Structural study of the therapeutic potential of protein-ligand interactions

Francisco Martínez-Jiménez

TESI DOCTORAL UPF / 2016

THESIS SUPERVISOR

Dr. Marc A. Marti-Renom

THESIS TUTOR

Dr. Baldo Oliva

Acknowledgments

It is very difficult to remember all the people to whom I owe gratitude. It is not because I am bad at memorizing names, which I am, it is because there were so many people contributing to this wonderful experience that I probably have forgotten some of you.
The first person that comes into my mind is my supervisor, Marc. Thank you for the scientific (and non-scientific) conversations, thank you for allowing me to attend to all those conferences, for the freedom I had to choose what I liked, thank you for your support, useful advice and your guidance; but yet more importantly, thank you for being my friend. Thank you, Marc, for making this thesis a wonderful experience.
I am also very grateful to all the people in the lab. I really enjoyed this experience with you guys. Thank you to Davide for helping me with the first steps in the lab, to Fransuuu for being so disastrous with names (we are not the exception but the rule!), to Yasmina for teaching Fransuuu how to properly dance, he is the master of the rhythm!; to Laia for showing us the single apple's diet, to Yannick for your good taste with wine and cheese, to Marco for patiently repeating the story of Sergio Roberto Alfredo, to David (*Castillo*) for surprisingly showing that people supporting Barça may have some knowledge about football, to Silvia for all the chewing gums I've stolen, to Gireesh for all non-scientific activities we have shared over the PhD, to Irene for teaching this beautiful language called Irenico, to Mike for being such a nice guy and more importantly, such a great cook!; to David (*Praga*) for our long and instructive conversations about our shared topic of interest, and finally, thank you to Pauli, for listening, helping and having such a big heart despite of being so incredibly small. THANK YOU ALL GUYS!.
Thank you also to the PRBB PhD community. Thank you to Lisa, Gireesh, Vero, Maria, Luis, Francesco, Jordi, Lorenzo, Nino, Marco... Life won't be the same without our insane PhD student retreats and parties. Special mention should be made to Juan (*el Gallego*), the best storyteller I have ever met! I'll miss you guys!
Thank you also to all the CNAG-CRG people who has helped me during this time. Thank you to Anna(s), Raul, Fernando, Steve, Santi, Angelika, Leo, Lukasz, Marie, Francesca, Miguel, Anibal... Your help, as significant as it might look, was very important to me.
Mención especial merece Giussele, tu fuiste el verdadero artífice y creador de SIGECO (con el beneplácito de Iñigo). Todos y cada uno de los integrantes de SIGECO sois responsables de que mi dia a dia sea mucho más ameno, de que tenga ganas de sentarme para leer el correo, y de que diga, orgulloso, que conservo 14 amigos (más satélites) desde la universidad. Gracias a Nobelisco

