s. The raise in the drug development cost has led to a dramatic shrinkage of the ef-

[8] A hit compound could have several definitions. Here we use the one from [122] where they defined a hit as being a compound which *has the desired activity in a compound screen and whose activity is confirmed upon retesting.*

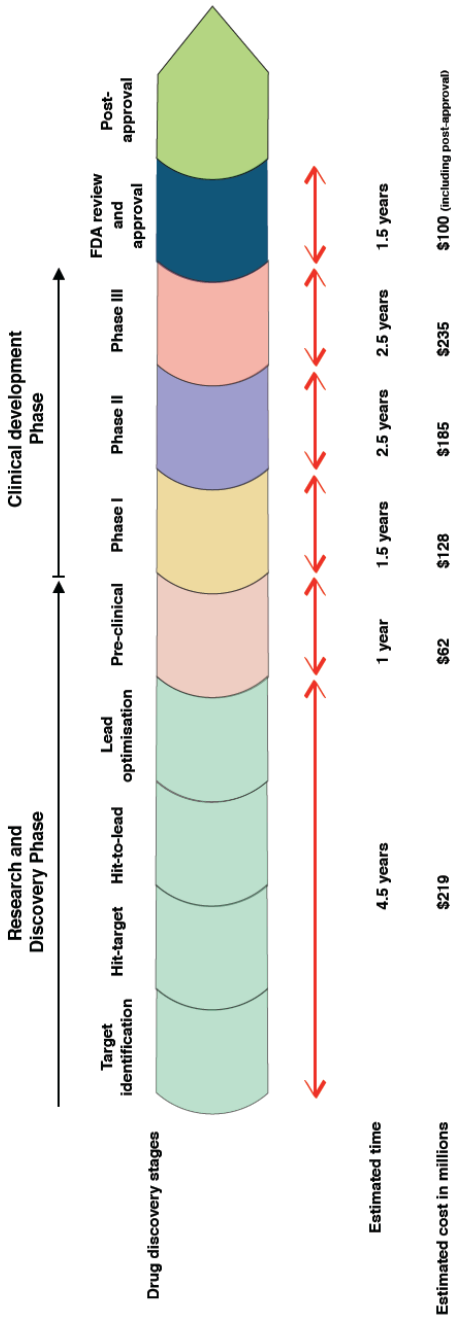


Figure 1.8: Drug discovery and development pipeline. For each stage the average cost and time are provided. Post-approval times not included in the time-line. Data extracted from [125].

ficiency, measured in terms of the number of new approved drugs per billion US dollars of research and discovery (R&D) spending [126]. Both research and development phases have significantly raised their expenses (Figure 1.9). Factors that have contributed to the raise of clinical costs include increased clinical trial complexity, larger clinical trial size, greater assessment of safety and toxicity drug profiles or evaluation on equivalent drugs to accommodate payer demands for comparative effectiveness data [126]. Similarly, factors such as the complexity of the target disease, expenses associated with the application of high-throughput technologies or the complexity of mechanism of action are increasing the prizes of pre-clinical stages. However, pre-clinical associated expenses may be narrowed down with a rational use of the state-of-the-art technologies. In this matter, computational methods are emerging as a tool to speed-up the process by enabling the management of the massive amount of data generated during the discovery stages. Next section introduces different computational methods currently applied during the drug discovery pipeline.

1.2.1. Computational drug discovery

Over the last thirthy years, computer-aided drug discovery (CADD) methods, have played a key role in the development of therapeutic drugs [127]. The modern drug discovery pipeline includes multiple CADD approaches that assist during the drug discovery process:

1. **Target identification and validation methods.** Many different computational approaches are used to identify and validate new targets. The *genomics revolution* caused by New-Generation Sequencing methods (NGS) have significantly increased the development of methods that primary rely on the genetic association between targets and the treating diseases. In some cases, the data is combined with additional information enabling a more precise evaluation of the target viability. Examples of the complementary data include structural data, such as experimental structure availability or druggability assessment; system-biology information such as protein-protein interactions, protein pathway analysis or sub-celullar target localization [128]. Recently, the inclusion of pharmacological data by *drug reporpousing or repositioning* methods became very popular [129,

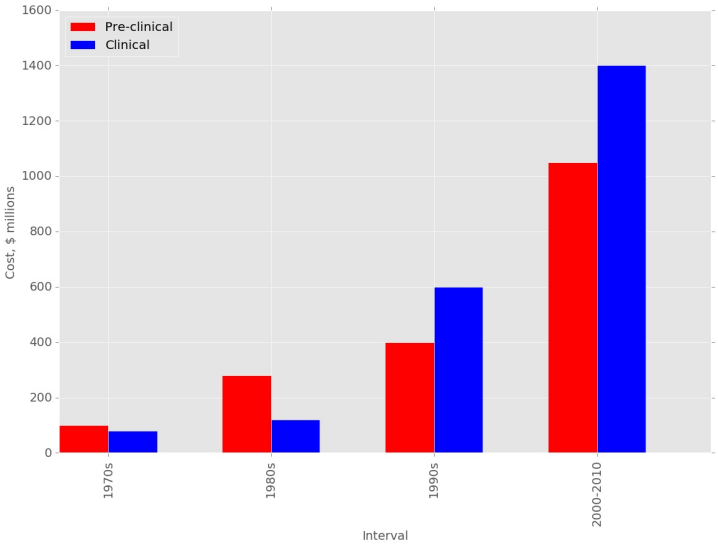


Figure 1.9: Cost of developing a new drug. Blue bars indicate expenses in clinical phases while red represents expenses in pre-clinical stages. Costs are shown in $ millions. Data extracted form: Tufts Center for the Study of Drug Development (`http://csdd.tufts.edu/news/complete_story/pr_tufts_csdd_2014_cost_study`).

130, 131]. These methods leverage information of whether the protein is targeted by any FDA approved drug, to prioritize those targets with annotated FDA approved drug(s). Such drug(s) are subsequently applied to the treating disease to validate the target testing hypothesis. Computational methods for target identification and validation have applied in great variety of diseases, including infectious diseases such as Tuberculosis [132] or Malaria [133], cancer [134] and neurogeneretive diseases [135].

2. **Ligand-target prediction**. Once the target has been validated, CADD methods can help in the search of potential target hits. This is one of the fields where CADD methods have been more successful either by making the predictions from scratch or in combination with phenotypic screenings [136]. Section 1.1.8 specifies the different methods and their current applications.

3. **Quantitative structure-activity relationship (QSAR)**. QSAR is an approach designed to find relationships between chemical structure and the biological activity of small molecules. QSAR methods are based on the assumption that variations in the biological activity of a series of chemicals targeting a particular protein are correlated with variations in their structural, physical, and chemical properties [137]. QSAR methods have become an essential tool in the pharmaceutical industry where they play a major role in the hit-to-lead and lead optimization stages. Traditionally, these methods have been used to improve compounds bioactivity. Recently, the applications have been extended to the improvement of ADMET (adsorption, distribution, metabolism, elimination, toxicity) properties [138, 139] and the oral bio-availability [140]. QSAR methods have undergone rapid changes over the last years. The first 2D-QSAR models were based on descriptors derived from a two-dimensional graph representation of a molecule. These descriptors tried to characterize the most important molecular properties for the molecular interaction. However 2D-QSAR had important limitations for designing new molecules due to the lack of consideration of the 3D structure. Later, 3D-QSAR methods integrated 3D properties of the ligands to predict their biological activity [141]. The first QSAR model that integrated the 3D geometry to perform the predictions was the Comparative Molecular Field Analysis (CoMFA) [142]. In CoMFA, steric and electrostatic features of protein target are

mapped onto a surface grid, which envelops a set of compounds superimposed in their active conformation. This grid acts as a surrogate of the binding site of the protein receptor and is frequently referred to as *pharmacophore*. However, this approach has an important limitation: a ligand molecule can only be represented by a single entity. Therefore, if a ligand binds with different conformations, only one of them can be represented in a 3D-QSAR model [141]. This limitation was overcome by 4D-QSAR methods, which include conformational flexibility and the freedom of alignment by ensemble averaging in the conventional three dimensional descriptors found in 3D-QSAR methods [143]. 4D-QSAR models have been succesfully applied to simulate binding to cytochrome P450 3A4 [144], HIV-1 protease [145] or to the p38-mitogen-activated protein kinase (p38-MAPK), [146] among others [147]. More recently, a 5D-QSAR model has been proposed [148]. This model includes a new degree of freedom, the fifth dimension, that allows for a multiple representation of the atomic topology of the receptor surrogate (i.e., representation of different induced-fit models of the receptor). Finally, in the 6D-QSAR methods, a greater representation of the different solvation scenarios is included [149]. This enables for an even more realistic simulation of the binding process, which is ultimately reflected in the development of better predictive models.

4. **Prediction and optimization of the ADMET properties**. Most of the drug discovery initiatives include a computational optimization of the compound's PK properties. As previously mentioned, QSAR methods have been extensively applied to predict the PK properties of compounds. However, there are other *in-silico* approaches that play a substantial role in the ADMET prediction field. One of the tools that have significantly contributed to the field is the *Lipinski's rule-of-five*, which aims to predict the odds of a compound to become a drug, the so-called *drug-likeness* [150]. The Lipinski's rule-of-five is a rule of thumb created by Christopher A. Lipinski based on the observation of chemical properties of drugs with favorable PK profile. It uses five arbitrary rules based of such number of chemical features to determine whether a compounds is likely to become a drug. If the compound fulfill, at least, four rules then it is considered as a drug-like candidate. However, assessment of compounds drug-likeness in absolute terms does not reflect adequately the whole range of compound

qualities. To address this issue, a computational method that quantitatively measures the drug-likeness of a compound has been recently published [151]. Optimization in the ADMET properties of a compound is generally performed during the hit-to-lead and lead-optimization stages, concurrently with the optimization of the compound's bio-activity. This multi-objective optimization process is accomplished in the computational model developed by Besnard and colleagues [152].

1.3. Drug discovery in Tuberculosis

About one-third of the world's population is infected with *Mycobacterium tuberculosis* (MTB), the causative agent of tuberculosis (TB) [153]. Approximately 90% of infected individuals have latent MTB infections, which remain dormant until activated by specific environmental and host response events. Remarkably, people with compromised immune systems, such as people with HIV, malnutrition or diabetes, or people who use tobacco, have a much higher risk of falling ill. Once the disease has been activated, when left untreated, kills more than half of the infected patients [154]. Despite of TB is considered as a treatable and curable disease, it remains as a top infectious disease killer worldwide. TB is usually treated with a standard 6 month course of combination of 4 antimicrobial drugs. Globally, the treatment success rate for people newly diagnosed with TB was 86% in 2013 [153]. Unfortunately, there is a increasing clinical occurrence of Multidrug-resistant tuberculosis (MDR-TB), which is a form of TB caused by bacteria that do not respond to first-line anti-TB drugs. Some infected patients develop extensively drug-resistant TB (XDR-TB), which is a form of MDR-TB tuberculosis that do no respond to any standard treatment, including the most effective second-line anti-TB drugs [155]. About 480,000 people developed MDR-TB in the world in 2014, while it is estimated that about 9.7% of MDR-TB cases had XDR-TB [153].

Infectious diseases in general, and TB in particular, are suffering from the lack of new innovative therapies [156]. The discovery and development of new antibiotics is widely recognized as one of the major global health emergencies. Most of the currently used antibiotics were discovered in the period from the 1930s to the 1960s [157]. Recently, a new class of antibiotics has been discovered [158].

However, estimations say that it could take more than five years until it is available in the market. The lack of innovation in the antibiotics field has caused the re-emergence of diseases such as TB, dengue, and *African trypanosomiasis.* These diseases predominantly affect poor populations in less developed countries [156]. Concretely, the highest TB incidence rates are found predominantly in low-income countries including most countries in central and southern Africa, southern Asia and some countries from central America (Figure 1.10). The high incident rates of TB in developing countries reflects the urgent need for new and affordable medicines for the treatment of TB, among other infectious diseases. This need has not been directly reflected in traditional R&D programs of the pharmaceutical industry, mainly because they do not offer sufficient financial returns for the pharmaceutical industry to engage in research and development. This fact has led to the development of alternative mechanism to fight against TB and others infectious diseases:

1. **Fostering research and development by philanthropic donations.** Charitable organizations, often private and corporate philanthropic foundations, donate money to drug research and development projects. In some cases, this money is assigned to public institutes to deeply investigate in the mechanism of bacterial infection and resistance. Such is the case of the $20 million project given to the Broad Institute in the fight against tuberculosis [159]. Other projects such as those funded by the Bill & Melinda Gates Foundation (`www.gatesfoundation.org`) seek for the development of less expensive and more effective diagnostic tools. These tools could reach higher TB target population and can be used at the point of care rather than requiring processing by a distant lab. Philanthropy is one of the major responsible of the important decrease in the TB mortality: the TB death rate dropped 47% between 1990 and 2015 [153].

2. **Nonprofit initiatives by big pharmaceutical companies.** Some pharmaceutical companies provide medicines and funds for medicines for developing countries or towards R&D for diseases that affect those countries. In some cases, the companies create specific institutes dedicated to the research and development of new medications against infectious diseases. Examples of this type of institutes include the Novartis Institute for Tropical Diseases (NITD) in Singapore, which focuses on dengue fever and

TB, or the Tres Cantos Open Lab Foundation in Madrid, which is an independent, not-for-profit foundation established by GlaxoSmithKine in 2010 focused on TB, Malaria and kinetoplastid infections. Unlike other type of projects, open-pharma initiatives have usually a very collaborative willingness, which many times results a with very fruitful partnerships between academia and the pharmaceutical institutes. An example of this type of collaborations is presented in chapter 3.2.

3. **Public-private Partnerships (PPP).** The PPP Knowledge Lab (`https://pppknowledgelab.org/`) defines a PPP as a *long-term contract between a private party and a government entity, for providing a public asset or service, in which the private party bears significant risk and management responsibility, and remuneration is linked to performance*. Therefore, in a PPP, a private entity, which develops a public service, ultimately assumes a substantial financial, technical and operational risk in the project. The advantages of these type of approaches resides in their ability to bring the private sector expertise into the delivery of certain of some services traditionally developed by the public sector. Moreover, a PPP is structured in such way that the public entity does not incur any borrowing. Rather, the PPP borrowing is incurred by the private sector implementing the project. Interestingly, PPPs have been applied to cope with TB epidemic worldwide [160, 161]. Overall, in-deep analysis of the outcome produced by PPPs in TB suggests that PPP may generally improve outcomes of a TB service. Specifically, the improvement is reflected throughout a earlier detection, better treatment administration, and broader service accessibility, especially in resource-limited areas [162]. The main beneficiary from this approach seems to be the final patient, who pays less for care while maintaining, or in some cases improving, the quality of treatment.

These strategies are essentially created to bridge the financing gap in Tuberculosis R&D. Next section focuses on specific methodologies and tools applied to perform research in this disease. Particular attention will be given to computer aided strategies applied to the TB research and discovery field.

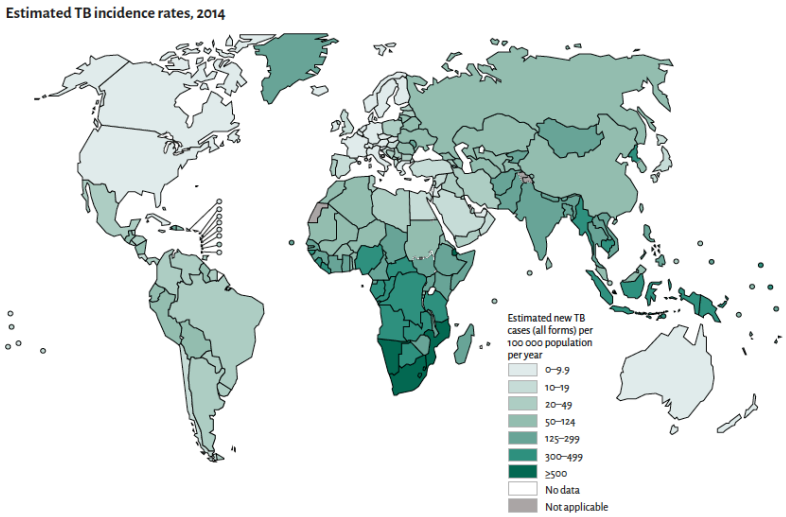


Figure 1.10: Estimated worldwide TB incidence rates in 2014. Figure extracted from [153].

1.3.1. Research strategies against MTB.

Beyond the funding problems of research against MTB, there are numerous technical challenges in identifying new antitubercular compounds [163]. One the main difficulties is the extremely slow growth rate of *Mycobacterium tuberculosis* as this is ultimately reflected in the rate of progress of discovery research. Another aspect is associated with the nature of the bacteria; MTB is a respiratory pathogen, and therefore has to be handled under strict safety conditions (Bio-safety Level 3) requiring expensive specialist facilities. Moreover, MTB have a very unusual cell wall that impedes many compounds from penetrating into the cell [164]. Additionally, this bacteria has efflux pumps that transport compounds out of the cell and that have been implicated in resistance to antibiotics [165]. To make things worse, anti-tubercular drugs need to be safe for periods over 6 months, or even longer periods when dealing with MDR-TB or XDR-TB, without significant side effects or drug-drug interactions.

The search for new anti-tubercular compounds is therefore a extremely chal-

lenging task. Researchers are employing many different approaches in parallel including HTS and computational methods. HTS aims to find new molecular entities that may lead to the development of new antibacterial treatments. One of these HTS approaches is the *cell based phenotypic screening*, which represents a powerful approach to identify anti-bacterial compounds and elucidate novel targets [166]. Some phenotypic screenings are also combined with toxicity assays to find those compounds with high anti-tubercular activity and positive PK profile [167]. Other HTS approaches aim at identifying highly potent molecules against an essential MTB target [168, 169]. Computational methods are essential in the analysis of the vast amount of data generated by HTS providing a very powerful tool to identity those candidate molecules amenable to be optimized in future stages.

1.3.2. In-silico approaches in TB

Similarly to many other diseases, CADD methods play a substantial role in the Tuberculosis R&D field. Uncountable *in-silico* methods have been published over the last decade, each of them applying different strategies to solve a specific biomedical question. However, all of them pursue the very same goal: fueling drug discovery against TB. Table 1.2 contains some remarkable *in-silico* resources for the fight against TB. The purpose of these resources is very diverse. One popular resource is targetTB, which consist on an open-source pipeline to identify targets in MTB [170]. Similarly, chapter 3.2 presents how the combination of three orthogonal approaches can help to identify the molecular targets of novel anti-tubercular compounds. Other resources, such as TDRtargets [171] (`http://tdrtargets.org/`) and CCD-TB [172], blend a target detection tool with a publicly accessible database of known existing targets and anti-microbial compounds. Some methods, on the other hand, are focused on providing insights into a specific problem in TB treatment. Such is the case of the computational detection of Comtan as a potential agent in the treatment of MDR-TB and XDR-TB [173], or the examples from [174] and [175]. Most of these resources take advantage of Tuberculist [176], a database of experimentally measured gene essentiality in MTB; and TuberQ [177], which contains information about MTB proteins druggability. Is such the importance of computational resources in TB research that recently a Mobile app, called TB-Mobile [178, 179],

was published. TB-Mobile provides an agile way to interact with TB data and it includes some chemoinformatics tools for clustering and finding new molecular targets to known anti-tubercular compounds. This app is therefore pushing the boundaries of science on mobile devices in several important ways, and could set up a milestone in bringing mobile apps into the computational biology research field.

Overall, *in-silico* methods play an important role in the research against tropical infectious diseases. Particularly, TB benefited enormously from the contribution of such methods and therefore they are partly responsible of the improvement in the prognosis of the disease.

Type of method	**Name**	**Resource description**	**Reference(s)**
Target identification pipeline	TargetTB	Target prioritization in TB thorough a computational pipeline	[132]
Database	TDRtargets	Database and method for identification of potential MTB targets	[171]
Application of bioinformatics tools	-	Drug repositioning applied to MDR-TB and XDR-TB	[173]
Database	CDD-TB	Database of anti-tubercular compounds reported from HTS alongside computational models to analyze the data	[172]

Application of bioinformatics and chemoinformatics tools	-	Identification of the MTB targets of bio-active antitubercular compounds using three orthogonal *in-silico* approaches	[136, 180]
Application of bioinformatics tools	-	Homology modelling and virtual doking applied to ligand-protein interaction prediction	[174]
Application of chemoinfomatics tools	TB Mobile	Mobile app that provides a platform to interact with data collected from CDD-TB	[178, 179]
Application of bioinformatics and chemoinformatics tools	-	Identification of Enoyl acyl carrier protein reductase binders using a 3D-QSAR approach	[175]
Application of bioinformatics tools	-	Interactome computational analysis to identify potential mechanisms of drug resistance to TB therapies	[132]
Database	Tuberculist	Database of experimentally measured gene essentiality	[176]
Database	TuberQ	MTB protein druggability database	[177]

Table 1.2: Table containing multiple computational resources used in the discovery and research against TB

1.4. Targeted cancer therapy

Cancer is one the leading causes of morbidity and mortality worldwide. In 2012 there were more than 14 million new cases and 8.2 million cancer related deaths. Moreover, the cancer global burden is expected to rise by about 70% over the next 20 years [181]. Intravenous cytotoxic chemotherapy has traditionally prevailed as the main therapeutic choice in cancer treatment. Chemotherapy drugs target rapidly dividing cells, including cancer cells and certain normal tissues. Hence, the lack of specificity of the chemotherapy treatment leads to strong side side effects such as hair loss, gastrointestinal symptoms, fatigue or myelosuppression, among others. In the past decade, however, the arrival of targeted cancer therapies have dramatically transformed cancer treatment. Targeted cancer therapies are drugs designed to specifically interfere with molecules necessary for tumorigenesis. The higher specificity associated to these drugs makes them a more powerful and less harming alternative for cancer treatment. Although chemotherapy remains the treatment of choice for many malignancies, targeted therapies are now a essential component of treatment for many types of cancer, including breast, colorectal, non-small cell lung cancer (NSCLC), as well as lymphoma, several classes of leukemia, and multiple myeloma. There are two main types of targeted cancer therapies, monoclonal antibodies and small molecule inhibitors.

1.4.1. Monoclonal antibodies

Monoclonal antibody-based therapy for cancer has become established over the past 15 years. Monoclonal antibodies are target specific, which means that they exclusively target only one protein. Moreover, their protein target has to be extra cellular, as the antibodies cannot enter the cell through the plasma membrane. Monoclonal antibodies can kill tumour cells throughout multiple mechanism of action [182]. One of the classic mechanism consist on direct action of the antibody on the target protein. An example of this class is the monoclonal antibody cetuximab, an epidermal growth factor receptor (EGFR) inhibitor used in EGFR-positive colorectal cancer [183] and squamous cell carcinoma of the head and neck (SCCHN) [184]. Another mechanism consist on the activation of the immune system response to kill cancer cells. Immunotherapies are be-

coming increasingly popular and its currently one of the most promising fields of cancer research. Examples of this class include the immune checkpoint inhibitors pembrolizumab (PD-1), atezolizumab (PDL-1) and ipilimumab (CTLA-4) [185]; or the CD52 antibody alemtuzumab [186].Tumour vascularization and stroma have also been targeted by antibody-based therapies. For example, bevacizumab is a monoclonal antibody that blocks angiogenesis by targeting the vascular endothelial growth factor receptor (VEGFR) [187]. It is currently used as a single agent or in combination with chemotherapy to treat certain types of advanced cancer, including colorectal, NSCLC, glioblastoma or kidney cancer [188]. Finally, several conjugated antibodies have been approved to treat cancer. An example of this class is ibritumomab tiuxetan, a yttrium-90-conjugated monoclonal antibody to CD20, for patients with relapsed B-cell non-Hodgkin's lymphomas. This drug combines the monoclonal antibody ibritumomab in conjunction with the chelator tiuxetan, to which radioactive isotope is added [189]. Undoubtedly, antibody-based cancer therapies have significantly contributed to the improvement of cancer survival. However, these therapies have still important limitations which prevents them for broader application. One of the major limitations is the temporally efficacy of some treatments. Patients with malignant tumours may not achieve a long-term therapeutic effect consequence of the multiple tumour escape mechanisms [182]. Deeper understanding of the tumor biology may provide insight into selection of patients who are suited to a specific antibody treatment. In summary, monoclonal antibodies has shown a great potential in the treatment of cancer. However, there are important limitations that need to be addressed to increase the clinical impact of this type of treatment.

1.4.2. Small molecule kinase inhibitors

Small molecule inhibitors is the second main class of targeted cancer therapy. Unlike monoclonal antibodies, they can penetrate into the cell through the plasma membrane. Small molecule targeted cancer therapies focus on inhibiting protein kinases. In fact, kinases have been established as promising drug targets for the treatment of various types of human disease because of their essential roles in signal transductions and regulation of a range of cellular activities. However, the vast majority of these targets are being investigated for the treatment of cancer [190]. Over the last years, many kinases have been found to be deeply

involved in the processes leading to tumorigenesis. Depending of their role in cancer progression we can classify small molecule kinase targets into different groups. First, there are kinases that have become insensitive to normal regulatory mechanisms. The altered activity of such kinases can be the consequence of genetic alterations (e.g., mutations or translocations) or epigenetic changes (e.g., gene amplification, increased expression) and are considered to be oncogenic. The constitutive activity of this class of kinase target makes them essential for survival and/or proliferation of the cancer cell. This phenomenon is known as oncogene addiction [191], and makes the cancer cell exceptionally susceptible to the oncogene kinase inhibitor. One of best examples of this phenomenon is the activating V600E BRAF mutation. About 50% of melanomas harbour this oncogenic mutation [192]. Currently, there are two small molecules FDA approved inhibitors that specifically target the BRAF V600E-mutated metastatic melanoma, vemurafenib [193] and dabrafenib [194]. Inhibiting mutationally activated kinases (i.e., oncogenic kinases) has resulted in the most dramatic clinical responses [190]. A second class of target kinases is composed by those non-oncogenic kinases whose presence is preferentially required for the survival and/or proliferation of tumour cells. These kinases are usually located in key signalling pathways downstream of cancer oncogenes. Examples of this type of targets include MEK1 and MEK2 (also known as MAP2K1 and MAP2K2), which are targeted by several small molecule inhibitors such as trametinib or cobimetinib. Combinations of these inhibitors with oncogene inhibitors led into a significant improvement in patient survival compared with single treatment regime in melanoma [195, 196]. Another class of kinases targets are those highly expressed in the tumour stroma and that are required for different stages of tumour formation and development in the human host. Examples of this class include the inhibition of VEGFR by pazopanib or by other small molecule inhibitors [197].

Protein kinases are defined by their ability to catalyse the transfer of the terminal phosphate of ATP to a substrate that usually contains a serine, threonine or tyrosine. They share a highly conserved arrangement of secondary structure elements that fold into a bi-lobed catalytic core structure (N-terminal lobe and C-terminal lobe), with ATP binding site located in a deep cleft located between the two lobes [198] (Figure 1.11). The ATP adenine ring forms hydrogen bonds with the kinase hinge region (*i.e.*, the segment that connects the amino and carboxy terminal kinase domains), while the ribose and triphosphate groups of ATP

bind in a hydrophilic channel adjacent to the ATP binding site that contains conserved residues essential to catalysis. Additionally, kinases have a conserved activation loop, which regulates the kinase activity and that contains a extremely conserved DFG motif (*i.e.*, aspartic acid, phenylalanine and glycine) at the start of the loop. The structural disposition of the activation loop switches between the active and inactive conformations of the protein kinase [198]. Since the catalytic mechanism requires the exact positioning of highly conserved active site residues, the kinase active state is rigid and highly conserved. In contrast, kinase inactive states are structurally highly diverse and dynamic [199]. Furthermore, the kinase ATP binding site contains a highly flexible phosphate-binding loop (P-loop). In many kinases the the P-loop contains an aromatic residue that points upward in the active kinase state, enabling the binding of ATP. Finally, kinases contain a key residue in the ATP-binding site known as *gatekeeper*. This residue is located close to the hinge region and controls the access of small molecule inhibitors to a hydrophobic pocket in the active site that is not contacted by ATP [200] (Figure 1.11).

Most of the current small molecule kinase inhibitors are ATP-competitive that mimics the ATP binding mode. However, depending of their specific binding mode small molecule protein kinase inhibitors can be classified into multiple classes:

1. **Type I inhibitors**. This type of ATP-competitive inhibitor binds the active conformation of the protein kinase. As mentioned above, the kinase active state is well defined and it is more rigid than inactive kinase states. Moreover, is very conserved among kinases making the development of selective type I inhibitors a very challenging task. The specificity is therefore given by unusual active site features such as rare amino acids in conserved positions, inserts/deletions, and, in some cases, residues that can be targeted by irreversible inhibitors. Additionally, small gatekeeper residues, such as threonine, can provide access to a hydrophobic *back pocket* not contacted by ATP [201]. Type 1 inhibitors typically consist of a heterocyclic core scaffold that occupies the adenine binding site alongside side chains that occupy the adjacent hydrophobic regions (Figure 1.12). Examples of this class include the EGFR inhibitors gefitinib and erlotinib, the BRAF V600E-mutant inhibitor vemurafenib, the anaplastic lymphoma ki-

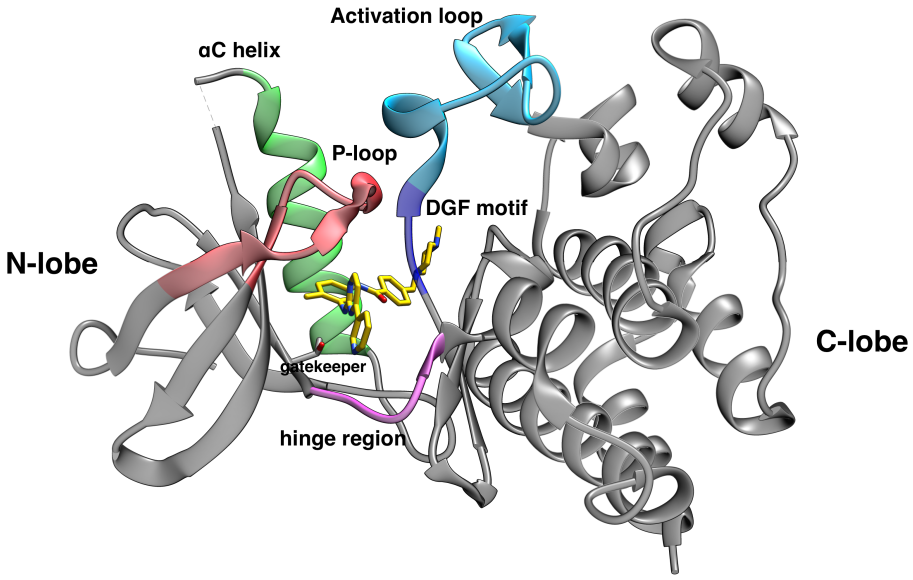


Figure 1.11: 3D structure of ABL1 kinase in complex with imatinib displaying the different structural regions of protein kinases. The structure represents the typical kinase inactive DGF-out conformation. The protein is represented as ribbons and the ligand as sticks. The activation loop is coloured in cyan, the DGF motif in blue, the P-loop is coloured in red, the hinge region in purple and the αC helix in green. PDB accession code 2HYY.

nase (ALK) inhibitor crizotinib or the Bcr-Abl tyrosine kinase inhibitor dasatinib. The complete list of type I inhibitors is shown in Table 1.3.

2. **Type II inhibitors**. Type 2 kinase inhibitors recognize the inactive conformation of the kinase. The most frequent conformation recognized by type 2 inhibitors is the so-called DFG-out. This conformation is created by a rearrangement of the activation loop that creates an extended and flexible binding pocket adjacent to the ATP binding site (Figure 1.12). The high degree of flexibility generated by this conformation suggests that inhibitors targeting such states should have a better chance of being selective. However, recent comprehensive analysis of type II selectivity revealed that many kinases can assume this inactive state and that type II inhibitors may not be intrinsically more selective than type I inhibitors [202]. The original discovery that inhibitors such as imatinib and sorafenib bind in the type 2 conformation was serendipitous, but subsequent analysis of multiple type 2 kinase inhibitor revealed that most of them share a similar binding pattern [202]. Other examples of type II kinase inhibitors include the BCR-ABL kinase inhibitors nilotinib or ponatinib. The complete list of FDA approved type II inhibitors is shown in Table 1.3.

3. **Targeting P-loop conformations**. In kinase-inhibitor complexes, the P-loop may fold into the ATP-binding site, forming aromatic stacking interactions with the inhibitor [203]. An additional characteristic of folded P-loop conformations is the induction of a large binding cavity between the P-loop and the αC helix 1.12. This binding cavity is present in many structures with folded P-loops and has been explored, for the first time, by the selective ERK1/2 inhibitor SCH772984 [204]. Multiple kinases can adopt a folded P-loop conformation, which unique geometric features of this binding mode may lead into the development of selective inhibitors for these kinases. Nevertheless, none of FDA approved drugs adopt this conformation, and therefore a broader general demonstration of this inhibitor binding mode is still necessary.

4. **Type III allosteric inhibitors**. Type III kinase inhibitors are non ATP-competitive inhibitors binding the kinase in a allosteric site (*i.e.*, a site distinct from the enzyme active site that can bind a ligand) and modulating kinase activity in an allosteric manner. Allosteric inhibitors tend to exhibit

the highest degree of selectivity since they exploit binding sites and regulatory mechanisms that are unique to each particular kinase (Figure 1.12). Most allosteric kinase inhibitors have been discovered serendipitously, and currently there is no general strategy for identifying such compounds. The best examples of this class are the MEK1/MEK2 allosteric inhibitors trametinib and cobimetinib, which occupy a pocket adjacent to the ATP binding site.