(como Giusele diría), a Eva, su futura novia; a Feliz, a los Excel de Feliz, a Soto, al hijo de Soto, al Isrraaa malo, a Juanin H.D.P., a Rafa, a los remos de Rafa, a Batxes, a Txula y a Argi; a Iñigo-GOD, a Markel, al solar de Markel, a Rocamora y su ejercito de esclavos, a Juan Lu, a Santiago Puigdemont, a Sebas y su amiga chilena, a Patrón, al dueño de la Forja, a Kaku, a mini Kakus, etc.
Agradecer tambien al resto de amigos, no necesariamente científicos, el hecho de haber compartido estos 4 años con vosotros. Gracias a la gente del pueblo (Lore, Gine, Peri, Belén, Mori, Rubio, Javi, Lorena, etc) por acogerme y tratarme tan bien cada vez que voy. Gracias a Carlos por soportar mis aventuras durante más de un año de convivencia. Gracias a Enric por compartir conmigo su infinita sabiduría y por su labor encomiable recogiendo balones perdidos en la jungla. A Juan Carlos por haber aguantado estóicamente durante las tres semana de viaje a Costa Rica. Gracias a Nuria por su bipolaridad. Gracias también a Iñigo, y a su familia, por haber sido parte de un viaje único e irrepetible. Y como no, un gracias infinito a mis misiles. Gracias Juanlu por ser el mejor compañero de piso, compañero de futbol, compañero de gimnasio, de fiesta en Plataforma, de fiesta en cualquier otro lugar del territorio español... Gracias por permitirme practicar mi catalán en la intimidad, por imitar tan bien al bueno de Josep, en fin, gracias por ser mi amigo. Un gracias enorme a mi *Weycito*, por haberme hecho vivir un año indescriptible, por las mil y una aventuras vividas juntos, por haberme hecho parte de tu familia (os quiero Gaby y Paty), por enseñarme que crear una amistad no requiere de mucho tiempo, te echo muchísimo de menos wey, tenemos que vernos pronto... Por último gracias a Vare, no hace falta que te diga mucho por qué, no? Gracias por estar ahí durante más de 14 años, gracias por conocerme tan bien, gracias por las miles de aventuras que hemos vivido, y sobre todo, gracias por las que nos quedan por vivir!
No creo en dios ni en el destino, pero estoy muy agradecido por haber conocido a Maria. Gracias, María, por todo tu apoyo y ayuda, por comprender mis errores y ensalzar mis virtudes, gracias por hacerme ver las cosas que para mi no son sencillas, gracias por haber compartido un viaje tan maravilloso como el que hicimos, gracias por reirte de mí y conmigo. En fin, gracias, María, por ser la persona tan maravillosa que eres.
Y por último lo más importante, gracias a mi familia. Gracias a mis tios y primos, a Juanita, a Alejandro y, sobretodo, a mis abuelos y a mi madre. Existen infinitas cosas por las que daros las gracias, pero sólo diré que gracias por todo lo que me habéis dado y por haberme educado para ser quien soy. Esta tesis es tan vuestra como mía.

Abstract

Most of the cellular functions are driven by small-molecules that selectively bind to their protein targets. Is such their importance, that the pharmacological intervention of proteins by small molecule drugs is frequently used 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 for predicting structurally detailed protein-ligand interactions at proteome scale. The method exemplified 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 suggested 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 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 todas las pequeñas moléculas 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 contra 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 sugiere 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 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

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**. Nucleic Acids Research. (*Submitted*).

2. **Martínez-Jiménez, F.***, & Marti-Renom, M. A. (2016). **Should network biology be used for drug discovery?** Expert Opinion on Drug Discovery, 1–3. http://doi.org/10.1080/17460441.2016.1236786

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" [1]. 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 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

[1] 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.

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 Subsection 1.1.6.

Protein structures are complex conformation of atoms organized in a hierarchical manner (Figure 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.
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 (Figure 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