5. **Covalent inhibitors**. The last class of kinase inhibitors are those capable of forming an irreversible, covalent bond to the kinase active site, most frequently thorough the reaction with a nucleophilic cysteine residue [205]. Most of the covalent kinase inhibitors have been developed by structure-guided incorporation of an electrophilic group into an inhibitor that already had sub-micromolar binding affinity [206]. Although a large number of kinases have cysteine residues in and around the ATP-binding site, there are not conserved cysteines residues across the human kinome [207]. This lack of conservation has been used to develop selective irreversible inhibitors of kinases habouring cysteine residues in the ATP-binding site. However, cross-reactivity of cysteine-reactive groups can lead to non-selective reactions with off-target proteins, which eventually gives rise to increased toxicity and lack of target specificity [208, 209]. Examples of FDA approved irreversible inhibitors include the Bruton's tyrosine kinase inhibitor (BTK) Ibrutinib or the EGFR inhibitor afatinib.

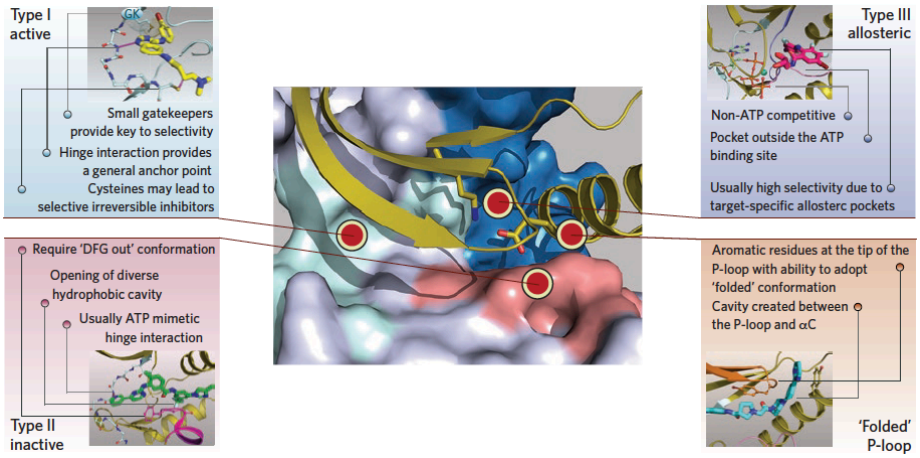


Figure 1.12: Structural features of the canonical classes of small molecule kinase inhibitors. The center panel shows the main interaction sites of different inhibitors types. The side panels show the specific structural features of each of the binding modes. Figure extracted from [199].

Compound name	Pharmacological Target	Binding mode	First FDA approval
Imatinib	ABL1	Type II	2001
Gefitinib	EGFR	Type I	2003
Erlotinib	EGFR	Type I	2005
Sorafenib	VEGFR, PDGFR, BRAF, etc.	Type II	2005
Dasatinib	ABL1	Type I	2006
Sunitinib	VEGFR	Type I	2006
Nilotinib	ABL1	Type II	2007
Lapatinib	EGFR, HER2	Type I and II	2007
Pazopanib	VEGFR, PDGFR, c-KIT, etc.	Type I and II	2009
Crizotinib	ALK, ROS1	Type I	2011
Vemurafenib	BRAF	Type I	2011
Ruxolitinib	JAK1/2	Type I	2011
Vandetanib	VEGFR	Type I	2011
Bosutinib	BCR-ABL1	Type I	2012
Tofacitinib	JAK3	Type I	2012
Axitinib	VEGFR, PDGFR, c-KIT	Type I	2012
Cabozantinib	c-MET	Type II	2012
Regorafenib	VEGFR, PDGFR, etc.	Type II	2012
Ponatinib	ABL1	Type II	2012
Dabrafenib	BRAF	Type I	2013
Trametinib	MEK1/2	Type III	2013
Afatinib	EGFR	Type I, Irreversible	2013
Ibrutinib	BTK	Type I, Irreversible	2013
Idelalisib	PI3K-delta	Type I	2014
Nintedanib	VEGFR, PDGFR, etc.	Type II	2014
Ceritinib	ALK, MET	Type II	2014
Lenvatinib	VEGFR, PDGFR, c-KIT, FGFR, etc.	Type V †	2015
Palbociclib	CDK4/6	Type I	2015

† In a recent publication, lenvatinib was proposed as a special class of kinase inhibitors (the so-called type V inhibitors). Compounds of this class are those binding both the ATP-binding site and the neighboring allosteric region in kinases with DFG-in conformation [210].

Table 1.3: FDA approved kinase inhibitors alonsgide their pharmacological target, binding mode and year of FDA approval.

The approval of imatinib in 2001 radically transformed the treatment of Philadelphia chromosome-positive (Ph+) chronic myelogenous leukemia (CML). Since then, more than 27 different small molecule kinase inhibitors have been approved by the FDA (Table 1.3), and many others are currently in clinical trials for the treatment of cancer. Despite of their great success in cancer treatment, small molecule kinase inhibitors suffer from major limitations that need to be addressed to improve their clinical impact. Next, I outline some of their most important challenges and limitations:

1. Of the total 538 estimated human kinases [198], only a few, and most of them belonging to the tyrosine kinase group, have been pharmacologically targeted by small molecule inhibitors. It is thus necessary to increase the spectrum of clinically targeted kinases. Moreover, the increment of the number of targeted kinases would create new therapeutic opportunities for disorders where kinases play an important, but yet clinically unexplored role.

2. Other important limitation is the lack of specificity of many small molecule kinase inhibitors. This is mainly consequence of the high structural conservation of the ATP-binding site in kinases, which causes that a large number of inhibitors interact with more than one target [211]. The multitarget nature of many kinase inhibitors gives rise to severe side effects that dramatically restricts its applicability in the clinics. Fostering the development of type III allosteric inhibitors would lead into more selective inhibitors preventing the appearance of unexpected toxicities.

3. Related to the previous point, the mechanistic basis of unexpected toxicities observed during the preclinical and clinical stages need to be studied more rigorously. Improved documentation of kinase inhibitor specificity and observed toxicities would provide a valuable database for understanding whether there are particular kinases of which inhibition should be particularly avoided [212].

4. The most important limitation of small-molecule kinase inhibitors, in particular, and targeted cancer therapies, in general, is the rapid acquisition of drug resistance. The duration of clinical benefits is frequently short, which dramatically restricts the utility of many targeted cancer therapies.

The next section will focus on the mechanistic basis of resistance to targeted cancer therapies, with particular emphasis on mutations of kinase targets altering the efficacy of the treatment.

1.4.3. Resistance to targeted cancer therapies

Drug resistance is one of the major problems in cancer treatment. Resistance to both chemotherapy agents and targeted cancer therapies is hampering the success of many anticancer treatments. The advent of new high throughput genomic technologies and its combination with bioinformatics and systems biology approaches, have enhanced the understanding of the molecular events underpinning treatment failure. However, we are still far from overcoming the emergence of drug resistance in cancer targeted therapies. One of the reasons leading to the complex drug resistance problem is the considerable amount of molecular mechanisms leading to drug resistance [213]. Some mechanisms of resistance for specific molecular targets share many features with the classic cytotoxic chemotherapy, while others, are genuine to the targeted cancer therapies. One of the classic mechanisms of resistance is caused by the pharmacokinetics properties (ADME) of the drugs, which added to the limited amount of drug that can be systemically administered confine the amount of drug that reaches the tumour. That means that the concentration of drug that eventually reaches the cancer cells is lower than the one required to perform the desired antiproliferative activity [214]. At the level of the tumour, various resistance mechanisms can operate, including activation of survival signalling pathways and the inactivation of downstream death signalling pathways [215], oncogenic bypass and pathway redundancy [216], factors associated to the tumour microenviroment [217] or epigenetic alterations [218].

Importantly, alterations in the drug target is one of the most frequent mechanism of resistance in targeted cancer therapies. Increased expression of the molecular target reduces the effectiveness of inhibitors of these targets because more target molecules be must be inhibited to have a effective therapeutic effect. For instance, the androgen receptor (AR) is genomically amplified in aproximately 30% of prostate cancers with acquired resistance to standard androgen deprivation therapy. In such cases, treatment using testosterone lowering drugs such as

leuprolide and AR antagonists such as bicalutamide, is not effective and alternative treatments are thus necessary [219].

Most of the small molecule targeted cancer therapies target oncogenic kinases that are responsible of the tumour proliferation and/or development (Subsection 1.4.2). Mutations of these oncogenic kinases can alter the binding of the small molecule kinase inhibitor giving rise to a reestablishment of the tumour proliferation activity. Moreover, evidence continues to emerge that cancers are characterized by extensive intratumour genetic heterogeneity (ITH), and that patients being considered for treatment with a targeted agent might, therefore, already possess resistance to the drug in a small population of cells [220]. This mechanism of resistance has been extensively reported over the last years [221, 222]. The first mutation identified in patients with CML who relapsed on treatment with imatinib was in the gatekeeper residue of BCR-ABL1, T315. This missense mutation hinders imatinib binding while preserving the catalytic activity that is needed for the oncogenic function of BCR-ABL1 [223]. Since then, more than one hundred BCR-ABL1 different mutations have been reported [224]. Second-generation BCR-ABL1 inhibitors (*e.g.*, nilotinib, dasatinib and bosutinib) were developed for the treatment of patients with acquired resistance to imatinib. However, the BCR-ABL1 T315I gatekeeper mutation confers resistance to all currently approved ABL1 TKIs other than the newest of these molecules, ponatinib [224]. Similarly to the BCR-ABL1 case, acquired resistance to EGFR inhibitors such as gefitinib or erlotinib is common (Subsection 1.4.2). Studies showed that more than 50% of the EGFR-gefitinib resistant cases harbored a secondary EGFR-T790M mutation [225]. Such is the impact of this mutation for the treatment of NSCLCs, that a third new generation of EGFR-T790M selective inhibitors have been designed to overcome resistance to EGFR-T790M positive patients [226, 227]. However, recent studies showed that third generation irreversible EGFR inhibitors also experience the emergence of resistance mutations [228]. Crizotinib is a small molecule kinase inhibitor approved for the treatment of some types of NSCLC. It performs its pharmacological activity by targeting the ALK and ROS-1 kinases (Subsection 1.4.2). Some studies in small cohorts of patients have already shown that mutations in the ALK kinase domain such as G1269A, L1198F, L1196M can drive acquired resistance to crizotinib [229, 230]. The mutations described to date span the entire ALK kinase domain and may also confer variable degrees of resistance to

second-generation ALK inhibitors [231].

The ABL1-imatinib, EGFR-gefitinib and ALK-crizotinib cases are probably the best studied examples of resistance to small molecule targeted cancer therapies. However, individual studies have shown that many other kinase mutations are drivers of drug resistance. Moreover, future improvement in the sensitivity of genomic high throughput technologies will, most likely, increase the number of these mutants [220]. To make things worse, treatments should be able to deal with ITH, which affects variation in drug response predominantly at the cellular level [232]. Hence, there is a need to rationally design cancer treatments able to overcome resistance due to mutations in drug targets. Fostering early detection of pre-existing or emerging drug resistance could enable more personalized use of targeted cancer therapy, as patients could be stratified to receive the treatments that are most likely to be effective. Another solution to the challenge of polygenic cancer drug resistance is rational combinatorial treatments, such as combinatorial targeted therapy [195], combination of chemotherapy with targeted therapy [233] or the promising combination of immunotherapy with targeted therapies [234]. Therefore, achieving the full potential of targeted cancer therapy is dependent on the identification of the best possible drug combinations. The resulting combinatorial explosion will require use of new technologies, including large-scale genomics and network biology with associated computational approaches [235]. In fact, computational methods are being applied to explain and predict therapeutic resistance [236, 237], tumour clonal evolution [238, 239] and potential drug combinations [240, 241]. Chapter 3.3 presents a computational approach that predicts mutations with potential to confer resistance to small molecule targeted cancer therapy. The computational framework exemplifies how computational methods can help to rationally design alternative non-resistant cancer targeted therapies.

1.5. Motivation

As we have shown over the Introduction, interaction between small molecule and proteins governs many of the cellular functions (Subsection 1.1.6). Such is the importance, that modulation of the protein function by a small molecules has been used by to treat multiple conditions (Subsection 1.2). In fact, the

discovery and pharmacological development of antibiotics for the treatment of infectious diseases such as TB, has dramatically improved our lifespan (Subsection 1.3.1). More recently, the emergence of targeted cancer therapies also transformed the landscape of cancer treatment, moving from the traditionally cytotoxic chemotherapy to more precise targeted therapies (Subsection 1.4). Research progresses are partly thanks to the development of methods to experimentally determine the 3D structure of proteins (Subsection 1.1.2). Furthermore, *in-silico* methods have contributed to characterize protein and ligand interactions, with the added value of providing new predicted interactions (Subsection 1.1.8). Indeed, the ability of computational methods to predict small molecule-protein interactions has significantly improved over the last decade. One of the main reasons for this improvement is the emergence of databases gathering large amount of structural and therapeutic information [95, 92, 242], which enables computational models to increase their predictive power by learning from new relationships and restrains. However, computational methods for ligand-target prediction should be able to tailor the requirements of drug discovery industry where 1) the 3D structure of the interaction is completely essential and 2) the screening process usually involves a very large set of candidate compounds. These two requirements are fulfilled by the method presented in Subsection 3.1, which exemplifies its applicability on a large set of antitubercular compounds in Subsection 3.2. //

We also discussed about how targeted cancer therapy has transformed cancer treatments (Subsection 1.4). Concretely, since the approval of imatinib in 2001, more than 25 small molecule kinase inhibitors have been approved by the FDA (Table 1.3), while many others are currently in clinical trials for the treatment of this devastating disease (Subsection 1.4.2). However, small molecule targeted cancer therapies suffer from a major limitation, the clinical benefit of patients receiving this therapies is often temporal (Subsection 1.4.3). Multiple tumor-intrinsic mechanisms confer resistance to drug targeted cancer therapies [213]. Among these mechanisms, mutations in drug targets is one of most frequently observed in the clinics. Numerous studies have been conducted to understand and overcome resistance due to mutations in drug targets. However, these studies are often limited to a small and clinically reported number of mutations. Therefore, there is a need for 1) a comprehensive characterization of the tumour mutational landscape with the potential to confer resistance and 2) providing al-

ternative treatments in those cases where the drug-resistant mutants are already present in the tumour burden. These two objectives are accomplished in the study introduced in Subsection 3.3.

2 Objectives

This thesis aims to fulfil the following specific objectives:

I To develop a publicly accessible network-based ligand target prediction method that provides large scale and structurally detailed predictions.

II To validate the method predicting the human targets of all small molecule FDA-approved drugs.

III To apply the method antitubercular compounds in order to identify their protein targets on the MTB structural proteome. The results should be combined with the predicted targets from other approaches exploring different methodological spaces.

IV To develop a model that predicts the cancer associated mutations with the highest chances to be responsible of resistance to a particular targeted cancer therapy.

V For those mutations classified as treatment-threatening, to identify alternative therapies overcoming resistance.

Objectives i) and ii) are presented in the 3.1 section. Concretely, this chapter presents nAnnolyze, a comparative docking approach that predicts structurally detailed ligand target interactions at proteome scale. nAnnolyze is a network-based version of the prior Annolyze [110]. The chapter also presents a virtual screening performed by nAnnolyze to predict the human targets of all FDA-approved drugs. Finally, the nAnnolyze network, method and predictions are publicly available at `http://nannolyze.cnag.cat/`.

Point iii) is discussed in chapter 3.2. More specifically, this section presents the computational predictions of three orthogonal approaches to identify new

protein targets that are likely to interact with a set of compounds with bioactivity against MTB. The resulting combination of the predictions, including the structural complexes by nAnnolyze, are publicly available online at [LINK NOT WORKING].

Finally, points iv) and v) are presented in chapter 1.4.3. Concretely, this chapter introduces a framework that 1) estimates the cancer associated likelihood of a mutation on a protein target 2) predicts the resistance potential of each of the target mutations using structural information of the interaction 3) provides alternative compounds for those mutations predicted to confer resistance to a given targeted cancer therapy [Actualizar].

3 Results

3.1. Ligand-Target Prediction by Structural Network Biology using nAnnolyze

This section presents nAnnolyze, a method for predicting large-scale and structurally detailed compound-protein interactions. nAnnolyze was applied to identify the human targets of all FDA-approved drugs. The method alongside all the predictions are available online in `http://nannolyze.cnag.cat/`.

Manuscripts presented in this section:

> Martínez-Jiménez, F., & Marti-Renom, M. a. (2015). **Ligand-Target Prediction by Structural Network Biology Using nAnnoLyze**. PloS Computational Biology, 11(3), e1004157. doi:10.1371/ journal.pcbi.1004

3.2. Target Prediction for two Open Acess Sets of Compounds Active against *Mycobacterium tuberculosis*

This section presents the application of nAnnolyze to predict the MTB targets of a set of compounds with antitubercular activity. The target predictions from nAnnolyze are combined with those resulting from the application of two other methods exploring different methodological spaces (i.e., the structural space, the chemical space and the historical space). The compounds and the predictions are publicly available at [MIRAR URL].

Manuscripts presented in this section:

Martínez-Jiménez, F., Papadatos, G., Yang, L., Wallace, I. M., Kumar, V., Pieper, U., ... Marti-Renom, M. a. (2013). **Target Prediction for an Open Access Set of Compounds Active against Mycobacterium tuberculosis**. PLoS Computational Biology, 9(10), e1003253. doi:10.1371/journal.pcbi.1003253

Rebollo-Lopez, M. J., Lelièvre, J., Alvarez-Gomez, D., Castro-Pichel, J., Martínez-Jiménez, F., Papadatos, G., ... Barros-Aguire, D. (2015). **Release of 50 new, drug-like compounds and their computational target predictions for open source anti-tubercular drug discovery**. PloS One, 10(12), e0142293. doi:10.1371/ journal.pone.0142293

3.3. Rational design of non-resistant targeted cancer therapies

Paper new. Paste here when submitted.

4 Discussion

This thesis presented a computational study of the structural interaction between small molecules and their protein targets with the main focus on extracting their therapeutic potential. Concretely, chapter 3.1 presented a comparative docking 1.1.9 method that predicts structurally detailed protein-ligand interactions at proteome scale. It exemplified its applicability by predicting the human targets of all small molecule FDA-approved drugs. A second application of nAnnolyze in MTB is presented in section 3.2. This chapter showed the computational identification of the MTB targets for two sets of compounds with known antitubercular activity. It used the combination of three methods exploring different methodological spaces (i.e., the structural space, the chemical space and the historical space) to give more robustness to the predictions. The open access profile of both nAnnolyze and the application in MTB, led to the development of a website that enables the interplay with the method and the results. Finally, chapter 1.4 introduced a computational model that predicts cancer associated mutations with the highest chances to confer resistance to a targeted therapy. Furthermore, it provided alternative treatments for those mutations identified as highly resistance-like. Each of the specific points presented in the studies are thoroughly analyzed in the pertinent discussion of the manuscripts. Hence, this discussion is focused on analyzing the impact to the scientific community, reviewing the main limitations and discussing future perspectives of the presented studies.

4.1. nAnnolyze: predicting large scale and structurally detailed ligand-target interaction using a network-based representation

4.1.1. Main findings

nAnnolyze is a network-based version of the Annolyze method [110]. It relies on a comparative docking approach that 1) predicts the protein targets of small molecules and 2) identifies the binding location on the 3D structure of the protein. The evaluation of the performance showed that nAnnolyze enables large-scale annotation and analysis of compound-protein pairs. The application of the method to predict the human targets of all small molecule FDA-approved showed

its ability to identify therapeutically relevant compound-target pairs. Moreover, nAnnolyze also predicted new unseen interactions between FDA-approved drugs and human proteins. Finally, the method alongside all the predictions are publicly available at `http://nannolyze.cnag.cat/`.

4.1.2. Impact of the presented research

To our knowledge, nAnnolyze is one the few methods that enables structurally detailed large-scale screening of compounds against an entire proteome. Unlike free-structure methods, which do not provide structural information about the binding, and virtual docking methods, which require considerable amount of resources for large-scale screenings, nAnnolyze fulfills two of the most important needs in the modern drug discovery paradigm 1) applicability to large-scale screenings and 2) the inclusion of structural information in the predictions.

The application to the human proteome provided an immense collection of compounds-target pairs amenable to me analyzed in future studies. Concretely, such information can be used to identify compound off-targets responsible of clinically reported side effects. The manuscript illustrated this possibility with the example of new predicted targets for the multikinase inhibitor sorafenib. Moreover, exploring the collection of compound-target pairs may give rise to the identification of new therapeutically relevant interactions with the potential to be further explored by drug repurposing approaches.

Finally, since the methods is fully available online, the scientific community can benefit from the usage by anonymously screening their own compounds against the human and MTB structural proteomes.

4.1.3. Limitations

One of the major limitations of the method is implicit in its own definition. nAnnolyze is a structure based approach and consequently its application is restricted to those proteins with either a experimentally determined 3D structure or a sequence amenable to be accurately modeled by comparative modeling ap-

proaches. Currently, approximately 40% of the human proteome fulfills these requirements.

The application of a comparative docking approach may also lead to the inclusion of bias towards structurally conserved protein pockets. Therefore, non-conserved allosteric pockets, which are often remarkably valuable to develop selective inhibitors 1.4.2, may be neglected by the method. Similarly, novel non-frequent compound scaffolds are also penalized in the search because of their limited availability in the explored structural space.

As mentioned above, comparative docking methods are usually faster than virtual docking approaches. However, they are generally slower than free-structure methods, which makes them a viable option only once the number of candidate compounds have narrowed down. Ideally, drug discovery early stages would choose the ligand-target prediction method that better fits to the characteristics of the screening (i.e., compound collection size, number of targets, stage of development, etc). Alternatively, the combination of different computational methods can increase both the predictive power and the confidence of the resulting predictions 3.2.

nAnnlyze does not include information about the type of interaction between the compound and the predicted targets (i.e., antagonist, agonist, inhibitor, etc.). Moreover, the graph does not include either information about the binding affinity of the compounds with their co-crystallized protein targets. Such information may play an important role in the decision of whether a predicted compound-target pair is suitable for further exploration.

Finally, one limitation of the website is related to the fact that it does not include the possibility to perform an screening against your own protein target. This application is frequently observed in academia, when the inhibition of the candidate target may validate the testing hypothesis.

4.1.4. Future perspectives

Future versions of nAnnolyze will benefit from the raise of publicly available structural data. Thanks to the initiatives such as the PSI [11] or the Structural

Genomics Consortium [12] the number of experimentally determined 3D structures will significantly increase over the next years. Therefore, the number of modellable proteins will raise simultaneously, which eventually will lead to a significant increase of the number of proteins to which structure-based methods can be applied. Additionally, the raise in the number of deposited structures in the PDB will likely increase the chemical spectrum of the co-crystallized compounds, decreasing thus the aforementioned compound's scaffold bias.

The flexibility of a network-based approach facilitates the integration of multiple sources of information. As discussed above, information as the type of interaction or compound binding affinity would improve both the level of detail and the quality of the predictions. Moreover, integration of protein-protein interaction (PPI) information, target-disease and target-side effect associations would enable more realistic selection of the molecular target to intervene.

Another feature amenable to be improved is the graph search algorithm. nAnnolyze uses the Dijkstra's algorithm [243] to find the shortest pathways between compounds and protein targets. Other popular network search algorithms include random walk [244] or network propagation [245].

One of the near future plan consist on applying nAnnolyze to alternative set of candidate molecules and protein targets. While chapter 3.2 presents the application in two different set of antitubercular compounds, future applications in other organisms and collection of compounds would significantly increase the value of the method. Moreover, the method would benefit from the feedback received after its application.

Finally, one of the most important goals and challenges of computational drug discovery is the translation into the experimental field. Experimental validation of the predictions would not only add more confidence to the method, but would also be useful to identify those cases where the approach is more suitable for.

4.2. Target prediction for two set of compounds active against MTB

4.2.1. Main findings

This chapter presented the application of three ligand-target prediction methods to identify the MTB targets of two sets of compounds with known antitubercular activity. The methods explored three different methodological spaces, including the structural space by nAnnolyze, the chemical space and the historical space. The final compound-target set was the result of combining the individual predictions by the three approaches. The first application on a set of 776 compounds resulted in the identification of 139 MTB targets involved in 71 unique pathways. The second application in a set of 50 antitubercular compounds identified 21 MTB targets involved in 13 different metabolic pathways. Subsequent analysis of the target essentially revealed a significant number of predicted targets previously annotated as essential for the survival of MTB. Moreover, study of the metabolic pathways associated with the predicted MTB targets revealed an significant enrichment in amino acid metabolism pathways, which are known to be essential for the survival of the bacterial. Finally, all the compounds alongside the predicted MTB were publicly delivered in both studies.

4.2.2. Impact of the presented research

To our knowledge this is the first virtual screening performed by three orthogonal approaches to systematically identify protein targets for small molecules. The resulting *metapredictor* is more robust than the individual methods, adding not only target and compound coverage, but also increasing confidence to the predictions. From the logistic perspective, this application also exemplified how computational methods can play a significant role in the drug discovery process. Moreover, it illustrated how drug discovery benefits from collaborations between big pharmaceutical companies and academia. From the clinical perspective, the delivery of the compounds and predictions fosters open access drug discovery against TB. Finally, future experimental validation and putative clinical development would significantly increase the value of the presented studies.

4.2.3. Limitations

Most of the methodological limitations are inherent to the applied ligand-target prediction methods. Concretely, nAnnolyze's limitations, which are discussed above, are also applicable to this study.

Some limitations and problems may emerge due to the combination of different methodologies. Although combination of multiple methods reduce individual biases limiting the amount of noise on the final predictions, it might also give rise to the loss of unique, and perhaps real, compound-target pairs predicted by a single method.

Interestingly, compounds with activity against human targets could be compromised by toxicity. However, the study did not specifically check for human off-targets because of two main reasons. First, the antitubercular compounds have been filtered by a human *in-vitro* toxicity assay. Second, empirical evidence suggests that antibiotics side effects are mostly due to high treatment doses associated with damage to the liver [246].

The study assumed that all the compounds perform their anti-infective activity through the modulation of a protein target. However, there are antibiotics that perform their activity through different mechanisms of action [158]. In such cases, the method will not identify the actual mechanism of action.

The study did not include any information about drug resistance. One of the major problems of bacterial infections is the emergence of resistant strains not responding to standard treatments. Such information was not considered in the model and may have a dramatic impact in the development of new non-resistant antibiotics.

Finally, none of the predictions have been experimentally validated in this study. Therefore, all the provided predictions need to be carefully considered.

4.2.4. Future perspectives

One of the near future goals consist on applying the same methodology to new sets of antitubercular compounds. Moreover, we are planning to apply similar combination of methods to other diseases and organisms.

Future applications would benefit from the improvement of each of the methods used in the study. Furthermore, including new features such as compound's predicted side effects or ADMET profile, would increase the level of detail of the predictions enabling the prioritization of those compounds with higher chances to become an approved drug.

Similarly to targeted cancer therapy, antibiotics suffer from a major limitation. The effect of the treatment is often temporary due to the emergence of drug-resistant strains. TB is not an exception, the emergence of MDR-TB and XDR-TB jeopardizes the prognosis of many TB patients. Combinatorial regimes are a promising alternative to overcome resistance to cancer 1.4 and bacterial infections 1.3.1 treatments. Therefore, computational identification of antibiotics combination can lead to the development of less resistant therapies. In our specific case, after the initial annotation of compound's targets, we would include a second layer identifying compounds combinations with positive resistance profiles.

Finally, as discussed above 4.1.4, experimental validation of the compound-target pairs would significantly increase the value of the presented work, taking a step forward in the fight against MTB infection.

4.3. Rational design of non-resistant targeted cancer therapies

4.3.1. Main findings

4.3.2. Impact of the presented research

4.3.3. Limitations

4.3.4. Future perspectives

5 Conclusions

- Conclusions.. (To extract from objetives –¿ Conclusions)

Bibliography

[1] A Kessel and N Ben-Tal. *Introduction to Proteins: Structure, Function, and Motion*. Chapman & Hall/CRC Mathematical and Computational Biology. CRC Press, 2010. ISBN: 9781439810729 (cit. on p. 1).

[2] B Alberts. *Molecular Biology of the Cell*. Molecular Biology of the Cell: Reference Edition v. 1. Garland Science, 2008. ISBN: 9780815341116 (cit. on p. 2).

[3] Wolfgang Kabsch and Christian Sander. «Dictionary of protein secondary structure: Pattern recognition of hydrogen-bonded and geometrical features». In: *Biopolymers* 22.12 (1983), pp. 2577–2637. ISSN: 1097-0282. DOI: 10.1002/bip.360221211 (cit. on p. 2).

[4] K A Dill. «Dominant forces in protein folding.» In: *Biochemistry* 29.31 (1990), pp. 7133–7155. ISSN: 0006-2960. DOI: 10.1021/bi00483a001 (cit. on p. 2).

[5] C Chothia and Arthur M Lesk. «The relation between the divergence of sequence and structure in proteins.» In: *The EMBO journal* 5.4 (1986), pp. 823–6. ISSN: 0261-4189. DOI: 060fehlt (cit. on p. 4).

[6] Buyong Ma et al. «Protein-protein interactions: structurally conserved residues distinguish between binding sites and exposed protein surfaces.» In: *Proceedings of the National Academy of Sciences of the United States of America* 100.10 (2003), pp. 5772–7. ISSN: 0027-8424. DOI: 10.1073/pnas.1030237100 (cit. on p. 4).

[7] Rajkumar Sasidharan and Cyrus Chothia. «The selection of acceptable protein mutations». In: *Proceedings of the National Academy of Sciences of the United States of America* 104.24 (2007), pp. 10080–10085. ISSN: 0027-8424. DOI: 10.1073/pnas.0703737104 (cit. on p. 4).

[8] Annabel E Todd, Christine A Orengo, and Janet M Thornton. «Evolution of Function in Protein Superfamilies , from a Structural Perspective». In: (2001). DOI: 10.1006/jmbi.2001.4513 (cit. on p. 4).

[9] J C KENDREW et al. «Structure of myoglobin: A three-dimensional Fourier synthesis at 2 A. resolution.» In: *Nature* 185.4711 (Feb. 1960), pp. 422–7. ISSN: 0028-0836 (cit. on p. 5).

[10] Helen M Berman et al. «The Protein Data Bank». In: *Nucleic Acids Research* 28.1 (Jan. 2000), pp. 235–242. DOI: 10.1093/nar/28.1.235 (cit. on p. 5).

[11] John C Norvell and Jeremy M Berg. «Update on the Protein Structure Initiative». In: *Structure* 15.12 (Dec. 2007), pp. 1519–1522. ISSN: 0969-2126. DOI: http://dx.doi.org/10.1016/j.str.2007.11.004 (cit. on pp. 5, 11, 13, 56).

[12] Opher Gileadi et al. *The scientific impact of the Structural Genomics Consortium: A protein family and ligand-centered approach to medically-relevant human proteins*. Sept. 2007. DOI: 10.1007/s10969-007-9027-2 (cit. on pp. 5, 57).

[13] M S Smyth and J H J Martin. «x Ray crystallography». In: *Molecular Pathology* 53.1 (Feb. 2000), pp. 8–14. DOI: 10.1136/mp.53.1.8 (cit. on p. 7).

[14] Roslyn M Bill et al. «Overcoming barriers to membrane protein structure determination». In: *Nature Biotechnology* 29.4 (2011), pp. 335–340. ISSN: 1087-0156. DOI: 10.1038/nbt.1833 (cit. on p. 7).

[15] Michael B Yaffe. «X-ray crystallography and structural biology». In: *Critical Care Medicine* 33.Suppl (Dec. 2005), S435–S440. ISSN: 0090-3493. DOI: 10.1097/01.CCM.0000191719.66383.01 (cit. on p. 7).

[16] A L Morris et al. «Stereochemical quality of protein structure coordinates.» In: *Proteins* 12.4 (Apr. 1992), pp. 345–64. ISSN: 0887-3585. DOI: 10.1002/prot.340120407 (cit. on p. 7).

[17] G. Wider. «Structure determination of biological macromolecules in solution using nuclear magnetic resonance spectroscopy.» In: *BioTechniques* 29.6 (Dec. 2000), 1278–82, 1284–90, 1292 passim. ISSN: 0736-6205 (cit. on pp. 7, 8).

[18] Ewen Callaway. «The Revolution Will Not Be Crystallized». In: *Nature* 525.7568 (Sept. 2015), pp. 172–174. ISSN: 0163-6545. DOI: 10.1215/01636545-2009-008 (cit. on p. 8).

[19] Heena Khatter et al. «Structure of the human 80S ribosome.» In: *Nature* 520.7549 (Apr. 2015), pp. 640–5. ISSN: 1476-4687. DOI: 10.1038/nature14427 (cit. on p. 8).

[20] Jianhua Zhao, Samir Benlekbir, and John L Rubinstein. «Electron cryomicroscopy observation of rotational states in a eukaryotic V-ATPase.» In: *Nature* 521.7551 (May 2015), pp. 241–5. ISSN: 1476-4687. DOI: 10.1038/nature14365 (cit. on p. 8).

[21] Maofu Liao et al. «Structure of the TRPV1 ion channel determined by electron cryo-microscopy.» In: *Nature* 504.7478 (Dec. 2013), pp. 107–12. ISSN: 1476-4687. DOI: 10.1038/nature12822 (cit. on p. 8).

[22] Xiao-chen Bai et al. «An atomic structure of human y-secretase.» In: *Nature* 525.7568 (Sept. 2015), pp. 212–7. ISSN: 1476-4687. DOI: 10.1038/nature14892 (cit. on p. 8).