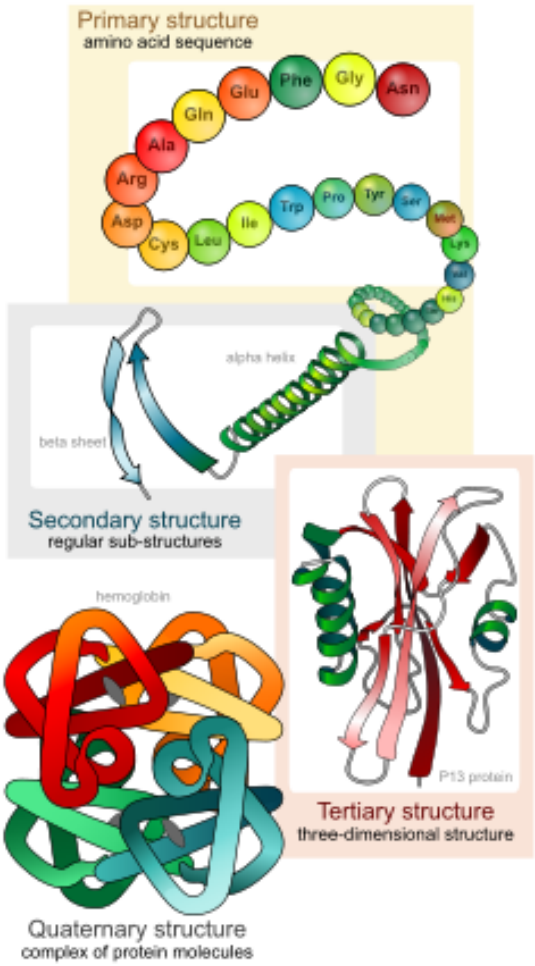


Figure 1.1: Hierarchical distribution of layers in protein structure. Image Credit: Mariana Ruiz Villarreal (`https://commons.wikimedia.org/wiki/File:Main_protein_structure_levels_en.svg`)

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.

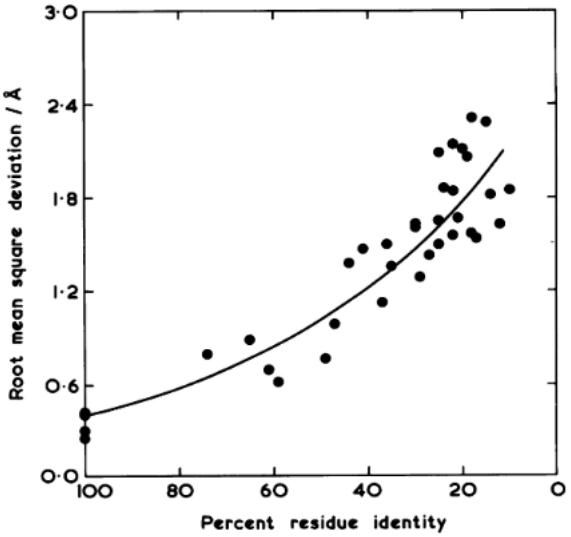


Figure 1.2: The original plot of the relation of residue identity and the RMSD deviation of the backbone atoms of the common cores of 32 pairs of homologous proteins. Figure extracted from [5]

1.1.2. Protein Structure Determination

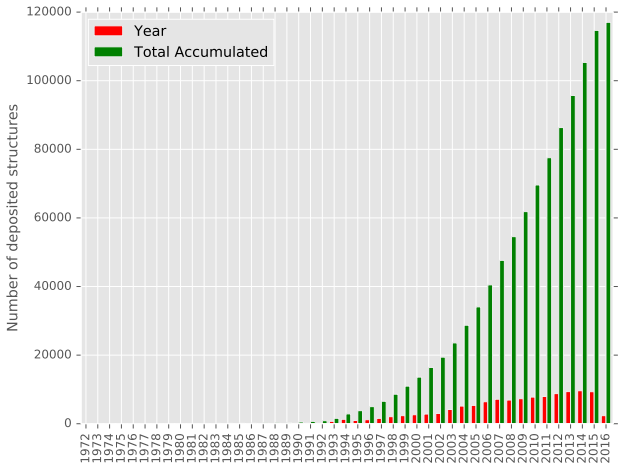
Since 1960, when the Brithis biochemist John Kendrew determined the myoglobin structure [9], more than 37,000 different protein structures have been deposited in the Protein Data Bank (PDB) [10]. The PDB is a repository created in the 1970s with the aim of storing all the 3D protein structures and unifying 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 tem-

plates, 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.

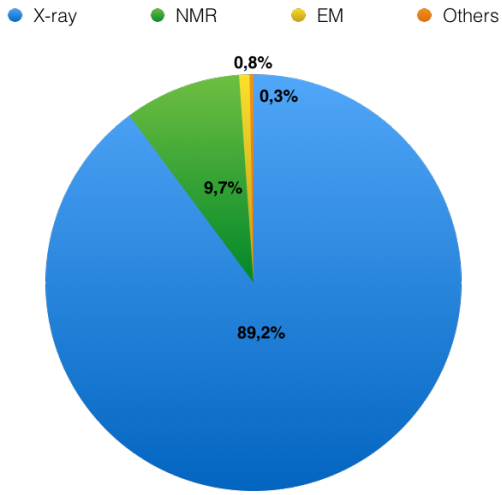
Several methods are currently used to experimentally determine the 3D structure of a protein. More than 99% of structures deposited in the PDB have been determined by the three main methods: X-ray crystallography (X-ray), nuclear magnetic resonance spectroscopy (NMR) and electron microscopy (EM) (Figure 1.3b). These methods provide experimental data that helps 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-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 several 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].