[23] B Rost and C Sander. «Bridging the Protein Sequence-Structure Gap by Structure Predictions». In: *Annual Review of Biophysics and Biomolecular Structure* 25.1 (June 1996), pp. 113–136. ISSN: 1056-8700. DOI: 10.1146/annurev.bb.25.060196.000553 (cit. on p. 9).

[24] D T Jones, W R Taylor, and J M Thornton. «A new approach to protein fold recognition.» In: *Nature* 358.6381 (July 1992), pp. 86–9. ISSN: 0028-0836. DOI: 10.1038/358086a0 (cit. on p. 9).

[25] J U Bowie, R Lüthy, and D Eisenberg. «A method to identify protein sequences that fold into a known three-dimensional structure.» In: *Science (New York, N.Y.)* 253.5016 (July 1991), pp. 164–70. ISSN: 0036-8075 (cit. on p. 9).

[26] Jooyoung Lee, Sitao Wu, and Yang Zhang. «From Protein Structure to Function with Bioinformatics». In: ed. by Daniel John Rigden. Dordrecht: Springer Netherlands, 2009. Chap. Ab Initio, pp. 3–25. ISBN: 978-1-4020-9058-5. DOI: 10.1007/978-1-4020-9058-5_1 (cit. on p. 9).

[27] Daniel Russel et al. «Putting the Pieces Together: Integrative Modeling Platform Software for Structure Determination of Macromolecular Assemblies». In: *PLoS Biology* 10.1 (Jan. 2012), e1001244. ISSN: 1544-9173. DOI: 10.1371/journal.pbio.1001244 (cit. on p. 9).

[28] Narayanan Eswar et al. «Comparative protein structure modeling using MODELLER.» en. In: *Current protocols in protein science / editorial board, John E. Coligan ... [et al.]* Chapter 2 (Dec. 2007), Unit 2.9. ISSN: 1934-3663. DOI: 10.1002/0471140864.ps0209s50 (cit. on pp. 10, 14).

[29] D Baker. «Protein structure prediction and structural genomics». In: *Science* 294 (2001), pp. 93–96. ISSN: 00368075. DOI: 10.1126/science.1065659 (cit. on p. 11).

[30] Su Yun Chung and S Subbiah. «A structural explanation for the twilight zone of protein sequence homology». In: *Structure* 4.10 (Oct. 1996), pp. 1123–1127. ISSN: 09692126. DOI: 10.1016/S0969-2126(96)00119-0 (cit. on p. 11).

[31] C Sander and R Schneider. «Database of homology-derived protein structures and the structural meaning of sequence alignment.» In: *Proteins* 9.1 (Jan. 1991), pp. 56–68. ISSN: 0887-3585. DOI: 10.1002/prot.340090107 (cit. on pp. 11, 13).

[32] Evandro Ferrada and Francisco Melo. «Nonbonded terms extrapolated from nonlocal knowledge-based energy functions improve error detection in near-native protein structure models». In: *Protein Science* 16.7 (2007), pp. 1410–1421. ISSN: 1469-896X. DOI: 10.1110/ps.062735907 (cit. on p. 12).

[33] A Fiser, R K Do, and A Sali. «Modeling of loops in protein structures.» In: *Protein science : a publication of the Protein Society* 9.9 (Sept. 2000), pp. 1753–73. ISSN: 0961-8368. DOI: 10.1110/ps.9.9.1753 (cit. on p. 12).

[34] A Kidera. «Enhanced conformational sampling in Monte Carlo simulations of proteins: application to a constrained peptide.» In: *Proceedings of the National Academy of Sciences of the United States of America* 92.21 (Oct. 1995), pp. 9886–9. ISSN: 0027-8424 (cit. on p. 12).

[35] D.B. McGarrah and R.S. Judson. «Analysis of the genetic algorithm method of molecular conformation determination». In: *Journal of Computational Chemistry* 14.11 (Nov. 1993), pp. 1385–1395. ISSN: 0192-8651. DOI: 10.1002/jcc.540141115 (cit. on p. 12).

[36] Domenico Cozzetto et al. «Assessment of predictions in the model quality assessment category». In: *Proteins: Structure, Function, and Bioinformatics* 69.S8 (2007), pp. 175–183. ISSN: 1097-0134. DOI: 10.1002/prot.21669 (cit. on p. 12).

[37] Francisco Melo, Roberto Sánchez, and Andrej Sali. «Statistical potentials for fold assessment.» In: *Protein science : a publication of the Protein Society* 11.2 (Mar. 2002), pp. 430–48. ISSN: 0961-8368. DOI: 10.1002/pro.110430 (cit. on p. 12).

[38] R Samudrala and J Moult. «An all-atom distance-dependent conditional probability discriminatory function for protein structure prediction.» In: *Journal of molecular biology* 275.5 (Feb. 1998), pp. 895–916. ISSN: 0022-2836. DOI: 10.1006/jmbi.1997.1479 (cit. on p. 12).

[39] DAVID A CASE et al. «The Amber Biomolecular Simulation Programs». In: *Journal of computational chemistry* 26.16 (Dec. 2005), pp. 1668–1688. ISSN: 0192-8651. DOI: 10.1002/jcc.20290 (cit. on p. 13).

[40] Bernard R. Brooks et al. «CHARMM: A program for macromolecular energy, minimization, and dynamics calculations». In: *Journal of Computational Chemistry* 4.2 (1983), pp. 187–217. ISSN: 0192-8651. DOI: 10.1002/jcc.540040211 (cit. on p. 13).

[41] Federico Fogolari, Alessandro Brigo, and Henriette Molinari. «Protocol for MM/PBSA molecular dynamics simulations of proteins.» In: *Biophysical journal* 85.1 (July 2003), pp. 159–66. ISSN: 0006-3495. DOI: 10.1016/S0006-3495(03)74462-2 (cit. on p. 13).

[42] P D Thomas and K A Dill. «Statistical potentials extracted from protein structures: how accurate are they?» In: *Journal of molecular biology* 257.2 (Mar. 1996), pp. 457–69. ISSN: 0022-2836. DOI: 10.1006/jmbi.1996.0175 (cit. on p. 13).

[43] Min-yi Shen and Andrej Sali. «Statistical potential for assessment and prediction of protein structures». In: *Protein Science : A Publication of the Protein Society* 15.11 (Nov. 2006), pp. 2507–2524. ISSN: 0961-8368. DOI: 10.1110/ps.062416606 (cit. on p. 13).

[44] Hongyi Zhou and Yaoqi Zhou. «Distance-scaled, finite ideal-gas reference state improves structure-derived potentials of mean force for structure selection and stability prediction.» In: *Protein science : a publication of the Protein Society* 11.11 (Dec. 2002), pp. 2714–26. ISSN: 0961-8368. DOI: 10.1110/ps.0217002 (cit. on p. 13).

[45] Manfred J Sippl. «Knowledge-based potentials for proteins». In: *Current Opinion in Structural Biology* 5.2 (Apr. 1995), pp. 229–235. ISSN: 0959440X. DOI: 10.1016/0959-440X(95)80081-6 (cit. on p. 13).

[46] Changsheng Du and Xin Xie. «G protein-coupled receptors as therapeutic targets for multiple sclerosis.» In: *Cell research* 22.7 (July 2012), pp. 1108–28. ISSN: 1748-7838. DOI: 10.1038/cr.2012.87 (cit. on p. 14).

[47] Kim R Kampen. «Membrane Proteins: The Key Players of a Cancer Cell». In: *The Journal of Membrane Biology* 242.2 (2011), pp. 69–74. ISSN: 1432-1424. DOI: 10.1007/s00232-011-9381-7 (cit. on p. 14).

[48] Julia Koehler Leman, Martin B Ulmschneider, and Jeffrey J Gray. «Computational modeling of membrane proteins». In: *Proteins* 83.1 (Jan. 2015), pp. 1–24. ISSN: 0887-3585. DOI: 10.1002/prot.24703 (cit. on p. 14).

[49] Marc A Martí-Renom et al. «Comparative Protein Structure Modeling of Genes and Genomes». In: *Annual Review of Biophysics and Biomolecular Structure* 29.1 (June 2000), pp. 291–325. ISSN: 1056-8700. DOI: 10.1146/annurev.biophys.29.1.291 (cit. on p. 14).

[50] Lars Malmström and David R Goodlett. «Protein structure modeling.» In: *Methods in molecular biology (Clifton, N.J.)* 673 (2010), pp. 63–72. ISSN: 1940-6029. DOI: 10.1007/978-1-60761-842-3_5 (cit. on p. 14).

[51] A Sali and T L Blundell. «Comparative protein modelling by satisfaction of spatial restraints.» In: *Journal of molecular biology* 234.3 (Dec. 1993), pp. 779–815. ISSN: 0022-2836. DOI: 10.1006/jmbi.1993.1626 (cit. on p. 14).

[52] Marco Biasini et al. «SWISS-MODEL: modelling protein tertiary and quaternary structure using evolutionary information». In: *Nucleic Acids Research* 42.Web Server issue (July 2014), W252–W258. ISSN: 0305-1048. DOI: 10.1093/nar/gku340 (cit. on p. 14).

[53] Johannes Söding, Andreas Biegert, and Andrei N Lupas. «The HHpred interactive server for protein homology detection and structure prediction». In: *Nucleic Acids Research* 33.Web Server issue (July 2005), W244–W248. ISSN: 0305-1048. DOI: 10.1093/nar/gki408 (cit. on p. 14).

[54] Jianyi Yang et al. «The I-TASSER Suite: protein structure and function prediction.» In: *Nature methods* 12.1 (Jan. 2015), pp. 7–8. ISSN: 1548-7105. DOI: 10.1038/nmeth.3213 (cit. on p. 14).

[55] Ambrish Roy, Alper Kucukural, and Yang Zhang. «I-TASSER: a unified platform for automated protein structure and function prediction». In: *Nature protocols* 5.4 (Apr. 2010), pp. 725–738. ISSN: 1754-2189. DOI: 10.1038/nprot.2010.5 (cit. on p. 14).

[56] Yang Zhang. «I-TASSER server for protein 3D structure prediction». In: *BMC Bioinformatics* 9 (Jan. 2008), p. 40. ISSN: 1471-2105. DOI: 10.1186/1471-2105-9-40 (cit. on p. 14).

[57] David E Kim, Dylan Chivian, and David Baker. «Protein structure prediction and analysis using the Robetta server». In: *Nucleic Acids Research* 32.Web Server issue (July 2004), W526–W531. ISSN: 0305-1048. DOI: 10.1093/nar/gkh468 (cit. on p. 14).

[58] Morten Källberg et al. «Protein Structure Prediction». In: ed. by Daisuke Kihara. New York, NY: Springer New York, 2014. Chap. RaptorX se, pp. 17–27. ISBN: 978-1-4939-0366-5. DOI: 10.1007/978-1-4939-0366-5_2 (cit. on p. 14).

[59] P A Bates et al. «Enhancement of protein modeling by human intervention in applying the automatic programs 3D-JIGSAW and 3D-PSSM.» In: *Proteins* Suppl 5 (Jan. 2001), pp. 39–46. ISSN: 0887-3585 (cit. on p. 14).

[60] G. Vriend. «WHAT IF: A molecular modeling and drug design program». In: *Journal of Molecular Graphics* 8.1 (Mar. 1990), pp. 52–56. ISSN: 02637855. DOI: 10.1016/0263-7855(90)80070-V (cit. on p. 14).

[61] Jane S. Richardson. *THE ANATOMY AND TAXONOMY OF PROTEIN STRUCTURE*. Vol. 34. 1981, pp. 167–339. ISBN: 9780120342341. DOI: 10.1016/S0065-3233(08)60520-3 (cit. on p. 15).

[62] Peer Bork. «Shuffled domains in extracellular proteins». In: *FEBS Letters* 286.1-2 (July 1991), pp. 47–54. ISSN: 1873-3468. DOI: 10.1016/0014-5793(91)80937-X (cit. on p. 15).

[63] D B Wetlaufer. «Nucleation, rapid folding, and globular intrachain regions in proteins.» In: *Proceedings of the National Academy of Sciences of the United States of America* 70.3 (1973), pp. 697–701. ISSN: 0027-8424. DOI: 10.1073/pnas.70.3.697 (cit. on p. 15).

[64] Suhail A Islam, Jingchu Luo, and M J Sternberg. «Identification and analysis of domains in proteins.» In: *Protein engineering* 8.6 (June 1995), pp. 513–25. ISSN: 0269-2139 (cit. on p. 15).

[65] Jung-Hoon Han et al. «The folding and evolution of multidomain proteins.» In: *Nature reviews. Molecular cell biology* 8.4 (2007), pp. 319–330. ISSN: 1471-0072. DOI: 10.1038/nrm2144 (cit. on p. 15).

[66] Cyrus Chothia et al. «Evolution of the protein repertoire.» In: *Science (New York, N.Y.)* 300.5626 (2003), pp. 1701–3. ISSN: 1095-9203. DOI: 10.1126/science.1085371 (cit. on p. 15).

[67] Christine Vogel et al. «Structure, function and evolution of multidomain proteins». In: *Current Opinion in Structural Biology* 14.2 (2004), pp. 208–216. ISSN: 0959440X. DOI: 10.1016/j.sbi.2004.03.011 (cit. on p. 15).

[68] G Apic, J Gough, and S a Teichmann. «Domain combinations in archaeal, eubacterial and eukaryotic proteomes.» In: *Journal of molecular biology* 310.2 (2001), pp. 311–325. ISSN: 0022-2836. DOI: 10.1006/jmbi.2001.4776 (cit. on p. 15).

[69] Alexey G. Murzin et al. «SCOP: A structural classification of proteins database for the investigation of sequences and structures». In: *Journal of Molecular Biology* 247.4 (1995), pp. 536–540. ISSN: 00222836. DOI: 10.1016/S0022-2836(05)80134-2 (cit. on p. 15).

[70] Ca Orengo et al. «CATH - a hierarchic classification of protein domain structures». In: *Structure* March (1997), pp. 1093–1109. ISSN: 09692126. DOI: 10.1016/S0969-2126(97)00260-8 (cit. on p. 15).

[71] A Bateman et al. «The Pfam protein families database». In: *Nucleic Acids Research* 28.1 (2002), pp. 276–280. ISSN: 0305-1048 (Print) 0305-1048 (Linking). DOI: gkd038[pii] (cit. on p. 15).

[72] Sarah Hunter et al. «InterPro: The integrative protein signature database». In: *Nucleic Acids Research* 37.SUPPL. 1 (2009), pp. 211–215. ISSN: 03051048. DOI: 10.1093/nar/gkn785 (cit. on p. 15).

[73] Friedrich Cramer. «Emil Fischer's Lock and Key Hypothesis after 100 years towards a Supracellular Chemistry». In: *Perspectives in Supramolecular Chemistry*. John Wiley & Sons, Ltd., 1994, pp. 1–23. ISBN: 9780470511411. DOI: 10.1002/9780470511411.ch1 (cit. on p. 16).

[74] D E Koshland. «Enzyme flexibility and enzyme action». In: *Journal of Cellular and Comparative Physiology* 54.S1 (Dec. 1959), pp. 245–258. ISSN: 1553-0809. DOI: 10.1002/jcp.1030540420 (cit. on p. 16).

[75] Jacque Monod, Jeffries Wyman, and Jean-Pierre Changeux. «On the nature of allosteric transitions: A plausible model». In: *Journal of Molecular Biology* 12.1 (May 1965), pp. 88–118. ISSN: 00222836. DOI: 10.1016/S0022-2836(65)80285-6 (cit. on p. 16).

[76] Akio Kitao, Steven Hayward, and Nobuhiro Go. «Energy landscape of a native protein: Jumping among minima model». In: *Proteins: Structure, Function, and Bioinformatics* 33.4 (Dec. 1998), pp. 496–517. ISSN: 1097-0134. DOI: 10.1002/(SICI)1097-0134(19981201)33:4<496::AID-PROT4>3.0.CO;2-1 (cit. on p. 16).

[77] G A Petsko and D Ringe. «Fluctuations in Protein Structure from X-Ray Diffraction». In: *Annual Review of Biophysics and Bioengineering* 13.1 (June 1984), pp. 331–371. ISSN: 0084-6589. DOI: 10.1146/annurev.bb.13.060184.001555 (cit. on p. 16).

[78] J Foote and C Milstein. «Conformational isomerism and the diversity of antibodies.» In: *Proceedings of the National Academy of Sciences of the United States of America* 91.22 (Oct. 1994), pp. 10370–4. ISSN: 0027-8424 (cit. on p. 16).

[79] Leo C James, Pietro Roversi, and Dan S Tawfik. «Antibody multispecificity mediated by conformational diversity.» In: *Science (New York, N.Y.)* 299.5611 (Feb. 2003), pp. 1362–7. ISSN: 1095-9203. DOI: 10.1126/science.1079731 (cit. on p. 16).

[80] Michael F Dunn. «Protein-Ligand Interactions: General Description». In: *eLS*. John Wiley & Sons, Ltd, 2001. ISBN: 9780470015902. DOI: 10.1038/npg.els.0001340 (cit. on p. 16).

[81] Julien Michel, Julian Tirado-Rives, and William L Jorgensen. «Energetics of Displacing Water Molecules from Protein Binding Sites: Consequences for Ligand Optimization». In: *Journal of the American Chemical Society* 131.42 (Oct. 2009), pp. 15403–15411. ISSN: 0002-7863. DOI: 10.1021/ja906058w (cit. on p. 19).

[82] Krishna Ravindranathan et al. «Improving MM-GB SA Scoring through the Application of the Variable Dielectric Model». In: *Journal of chemical theory and computation* 7.12 (Dec. 2011), pp. 3859–3865. ISSN: 1549-9618. DOI: 10.1021/ct200565u (cit. on p. 19).

[83] Julien Michel, Marcel L Verdonk, and Jonathan W Essex. «Protein Ligand Binding Affinity Predictions by Implicit Solvent Simulations: A Tool for Lead Optimization». In: *Journal of Medicinal Chemistry* 49.25 (Dec. 2006), pp. 7427–7439. ISSN: 0022-2623. DOI: 10.1021/jm061021s (cit. on p. 19).

[84] Hao-Yang Liu, Sam Z Grinter, and Xiaoqin Zou. «Multiscale generalized Born modeling of ligand binding energies for virtual database screening». In: *The journal of physical chemistry. B* 113.35 (Sept. 2009), pp. 11793–11799. ISSN: 1520-6106. DOI: 10.1021/jp901212t (cit. on p. 19).

[85] Yipin Lu et al. «Analysis of Ligand-Bound Water Molecules in High-Resolution Crystal Structures of Protein-Ligand Complexes». In: *Journal of Chemical Information and Modeling* 47.2 (Mar. 2007), pp. 668–675. ISSN: 1549-9596. DOI: 10.1021/ci6003527 (cit. on p. 19).

[86] Kim A Sharp and Barry. Honig. «Calculating total electrostatic energies with the nonlinear Poisson-Boltzmann equation». In: *The Journal of Physical Chemistry* 94.19 (Sept. 1990), pp. 7684–7692. ISSN: 0022-3654. DOI: 10.1021/j100382a068 (cit. on p. 19).

[87] Donald Bashford and David A Case. «GENERALIZED BORN MODELS OF MACROMOLECULAR SOLVATION EFFECTS». In: *Annual Review of Physical Chemistry* 51.1 (Oct. 2000), pp. 129–152. ISSN: 0066-426X. DOI: 10.1146/annurev.physchem.51.1.129 (cit. on p. 19).

[88] Richard A. Friesner et al. «Glide: A New Approach for Rapid, Accurate Docking and Scoring». In: *Journal of Medicinal Chemistry* 47.7 (2004), pp. 1739–1749. ISSN: 00222623. DOI: 10.1021/jm0306430. arXiv: arXiv:1011.1669v3 (cit. on pp. 19, 21).

[89] Claudia Steffen et al. «TmoleX a graphical user interface for TURBOMOLE.» In: *Journal of computational chemistry* 31.16 (2010), pp. 2967–2970. ISSN: 1096-987X. DOI: 10.1002/jcc. arXiv: NIHMS150003 (cit. on p. 19).

[90] Peter Csermely et al. «Structure and dynamics of molecular networks: A novel paradigm of drug discovery: A comprehensive review». In: *Pharmacology and Therapeutics* 138.3 (2013), pp. 333–408. ISSN: 01637258. DOI: 10.1016/j.pharmthera.2013.01.016. arXiv: 1210.0330 (cit. on p. 20).

[91] Philip Prathipati and Kenji Mizuguchi. «Systems biology approaches to a rational drug discovery paradigm». In: *Current Topics in Medicinal Chemistry* 15.999 (2015), pp. 1–1. ISSN: 15680266. DOI: 10.2174/1568026615666150826114524 (cit. on p. 20).

[92] A Patrícia Bento et al. «The ChEMBL bioactivity database: an update». In: *Nucleic acids research* 42.Database issue (Jan. 2014), pp. D1083–90. ISSN: 0305-1048. DOI: 10.1093/nar/gkt1031 (cit. on pp. 20, 49).

[93] Feng Zhu et al. «Therapeutic target database update 2012: a resource for facilitating target-oriented drug discovery». In: *Nucleic Acids Research* 40.Database issue (Jan. 2012), pp. D1128–D1136. ISSN: 0305-1048. DOI: 10.1093/nar/gkr797 (cit. on p. 20).

[94] Liegi Hu et al. «Binding MOAD (Mother Of All Databases).» In: *Proteins* 60.3 (Aug. 2005), pp. 333–40. ISSN: 1097-0134. DOI: 10.1002/prot.20512 (cit. on p. 20).

[95] Tiqing Liu et al. «BindingDB: a web-accessible database of experimentally determined protein–ligand binding affinities». In: *Nucleic Acids Research* 35.Database issue (Jan. 2007), pp. D198–D201. ISSN: 0305-1048. DOI: 10.1093/nar/gkl999 (cit. on pp. 20, 49).

[96] Sunghwan Kim et al. «PubChem Substance and Compound databases». In: *Nucleic Acids Research* 44.D1 (Jan. 2016), pp. D1202–D1213. DOI: 10.1093/nar/gkv951 (cit. on p. 20).

[97] Yanli Wang et al. «PubChem BioAssay: 2014 update». In: *Nucleic Acids Research* 42.Database issue (Jan. 2014), pp. D1075–D1082. ISSN: 0305-1048. DOI: 10.1093/nar/gkt978 (cit. on p. 20).

[98] John J Irwin et al. «ZINC: A Free Tool to Discover Chemistry for Biology». In: *Journal of Chemical Information and Modeling* 52.7 (July 2012), pp. 1757–1768. ISSN: 1549-9596. DOI: 10.1021/ci3001277 (cit. on p. 20).

[99] Hernán Alonso, Andrey A. Bliznyuk, and Jill E. Gready. «Combining docking and molecular dynamic simulations in drug design». In: *Medicinal Research Reviews* 26.5 (2006), pp. 531–568. ISSN: 01986325. DOI: 10.1002/med.20067 (cit. on p. 21).

[100] Garrett M Morris et al. «AutoDock4 and AutoDockTools4: Automated Docking with Selective Receptor Flexibility». In: *Journal of computational chemistry* 30.16 (Dec. 2009), pp. 2785–2791. ISSN: 0192-8651. DOI: 10.1002/jcc.21256 (cit. on p. 21).

[101] Oleg Trott and Arthur J Olson. «AutoDock Vina: improving the speed and accuracy of docking with a new scoring function, efficient optimization and multithreading». In: *Journal of computational chemistry* 31.2 (Jan. 2010), pp. 455–461. ISSN: 0192-8651. DOI: 10.1002/jcc.21334 (cit. on p. 21).

[102] Todd J a. Ewing and Irwin D Kuntz. «Critical evaluation of search algorithms for automated molecular docking and database screening». In: *Journal of Computational Chemistry* 18 (1997), pp. 1175–1189. ISSN: 0192-8651 (cit. on p. 21).

[103] Matthias Rarey et al. «A Fast Flexible Docking Method using an Incremental Construction Algorithm». In: *Journal of Molecular Biology* 261.3 (Aug. 1996), pp. 470–489. ISSN: 0022-2836. DOI: http://dx.doi.org/10.1006/jmbi.1996.0477 (cit. on p. 21).

[104] G Jones et al. «Development and validation of a genetic algorithm for flexible docking.» In: *Journal of molecular biology* 267.3 (Apr. 1997), pp. 727–48. ISSN: 0022-2836. DOI: 10.1006/jmbi.1996.0897 (cit. on p. 21).

[105] Julie R Schames et al. «Discovery of a novel binding trench in HIV integrase.» In: *Journal of medicinal chemistry* 47.8 (Apr. 2004), pp. 1879–81. ISSN: 0022-2623. DOI: 10.1021/jm0341913 (cit. on p. 21).

[106] Istvan J Enyedy et al. «Discovery of Small-Molecule Inhibitors of Bcl-2 through Structure-Based Computer Screening». In: *Journal of Medicinal Chemistry* 44.25 (Dec. 2001), pp. 4313–4324. ISSN: 0022-2623. DOI: 10.1021/jm010016f (cit. on p. 21).

[107] Eric Vangrevelinghe et al. «Discovery of a Potent and Selective Protein Kinase CK2 Inhibitor by High-Throughput Docking». In: *Journal of Medicinal Chemistry* 46.13 (June 2003), pp. 2656–2662. ISSN: 0022-2623. DOI: 10.1021/jm030827e (cit. on p. 21).

[108] D Kitchen et al. «Docking and scoring in virtual screening for drug discovery: methods and applications». In: *Nature Reviews Drug Discovery* 3.11 (2004), pp. 935–949. ISSN: 1474-1784. DOI: 10.1038/nrd1549 (cit. on p. 21).

[109] Sara Reardon. «Project ranks billions of drug interactions.» In: *Nature* 503.7477 (2013), pp. 449–50. ISSN: 1476-4687. DOI: 10.1038/503449a (cit. on p. 21).

[110] Marc a Marti-Renom et al. «The AnnoLite and AnnoLyze programs for comparative annotation of protein structures.» In: *BMC bioinformatics* 8 Suppl 4 (Jan. 2007), S4. ISSN: 1471-2105. DOI: 10.1186/1471-2105-8-S4-S4 (cit. on pp. 21, 23, 50, 54).

[111] Roman A Laskowski, James D Watson, and Janet M Thornton. «ProFunc: a server for predicting protein function from 3D structure». In: *Nucleic Acids Research* 33.suppl 2 (July 2005), W89–W93. DOI: 10.1093/nar/gki414 (cit. on p. 21).

[112] David Lee, Oliver Redfern, and Christine Orengo. «Predicting protein function from sequence and structure». In: *Nat Rev Mol Cell Biol* 8.12 (Dec. 2007), pp. 995–1005. ISSN: 1471-0072 (cit. on p. 21).

[113] Liisa Holm and Päivi Rosenström. «Dali server: conservation mapping in 3D». In: *Nucleic Acids Research* 38.suppl 2 (July 2010), W545–W549. DOI: 10.1093/nar/gkq366 (cit. on p. 21).

[114] Brice Hoffmann et al. «A new protein binding pocket similarity measure based on comparison of clouds of atoms in 3D: application to ligand prediction.» In: *BMC bioinformatics* 11 (Jan. 2010), p. 99. ISSN: 1471-2105. DOI: 10.1186/1471-2105-11-99 (cit. on p. 23).

[115] Mark N Wass, Lawrence A Kelley, and Michael J E Sternberg. «3DLigandSite: predicting ligand-binding sites using similar structures». In: *Nucleic Acids Research* 38.Web Server issue (July 2010), W469–W473. ISSN: 0305-1048. DOI: 10.1093/nar/gkq406 (cit. on p. 23).

[116] John A Capra et al. «Predicting protein ligand binding sites by combining evolutionary sequence conservation and 3D structure.» In: *PLoS computational biology* 5.12 (Dec. 2009), e1000585. ISSN: 1553-7358. DOI: 10.1371/journal.pcbi.1000585 (cit. on p. 23).

[117] Olga V Kalinina et al. «Combinations of protein-chemical complex structures reveal new targets for established drugs.» In: *PLoS computational biology* 7.5 (May 2011), e1002043. ISSN: 1553-7358. DOI: 10.1371/journal.pcbi.1002043 (cit. on p. 23).

[118] J Drews. «Drug discovery: a historical perspective.» In: *Science* 287.5460 (2000), pp. 1960–64. ISSN: 0036-8075. DOI: 10.1126/science.287.5460.1960 (cit. on p. 23).

[119] Graham L Patrick. «History of Drug Discovery». In: *eLS*. John Wiley & Sons, Ltd, 2001. ISBN: 9780470015902. DOI: 10.1002/9780470015902.a0003090.pub2 (cit. on p. 23).

[120] Lisa Hutchinson and Rebecca Kirk. «High drug attrition rates–where are we going wrong?» In: *Nature reviews. Clinical oncology* 8.4 (Apr. 2011), pp. 189–90. ISSN: 1759-4782. DOI: 10.1038/nrclinonc.2011.34 (cit. on p. 24).

[121] Mark A Lindsay. «Target discovery.» In: *Nature reviews. Drug discovery* 2.10 (Oct. 2003), pp. 831–8. ISSN: 1474-1776. DOI: 10.1038/nrd1202 (cit. on p. 24).

[122] J. P. Hughes et al. «Principles of early drug discovery». In: *British Journal of Pharmacology* 162.6 (2011), pp. 1239–1249. ISSN: 00071188. DOI: 10.1111/j.1476-5381.2010.01127.x (cit. on p. 24).

[123] Peter Imming, Christian Sinning, and Achim Meyer. «Drugs, their targets and the nature and number of drug targets.» In: *Nature reviews. Drug discovery* 5.10 (2006), pp. 821–834. ISSN: 1474-1776. DOI: 10.1038/nrd2132 (cit. on p. 24).

[124] J E Klees and R Joines. «Occupational health issues in the pharmaceutical research and development process.» eng. In: *Occupational medicine (Philadelphia, Pa.)* 12.1 (1997), pp. 5–27. ISSN: 0885-114X (cit. on p. 24).

[125] Steven M Paul et al. «How to improve R&D productivity: the pharmaceutical industry's grand challenge.» In: *Nature reviews. Drug discovery* 9.3 (2010), pp. 203–214. ISSN: 1474-1776. DOI: 10.1038/nrd3078 (cit. on p. 25).

[126] Jack W Scannell et al. «Diagnosing the decline in pharmaceutical R&D efficiency.» In: *Nature reviews. Drug discovery* 11.3 (2012), pp. 191–200. ISSN: 1474-1784. DOI: 10.1038/nrd3681 (cit. on p. 26).

[127] Gregory Sliwoski et al. «Computational methods in drug discovery.» In: *Pharmacological reviews* 66.1 (2014), pp. 334–95. ISSN: 1521-0081. DOI: 10.1124/pr.112.007336 (cit. on p. 26).

[128] Yongliang Yang, S. James Adelstein, and Amin I. Kassis. «Target discovery from data mining approaches». In: *Drug Discovery Today* 14.3-4 (2009), pp. 147–154. ISSN: 13596446. DOI: 10.1016/j.drudis.2008.12.005 (cit. on p. 26).

[129] Assaf Gottlieb et al. «PREDICT: a method for inferring novel drug indications with application to personalized medicine.» In: *Molecular systems biology* 7.1 (Apr. 2011), p. 496. ISSN: 1744-4292. DOI: 10.1038/msb.2011.26 (cit. on p. 26).

[130] M Zhang et al. «The orphan disease networks». In: *Am J Hum Genet* 88 (2011). DOI: 10.1016/j.ajhg.2011.05.006 (cit. on p. 28).

[131] Yutaka Fukuoka, Daiki Takei, and Hisamichi Ogawa. «A two-step drug repositioning method based on a protein-protein interaction network of genes shared by two diseases and the similarity of drugs.» In: *Bioinformation* 9.2 (2013), pp. 89–93. ISSN: 0973-2063. DOI: 10.6026/97320630009089 (cit. on p. 28).

[132] Karthik Raman and Nagasuma Chandra. «Mycobacterium tuberculosis interactome analysis unravels potential pathways to drug resistance.» In: *BMC microbiology* 8 (2008), p. 234. ISSN: 1471-2180. DOI: 10.1186/1471-2180-8-234 (cit. on pp. 28, 35, 36).

[133] Suthat Phaiphinit et al. «In silico multiple-targets identification for heme detoxification in the human malaria parasite Plasmodium falciparum». In: *Infection, Genetics and Evolution* 37 (Jan. 2016), pp. 237–244. ISSN: 1567-1348. DOI: http://dx.doi.org/10.1016/j.meegid.2015.11.025 (cit. on p. 28).

[134] Jouhyun Jeon et al. «A systematic approach to identify novel cancer drug targets using machine learning, inhibitor design and high-throughput screening». In: *Genome Medicine* 6.7 (2014), pp. 1–18. ISSN: 1756-994X. DOI: 10.1186/s13073-014-0057-7 (cit. on p. 28).

[135] Regina Augustin et al. «Computational identification and experimental validation of microRNAs binding to the Alzheimer-related gene ADAM10». In: *BMC Medical Genetics* 13.1 (2012), pp. 1–12. ISSN: 1471-2350. DOI: 10.1186/1471-2350-13-35 (cit. on p. 28).

[136] Francisco Martínez-Jiménez et al. «Target Prediction for an Open Access Set of Compounds Active against Mycobacterium tuberculosis». In: *PLoS Computational Biology* 9.10 (Oct. 2013). Ed. by Alexander Donald MacKerell, e1003253. ISSN: 1553-7358. DOI: 10.1371/journal.pcbi.1003253 (cit. on pp. 28, 36).