(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.

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 has traditionally been applied to large molecules such as ribosomes [19] or the V-ATPase [20], but it has also shown 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 supervisorthe 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].

1.1.4. Homology modeling

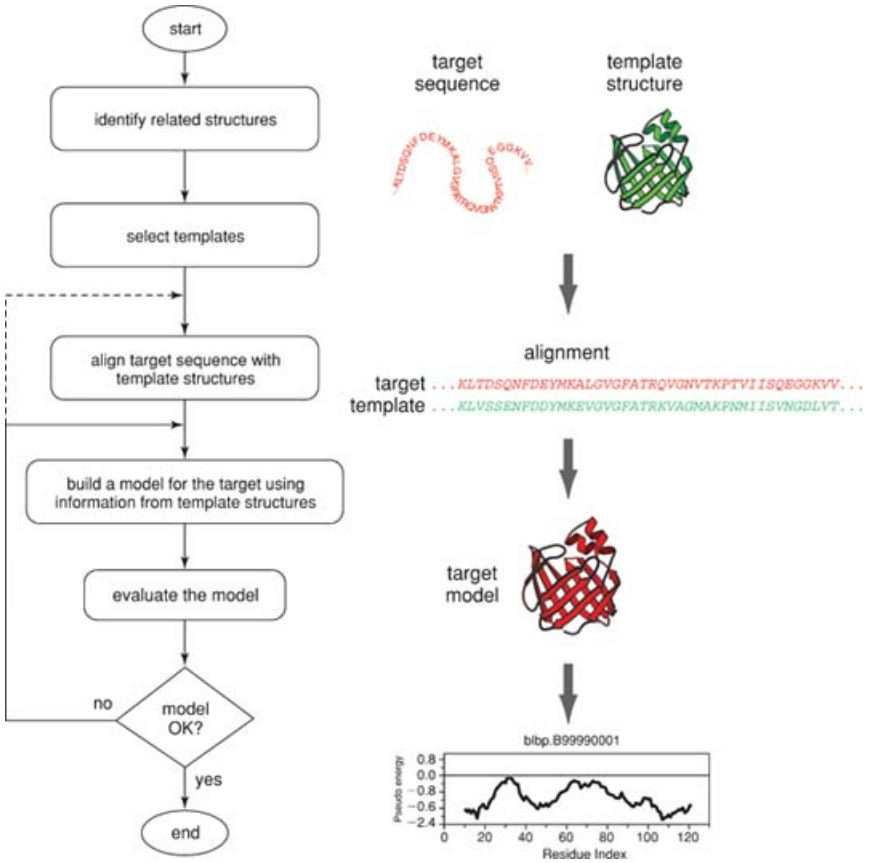


Figure 1.4: Workflow in comparative protein structure modeling. The figure has been extracted from [28]

This type of protein structure prediction method 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 consists 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 lacking 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 [2] 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. **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 identifica-

[2]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).