[137] Roger Perkins et al. «Quantitative structure-activity relationship methods: perspectives on drug discovery and toxicology.» In: *Environmental toxicology and chemistry / SETAC* 22.8 (2003), pp. 1666–1679. ISSN: 1092-874X. DOI: 10.1897/01-171 (cit. on p. 28).

[138] Julie E Penzotti, Gregory A Landrum, and Santosh Putta. «Building predictive ADMET models for early decisions in drug discovery.» In: *Current opinion in drug discovery & development* 7.1 (Jan. 2004), pp. 49–61. ISSN: 1367-6733 (cit. on p. 28).

[139] Olga Obrezanova et al. «Gaussian Processes: A Method for Automatic QSAR Modeling of ADME Properties». In: *Journal of Chemical Information and Modeling* 47.5 (Sept. 2007), pp. 1847–1857. ISSN: 1549-9596. DOI: 10.1021/ci7000633 (cit. on p. 28).

[140] Fumitaka Yoshida and John G Topliss. «QSAR Model for Drug Human Oral Bioavailability». In: *Journal of Medicinal Chemistry* 43.13 (June 2000), pp. 2575–2585. ISSN: 0022-2623. DOI: 10.1021/jm0000564 (cit. on p. 28).

[141] Jitender Verma, Vijay M. Khedkar, and Evans C. Coutinho. «3D-QSAR in Drug Design». In: *Current Topics in Medicinal Chemistry* 10.1 (2010), pp. 95–115. ISSN: 15680266. DOI: 10.2174/156802610790232260 (cit. on pp. 28, 29).

[142] R D Cramer, D E Patterson, and J D Bunce. «Comparative molecular field analysis (CoMFA)». In: *Journal of the American Chemical Society* 110.18 (1988), pp. 5959–5967. ISSN: 0002-7863. DOI: 10.1021/ja00226a005 (cit. on p. 28).

[143] A J Hopfinger et al. «Construction of 3D-QSAR Models Using the 4D-QSAR Analysis Formalism». In: *Journal of the American Chemical Society* 119.43 (Oct. 1997), pp. 10509–10524. ISSN: 0002-7863. DOI: 10.1021/ja9718937 (cit. on p. 29).

[144] S. Ekins et al. «Three- and four-dimensional-quantitative structure activity relationship (3D/4D-qsar) analyses of CYP2C9 inhibitors». In: *Drug Metabolism and Disposition* 28.8 (Aug. 2000), pp. 994–1002. ISSN: 00909556 (cit. on p. 29).

[145] Manisha Iyer and A J Hopfinger. «Treating Chemical Diversity in QSAR Analysis Modeling Diverse HIV-1 Integrase Inhibitors Using 4D Fingerprints». In: *Journal of Chemical Information and Modeling* 47.5 (Sept. 2007), pp. 1945–1960. ISSN: 1549-9596. DOI: 10.1021/ci700153g (cit. on p. 29).

[146] Nelilma Correia Romeiro et al. «Construction of 4D-QSAR Models for Use in the Design of Novel p38-MAPK Inhibitors». In: *Journal of Computer-Aided Molecular Design* 19.6 (), pp. 385–400. ISSN: 1573-4951. DOI: 10.1007/s10822-005-7927-4 (cit. on p. 29).

[147] Carolina H. Andrade et al. «4D-QSAR: Perspectives in drug design». In: *Molecules* 15.5 (2010), pp. 3281–3294. ISSN: 14203049. DOI: 10.3390/molecules15053281 (cit. on p. 29).

[148] Angelo Vedani and Max Dobler. «5D-QSAR: The Key for Simulating Induced Fit». In: *Journal of Medicinal Chemistry* 45.11 (May 2002), pp. 2139–2149. ISSN: 0022-2623. DOI: 10.1021/jm011005p (cit. on p. 29).

[149] Angelo Vedani, Max Dobler, and Markus A Lill. «Combining Protein Modeling and 6D-QSAR. Simulating the Binding of Structurally Diverse Ligands to the Estrogen Receptor». In: *Journal of Medicinal Chemistry* 48.11 (June 2005), pp. 3700–3703. ISSN: 0022-2623. DOI: 10.1021/jm050185q (cit. on p. 29).

[150] Christopher A Lipinski. «Lead and drug-like compounds: the rule of five revolution.» In: *Drug discovery today. Technologies* 1.4 (Dec. 2004), pp. 337–41. ISSN: 1740-6749. DOI: 10.1016/j.ddtec.2004.11.007 (cit. on p. 29).

[151] G Richard Bickerton et al. «Quantifying the chemical beauty of drugs.» en. In: *Nature chemistry* 4.2 (Feb. 2012), pp. 90–8. ISSN: 1755-4349. DOI: 10.1038/nchem.1243 (cit. on p. 30).

[152] Jérémy Besnard et al. «Automated design of ligands to polypharmacological profiles». In: *Nature* 492.7428 (Dec. 2012), pp. 215–220. ISSN: 0028-0836. DOI: 10.1038/nature11691 (cit. on p. 30).

[153] Clare M. Lewandowski, New Co-investigator, and Clare M. Lewandowski. «WHO Global tuberculosis report 2015». In: *WHO Global tuberculosis report 2015* 1 (2015), pp. 1689–1699. ISSN: 1098-6596. DOI: 10.1017/CBO9781107415324.004. arXiv: arXiv:1011.1669v3 (cit. on pp. 30, 31, 33).

[154] D W Connell et al. «Update on tuberculosis: TB in the early 21st century.» In: *European respiratory review : an official journal of the European Respiratory Society* 20.120 (June 2011), pp. 71–84. ISSN: 1600-0617. DOI: 10.1183/09059180.00000511 (cit. on p. 30).

[155] M Berry and O M Kon. «Multidrug- and extensively drug-resistant tuberculosis: an emerging threat.» In: *European respiratory review : an official journal of the European Respiratory Society* 18.114 (Dec. 2009), pp. 195–7. ISSN: 1600-0617. DOI: 10.1183/09059180.00005209 (cit. on p. 30).

[156] Patrice Trouiller et al. «Drug development for neglected diseases: a deficient market and a public-health policy failure». In: *The Lancet* 359.9324 (June 2002), pp. 2188–2194. ISSN: 01406736. DOI: 10.1016/S0140-6736(02)09096-7 (cit. on pp. 30, 31).

[157] J. M. Conly and B. L. Johnston. «Where are all the new antibiotics? The new antibiotic paradox». In: *Canadian Journal of Infectious Diseases and Medical Microbiology* 16.3 (May 2005), pp. 159–160. ISSN: 17129532 (cit. on p. 30).

[158] Losee L Ling et al. «A new antibiotic kills pathogens without detectable resistance». In: *Nature* 517.7535 (2015), pp. 455–459. ISSN: 1476-4687. DOI: 10.1038/nature14098 (cit. on pp. 30, 59).

[159] *Philanthropists unite to accelerate global fight against tuberculosis with combined $20 million gift to Broad Institute* (cit. on p. 31).

[160] Deepak K Karki et al. «Costs of a successful public-private partnership for TB control in an urban setting in Nepal». In: *BMC Public Health* 7.1 (2007), pp. 1–12. ISSN: 1471-2458. DOI: 10.1186/1471-2458-7-84 (cit. on p. 32).

[161] K J Murthy et al. «Public-private partnership in tuberculosis control: experience in Hyderabad, India.» In: *The international journal of tuberculosis and lung disease : the official journal of the International Union against Tuberculosis and Lung Disease* 5.4 (Apr. 2001), pp. 354–9. ISSN: 1027-3719 (cit. on p. 32).

[162] Xun Lei et al. «Public–private mix for tuberculosis care and control: a systematic review». In: *International Journal of Infectious Diseases* 34 (May 2015), pp. 20–32. ISSN: 1201-9712. DOI: http://dx.doi.org/10.1016/j.ijid.2015.02.015 (cit. on p. 32).

[163] Edison S Zuniga, Julie Early, and Tanya Parish. «The future for early-stage tuberculosis drug discovery». In: *Future Microbiology* 10.2 (2015), pp. 217–229. ISSN: 1746-0913. DOI: 10.2217/fmb.14.125 (cit. on p. 33).

[164] P J Brennan. «Structure, function, and biogenesis of the cell wall of Mycobacterium tuberculosis». In: *Tuberculosis* 83.1–3 (Feb. 2003), pp. 91–97. ISSN: 1472-9792. DOI: http://dx.doi.org/10.1016/S1472-9792(02)00089-6 (cit. on p. 33).

[165] Liliana Rodrigues et al. «Contribution of efflux activity to isoniazid resistance in the Mycobacterium tuberculosis complex». In: *Infection, Genetics and Evolution* 12.4 (June 2012), pp. 695–700. ISSN: 1567-1348. DOI: http://dx.doi.org/10.1016/j.meegid.2011.08.009 (cit. on p. 33).

[166] Ujjini H Manjunatha and Paul W Smith. «Perspective: Challenges and opportunities in TB drug discovery from phenotypic screening». In: *Bioorganic & Medicinal Chemistry* 23.16 (Aug. 2015), pp. 5087–5097. ISSN: 0968-0896. DOI: http://dx.doi.org/10.1016/j.bmc.2014.12.031 (cit. on p. 34).

[167] Lluís Ballell et al. «Fueling open-source drug discovery: 177 small-molecule leads against tuberculosis.» In: *ChemMedChem* 8.2 (Mar. 2013), pp. 313–21. ISSN: 1860-7187. DOI: 10.1002/cmdc.201200428 (cit. on p. 34).

[168] Sae Woong Park et al. «Target-Based Identification of Whole-Cell Active Inhibitors of Biotin Biosynthesis in Mycobacterium tuberculosis». In: *Chemistry & Biology* 22.1 (Jan. 2015), pp. 76–86. ISSN: 1074-5521. DOI:

http://dx.doi.org/10.1016/j.chembiol.2014.11.012 (cit. on p. 34).

[169] Garima Arora et al. «High Throughput Screen Identifies Small Molecule Inhibitors Specific for Mycobacterium tuberculosis Phosphoserine Phosphatase». In: *Journal of Biological Chemistry* (July 2014). DOI: 10.1074/jbc.M114.597682 (cit. on p. 34).

[170] Karthik Raman and Nagasuma Chandra. «Mycobacterium tuberculosis interactome analysis unravels potential pathways to drug resistance.» In: *BMC microbiology* 8.1 (2008), p. 234. ISSN: 1471-2180. DOI: 10.1186/1471-2180-8-234 (cit. on p. 34).

[171] Gregory J Crowther et al. «Identification of attractive drug targets in neglected-disease pathogens using an in silico approach.» In: *PLoS neglected tropical diseases* 4.8 (Aug. 2010), e804. ISSN: 1935-2735. DOI: 10.1371/journal.pntd.0000804 (cit. on pp. 34, 35).

[172] Sean Ekins et al. «A collaborative database and computational models for tuberculosis drug discovery.» In: *Molecular bioSystems* 6.5 (2010), pp. 840–851. ISSN: 1742-206X. DOI: 10.1039/b917766c (cit. on pp. 34, 35).

[173] Sarah L Kinnings et al. «Drug discovery using chemical systems biology: repositioning the safe medicine Comtan to treat multi-drug and extensively drug resistant tuberculosis.» In: *PLoS computational biology* 5.7 (July 2009), e1000423. ISSN: 1553-7358. DOI: 10.1371/journal.pcbi.1000423 (cit. on pp. 34, 35).

[174] Marc R de Jonge et al. «A computational model of the inhibition of Mycobacterium tuberculosis ATPase by a new drug candidate R207910». In: *Proteins: Structure, Function, and Bioinformatics* 67.4 (June 2007), pp. 971–980. ISSN: 1097-0134. DOI: 10.1002/prot.21376 (cit. on pp. 34, 36).

[175] Ashutosh Kumar and Mohammad Imran Siddiqi. «Receptor based 3D-QSAR to identify putative binders of Mycobacterium tuberculosis Enoyl acyl carrier protein reductase». In: *Journal of Molecular Modeling* 16.5 (2009), pp. 877–893. ISSN: 0948-5023. DOI: 10.1007/s00894-009-0584-0 (cit. on pp. 34, 36).

[176] Jocelyne M Lew et al. «TubercuList – 10 years after». In: *Tuberculosis* 91.1 (Jan. 2011), pp. 1–7. ISSN: 1472-9792. DOI: http://dx.doi.org/10.1016/j.tube.2010.09.008 (cit. on pp. 34, 36).

[177] Leandro Radusky et al. «TuberQ: a Mycobacterium tuberculosis protein druggability database». In: *Database* 2014 (Jan. 2014). DOI: 10.1093/database/bau035 (cit. on pp. 34, 36).

[178] Sean Ekins, Alex M Clark, and Malabika Sarker. «TB Mobile: a mobile app for anti-tuberculosis molecules with known targets». In: *Journal of Cheminformatics* 5.1 (2013), p. 13. ISSN: 1758-2946. DOI: 10.1186/1758-2946-5-13 (cit. on pp. 34, 36).

[179] Alex M. Clark, Malabika Sarker, and Sean Ekins. «New target prediction and visualization tools incorporating open source molecular fingerprints for TB Mobile 2.0». In: *Journal of Cheminformatics* 6.1 (2014), pp. 1–17. ISSN: 17582946. DOI: 10.1186/s13321-014-0038-2 (cit. on pp. 34, 36).

[180] María Jose Rebollo-Lopez et al. «Release of 50 new, drug-like compounds and their computational target predictions for open source anti-tubercular drug discovery.» In: *PloS one* 10.12 (2015), e0142293. ISSN: 1932-6203. DOI: 10.1371/journal.pone.0142293 (cit. on p. 36).

[181] International Agency for Research on Cancer. *World Cancer Report 2014*. Tech. rep. 2014 (cit. on p. 37).

[182] A M Scott, J D Wolchok, and L J Old. «Antibody therapy of cancer». In: *Nat Rev Cancer* 12.4 (2012), pp. 278–287. ISSN: 1474-1768. DOI: 10.1038/nrc3236 (cit. on pp. 37, 38).

[183] D J Jonker et al. «Cetuximab for the treatment of colorectal cancer». In: *N Engl J Med* 357.20 (Nov. 2007), pp. 2040–2048. ISSN: 0028-4793. DOI: 10.1056/NEJMoa071834 (cit. on p. 37).

[184] Mohamedtaki a Tejani, Roger B Cohen, and Ranee Mehra. «The contribution of cetuximab in the treatment of recurrent and/or metastatic head and neck cancer.» In: *Biologics : targets & therapy* 4 (Aug. 2010), pp. 173–185. ISSN: 1177-5475. DOI: 10.2147/BTT.S3050 (cit. on p. 37).

[185] Michael A Postow, Margaret K Callahan, and Jedd D Wolchok. «Immune Checkpoint Blockade in Cancer Therapy.» In: *Journal of clinical oncology : official journal of the American Society of Clinical Oncology* 33.17 (June 2015), JCO.2014.59.4358–. ISSN: 1527-7755. DOI: 10.1200/JCO.2014.59.4358 (cit. on p. 38).

[186] S. Demko et al. «FDA Drug Approval Summary: Alemtuzumab as Single-Agent Treatment for B-Cell Chronic Lymphocytic Leukemia». In: *The Oncologist* 13.2 (Feb. 2008), pp. 167–174. ISSN: 1083-7159. DOI: 10.1634/theoncologist.2007-0218 (cit. on p. 38).

[187] Napoleone Ferrara et al. «Discovery and development of bevacizumab, an anti-VEGF antibody for treating cancer.» In: *Nature reviews. Drug discovery* 3.5 (May 2004), pp. 391–400. ISSN: 1474-1776. DOI: 10.1038/nrd1381 (cit. on p. 38).

[188] Gillian M. Keating. «Bevacizumab: A review of its use in advanced cancer». In: *Drugs* 74.16 (Oct. 2014), pp. 1891–1925. ISSN: 11791950. DOI: 10.1007/s40265-014-0302-9 (cit. on p. 38).

[189] Thomas E. Witzig et al. «Treatment with ibritumomab tiuxetan radioimmunotherapy in patients with rituximab-refractory follicular non-Hodgkin's lymphoma». In: *Journal of Clinical Oncology* 20.15 (Aug. 2002), pp. 3262–3269. ISSN: 0732183X. DOI: 10.1200/JCO.2002.11.017 (cit. on p. 38).

[190] Jianming Zhang, Priscilla L Yang, and Nathanael S Gray. «Targeting cancer with small molecule kinase inhibitors.» In: *Nature reviews. Cancer* 9.1 (2009), pp. 28–39. ISSN: 1474-175X. DOI: 10.1038/nrc2559 (cit. on pp. 38, 39).

[191] I Bernard Weinstein and Andrew K Joe. «Mechanisms of disease: Oncogene addiction–a rationale for molecular targeting in cancer therapy.» In: *Nature clinical practice. Oncology* 3.8 (Aug. 2006), pp. 448–57. ISSN: 1743-4254. DOI: 10.1038/ncponc0558 (cit. on p. 39).

[192] Paolo A Ascierto et al. «The role of BRAF V600 mutation in melanoma». In: *Journal of Translational Medicine* 10 (July 2012), p. 85. ISSN: 1479-5876. DOI: 10.1186/1479-5876-10-85 (cit. on p. 39).

[193] Gideon Bollag et al. «Clinical efficacy of a RAF inhibitor needs broad target blockade in BRAF-mutant melanoma». In: *Nature* 467.7315 (Sept. 2010), pp. 596–599. ISSN: 0028-0836. DOI: 10.1038/nature09454 (cit. on p. 39).

[194] Geoffrey T Gibney and Jonathan S Zager. «Clinical development of dabrafenib in BRAF mutant melanoma and other malignancies». In: *Expert Opinion on Drug Metabolism & Toxicology* 9.7 (July 2013), pp. 893–899. ISSN: 1742-5255. DOI: 10.1517/17425255.2013.794220 (cit. on p. 39).

[195] Keith T Flaherty et al. «Combined BRAF and MEK Inhibition in Melanoma with BRAF V600 Mutations». In: *New England Journal of Medicine* 367.18 (Sept. 2012), pp. 1694–1703. ISSN: 0028-4793. DOI: 10.1056/NEJMoa1210093 (cit. on pp. 39, 48).

[196] Keith T Flaherty et al. «Improved Survival with MEK Inhibition in BRAF-Mutated Melanoma». In: *New England Journal of Medicine* 367.2 (June 2012), pp. 107–114. ISSN: 0028-4793. DOI: 10.1056/NEJMoa1203421 (cit. on p. 39).

[197] S. Percy Ivy, Jeannette Y. Wick, and Bennett M. Kaufman. «An overview of small-molecule inhibitors of VEGFR signaling.» In: *Nature reviews. Clinical oncology* 6.10 (2009), pp. 569–79. ISSN: 1759-4782. DOI: 10.1038/nrclinonc.2009.130 (cit. on p. 39).

[198] G Manning. «The Protein Kinase Complement of the Human Genome». In: *Science* 298.5600 (2002), pp. 1912–1934. ISSN: 00368075. DOI: 10.1126/science.1075762 (cit. on pp. 39, 40, 45).

[199] Susanne Müller et al. «The ins and outs of selective kinase inhibitor development.» In: *Nature chemical biology* 11.11 (Nov. 2015), pp. 818–21. ISSN: 1552-4469. DOI: 10.1038/nchembio.1938 (cit. on pp. 40, 44).

[200] Y. Liu et al. «A molecular gate which controls unnatural ATP analogue recognition by the tyrosine kinase v-Src». In: *Bioorganic & Medicinal Chemistry* 6.8 (1998), pp. 1219–1226. ISSN: 09680896. DOI: 10.1016/S0968-0896(98)00099-6 (cit. on p. 40).

[201] Martin E M Noble, Jane A Endicott, and Louise N Johnson. «Protein kinase inhibitors: insights into drug design from structure.» In: *Science (New York, N.Y.)* 303.5665 (Mar. 2004), pp. 1800–5. ISSN: 1095-9203. DOI: 10.1126/science.1095920 (cit. on p. 40).

[202] Kinase Inhibitor et al. «Exploration of Type II Binding Mode : A Privileged Approach for». In: (2016). DOI: 10.1021/cb500129t (cit. on p. 42).

[203] Cristiano R W Guimarães et al. «Understanding the Impact of the P-loop Conformation on Kinase Selectivity». In: *Journal of Chemical Information and Modeling* 51.6 (June 2011), pp. 1199–1204. ISSN: 1549-9596. DOI: 10.1021/ci200153c (cit. on p. 42).

[204] Erick J Morris et al. «Discovery of a Novel ERK Inhibitor with Activity in Models of Acquired Resistance to BRAF and MEK Inhibitors». In: *Cancer Discovery* 3.7 (July 2013), pp. 742–750. DOI: 10.1158/2159-8290.CD-13-0070 (cit. on p. 42).

[205] Michael S Cohen et al. «Structural bioinformatics-based design of selective, irreversible kinase inhibitors.» In: *Science (New York, N.Y.)* 308.5726 (May 2005), pp. 1318–21. ISSN: 1095-9203. DOI: 10.1126/science1108367 (cit. on p. 43).

[206] Michele H Potashman and Mark E Duggan. «Covalent Modifiers: An Orthogonal Approach to Drug Design». In: *Journal of Medicinal Chemistry* 52.5 (Mar. 2009), pp. 1231–1246. ISSN: 0022-2623. DOI: 10.1021/jm8008597 (cit. on p. 43).

[207] Qingsong Liu et al. «Developing irreversible inhibitors of the protein kinase cysteinome». In: *Chemistry & biology* 20.2 (Feb. 2013), pp. 146–159. ISSN: 1074-5521. DOI: 10.1016/j.chembiol.2012.12.006 (cit. on p. 43).

[208] Tjeerd Barf and Allard Kaptein. «Irreversible Protein Kinase Inhibitors: Balancing the Bene fi ts and Risks». In: *Journal of medicinal chemistry* 55 (2012), pp. 6243–6262. DOI: 10.1021/jm3003203 (cit. on p. 43).

[209] Daniel C Liebler. «Protein Damage by Reactive Electrophiles: Targets and Consequences». In: *Chemical Research in Toxicology* 21.1 (Jan. 2008), pp. 117–128. ISSN: 0893-228X. DOI: 10.1021/tx700235t (cit. on p. 43).

[210] Kiyoshi Okamoto et al. «Distinct Binding Mode of Multikinase Inhibitor Lenvatinib Revealed by Biochemical Characterization». In: *ACS Medicinal Chemistry Letters* 6.1 (Jan. 2015), pp. 89–94. ISSN: 1948-5875. DOI: 10.1021/ml500394m (cit. on p. 44).

[211] Mindy I Davis et al. «Comprehensive analysis of kinase inhibitor selectivity». In: *Nat Biotech* 29.11 (Nov. 2011), pp. 1046–1051. ISSN: 1087-0156. DOI: 10.1038/nbt.1990 (cit. on p. 45).

[212] Jianming Zhang, Priscilla L Yang, and Nathanael S Gray. «Targeting cancer with small molecule kinase inhibitors». In: *Nat Rev Cancer* 9.1 (Jan. 2009), pp. 28–39. ISSN: 1474-175X (cit. on p. 45).

[213] Caitriona Holohan et al. «Cancer drug resistance: an evolving paradigm.» In: *Nature reviews. Cancer* 13.10 (Oct. 2013), pp. 714–26. ISSN: 1474-1768. DOI: 10.1038/nrc3599 (cit. on pp. 46, 49).

[214] Michael M Gottesman. «Mechanisms of Cancer Drug Resistance». In: *Annual Review of Medicine* 53.1 (Feb. 2002), pp. 615–627. ISSN: 0066-4219. DOI: 10.1146/annurev.med.53.082901.103929 (cit. on p. 46).

[215] Scott W Lowe, Enrique Cepero, and Gerard Evan. «Intrinsic tumour suppression.» In: *Nature* 432.7015 (Nov. 2004), pp. 307–15. ISSN: 1476-4687. DOI: 10.1038/nature03098 (cit. on p. 46).

[216] Jeremy S. Logue and Deborah K. Morrison. «Complexity in the signaling network: Insights from the use of targeted inhibitors in cancer therapy». In: *Genes and Development* 26.7 (2012), pp. 641–650. ISSN: 08909369. DOI: 10.1101/gad.186965.112 (cit. on p. 46).

[217] Douglas W McMillin, Joseph M Negri, and Constantine S Mitsiades. «The role of tumour-stromal interactions in modifying drug response: challenges and opportunities». In: *Nat Rev Drug Discov* 12.3 (Mar. 2013), pp. 217–228. ISSN: 1474-1776. DOI: 10.1038/nrd3870 (cit. on p. 46).

[218] Sabine Maier et al. «Identifying DNA Methylation Biomarkers of Cancer Drug Response». In: *American Journal of Pharmacogenomics* 5.4 (2005), pp. 223–232. ISSN: 1175-2203. DOI: 10.2165/00129785-200505040-00003 (cit. on p. 46).

[219] Pasi Koivisto et al. «Androgen Receptor Gene Amplification: A Possible Molecular Mechanism for Androgen Deprivation Therapy Failure in Prostate Cancer». In: *Cancer Research* 57.2 (Jan. 1997), pp. 314–319 (cit. on p. 47).

[220] Michael W. Schmitt, Lawrence A. Loeb, and Jesse J. Salk. «The influence of subclonal resistance mutations on targeted cancer therapy.» In: *Nature reviews. Clinical oncology* (2015). ISSN: 1759-4782. DOI: 10.1038/nrclinonc.2015.175 (cit. on pp. 47, 48).

[221] Yi-fan Chen and Li-wu Fu. «Mechanisms of acquired resistance to tyrosine kinase inhibitors». In: *Acta Pharmaceutica Sinica B* 1.4 (Dec. 2011), pp. 197–207. ISSN: 2211-3835. DOI: http://dx.doi.org/10.1016/j.apsb.2011.10.007 (cit. on p. 47).

[222] Rina Barouch-Bentov and Karsten Sauer. «Mechanisms of Drug-Resistance in Kinases». In: *Expert opinion on investigational drugs* 20.2 (Feb. 2011), pp. 153–208. ISSN: 1354-3784. DOI: 10.1517/13543784.2011.546344 (cit. on p. 47).

[223] Mercedes E Gorre et al. «Clinical Resistance to STI-571 Cancer Therapy Caused by BCR-ABL Gene Mutation or Amplification». In: *Science* 293.5531 (Aug. 2001), pp. 876–880 (cit. on p. 47).

[224] Simona Soverini et al. «Implications of BCR-ABL1 kinase domain-mediated resistance in chronic myeloid leukemia». In: *Leukemia Research* 38.1 (July 2016), pp. 10–20. ISSN: 0145-2126. DOI: 10.1016/j.leukres.2013.09.011 (cit. on p. 47).

[225] Jin-Yuan Shih, Chien-Hung Gow, and Pan-Chyr Yang. «EGFR Mutation Conferring Primary Resistance to Gefitinib in Non–Small-Cell Lung Cancer». In: *New England Journal of Medicine* 353.2 (July 2005), pp. 207–208. ISSN: 0028-4793. DOI: 10.1056/NEJM200507143530217 (cit. on p. 47).

[226] Annette O Walter et al. «Discovery of a Mutant-Selective Covalent Inhibitor of EGFR that Overcomes T790M-Mediated Resistance in NSCLC». In: *Cancer Discovery* 3.12 (Dec. 2013), pp. 1404–1415. DOI: 10.1158/2159-8290.CD-13-0314 (cit. on p. 47).

[227] Darren A E Cross et al. «AZD9291, an irreversible EGFR TKI, overcomes T790M-mediated resistance to EGFR inhibitors in lung cancer». In: *Cancer discovery* 4.9 (Sept. 2014), pp. 1046–1061. ISSN: 2159-8274. DOI: 10.1158/2159-8290.CD-14-0337 (cit. on p. 47).

[228] Dalia Ercan et al. «EGFR Mutations and Resistance to Irreversible Pyrimidine-Based EGFR Inhibitors». In: *American Association for Cancer Research* 21.17 (Aug. 2015), pp. 3913–3923. DOI: 10.1158/1078-0432.CCR-14-2789 (cit. on p. 47).

[229] Robert C Doebele et al. «Mechanisms of Resistance to Crizotinib in Patients with ALK Gene Rearranged Non-Small Cell Lung Cancer». In: *Clinical cancer research : an official journal of the American Association for Cancer Research* 18.5 (Mar. 2012), pp. 1472–1482. ISSN: 1078-0432. DOI: 10.1158/1078-0432.CCR-11-2906 (cit. on p. 47).

[230] Alice T Shaw et al. «Resensitization to Crizotinib by the Lorlatinib ALK Resistance Mutation L1198F». In: *New England Journal of Medicine* 374.1 (Dec. 2015), pp. 54–61. ISSN: 0028-4793. DOI: 10.1056/NEJMoa1508887 (cit. on p. 47).

[231] Peng Wu, Thomas E. Nielsen, and Mads H. Clausen. «Small-molecule kinase inhibitors: An analysis of FDA-approved drugs». In: *Drug Discovery Today* 21.1 (2016), pp. 5–10. ISSN: 18785832. DOI: 10.1016/j.drudis.2015.07.008 (cit. on p. 48).

[232] Rebecca a. Burrell and Charles Swanton. «Tumour heterogeneity and the evolution of polyclonal drug resistance». In: *Molecular Oncology* 8.6 (2014), pp. 1095–1111. ISSN: 15747891. DOI: 10.1016/j.molonc.2014.06.005 (cit. on p. 48).

[233] Javier Cortes and Henri Roché. «Docetaxel combined with targeted therapies in metastatic breast cancer». In: *Cancer Treatment Reviews* 38.5 (July 2016), pp. 387–396. ISSN: 0305-7372. DOI: 10.1016/j.ctrv.2011.08.001 (cit. on p. 48).

[234] Matthew Vanneman and Glenn Dranoff. «Combining immunotherapy and targeted therapies in cancer treatment». In: *Nat Rev Cancer* 12.4 (Apr. 2012), pp. 237–251. ISSN: 1474-175X. DOI: 10.1038/nrc3237 (cit. on p. 48).

[235] Bissan Al-Lazikani, Udai Banerji, and Paul Workman. «Combinatorial drug therapy for cancer in the post-genomic era.» In: *Nature biotechnology* 30.7 (July 2012), pp. 679–92. ISSN: 1546-1696. DOI: 10.1038/nbt.2284 (cit. on p. 48).

[236] Ivana Bozic et al. «Evolutionary dynamics of cancer in response to targeted combination therapy.» In: *eLife* 2 (Jan. 2013), e00747. ISSN: 2050-084X. DOI: 10.7554/eLife.00747 (cit. on p. 48).

[237] Natalia L Komarova, Jan a Burger, and Dominik Wodarz. «Evolution of ibrutinib resistance in chronic lymphocytic leukemia (CLL).» In: *Proceedings of the National Academy of Sciences of the United States of America* 111.38 (2014), pp. 13906–11. ISSN: 1091-6490. DOI: 10.1073/pnas.1409362111 (cit. on p. 48).

[238] Marc J Williams et al. «Identification of neutral tumor evolution across cancer types». In: *Nature Genetics* August 2015 (2016). ISSN: 1061-4036. DOI: 10.1038/ng.3489 (cit. on p. 48).

[239] Camille Stephan-Otto Attolini et al. «A mathematical framework to determine the temporal sequence of somatic genetic events in cancer.» In: *Proceedings of the National Academy of Sciences of the United States of America* 107.41 (2010), pp. 17604–9. ISSN: 1091-6490. DOI: 10.1073/pnas.1009117107 (cit. on p. 48).

[240] J Chmielecki et al. «Optimization of dosing for EGFR-mutant non-small cell lung cancer with evolutionary cancer modeling». In: *Sci Transl Med* 3.90 (2011), 90ra59. ISSN: 1946-6242. DOI: 10.1126/scitranslmed.3002356 (cit. on p. 48).

[241] Paul H Huang et al. «Quantitative analysis of EGFRvIII cellular signaling networks reveals a combinatorial therapeutic strategy for glioblastoma». In: *Proceedings of the National Academy of Sciences* 104.31 (July 2007), pp. 12867–12872. DOI: 10.1073/pnas.0705158104 (cit. on p. 48).

[242] Feng Zhu et al. «Therapeutic target database update 2012: a resource for facilitating target-oriented drug discovery.» In: *Nucleic acids research* 40.Database issue (Jan. 2012), pp. D1128–36. ISSN: 1362-4962. DOI: 10.1093/nar/gkr797 (cit. on p. 49).

[243] E W Dijkstra. «A note on two problems in connexion with graphs». In: *Numerische Mathematik* 1.1 (1959), pp. 269–271. ISSN: 0945-3245. DOI: 10.1007/BF01386390 (cit. on p. 57).

[244] Xing Chen, Ming-Xi Liu, and Gui-Ying Yan. «Drug-target interaction prediction by random walk on the heterogeneous network.» In: *Molecular bioSystems* 8.7 (July 2012), pp. 1970–8. ISSN: 1742-2051. DOI: 10.1039/c2mb00002d (cit. on p. 57).

[245] Yu-Fen Huang, Hsiang-Yuan Yeh, and Von-Wun Soo. «Inferring drug-disease associations from integration of chemical, genomic and phenotype data using network propagation.» In: *BMC medical genomics* 6 Suppl 3.Suppl 3 (Jan. 2013), S4. ISSN: 1755-8794. DOI: 10.1186/1755-8794-6-S3-S4 (cit. on p. 57).

[246] J F Westphal, D Vetter, and J M Brogard. «Hepatic side-effects of antibiotics». In: *Journal of Antimicrobial Chemotherapy* 33.3 (Mar. 1994), pp. 387–401. DOI: 10.1093/jac/33.3.387 (cit. on p. 59).