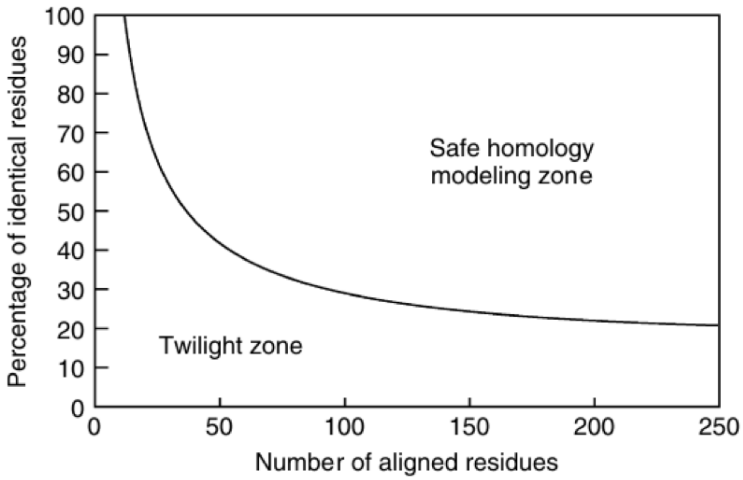


Figure 1.5: Homology threshold curve as a function of alignment length. Data extracted from [31]

tion 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].

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 their 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 restricting 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[3].

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 outperform the others for all cases.

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

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

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.

protein 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 range 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 contain 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* [4]. The next section is focused on protein-compound interaction presenting the basis for all the work developed during the thesis.

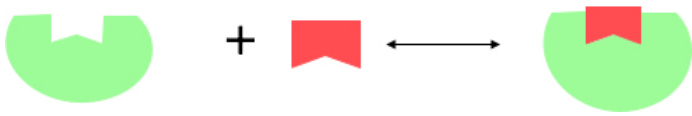
[4]For simplicity, in this thesis, unless otherwise indicated, the term ligand will only refer to small molecules ligands, while proteins ligands will be explicitly named as protein-protein interactions.

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 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].

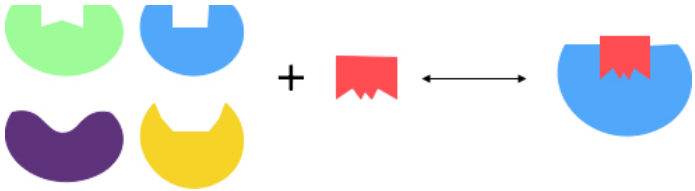
1.1.7. Protein-ligand binding energetics

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



(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.

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 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 pre-clinical 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[5]

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 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. Virtual docking methods 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. Several entropic, enthalpic and environmental factors have

[5]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.

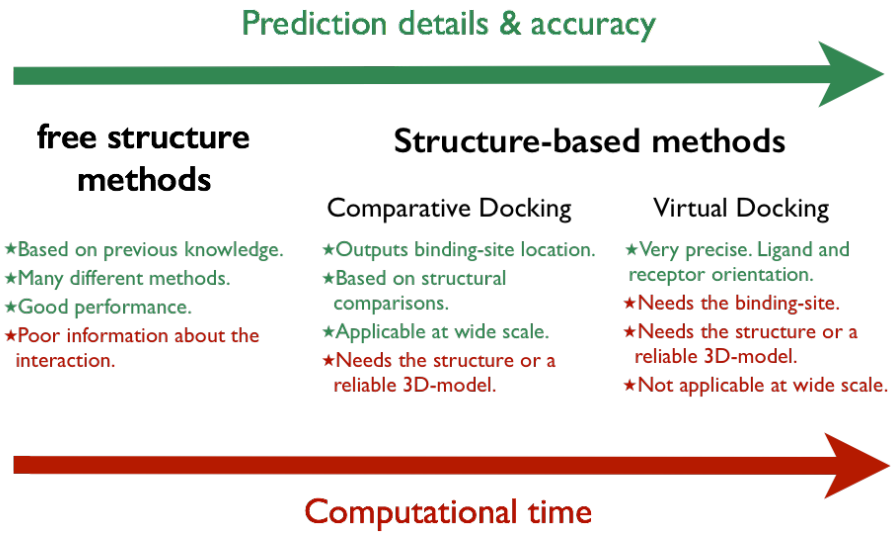


Figure 1.7: Classification of the methods for ligand-target interaction prediction alongside their advantages and disadvantages. The red arrow represent the increase in the computational time of the calculus needed for each prediction. The green arrow represents the level of detail of the given output.

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] [6]. 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 simplistic 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, ad-

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

vantages, 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 whose 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 druggability*. 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)[7] 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 efficiency, 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.

[7] 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.*

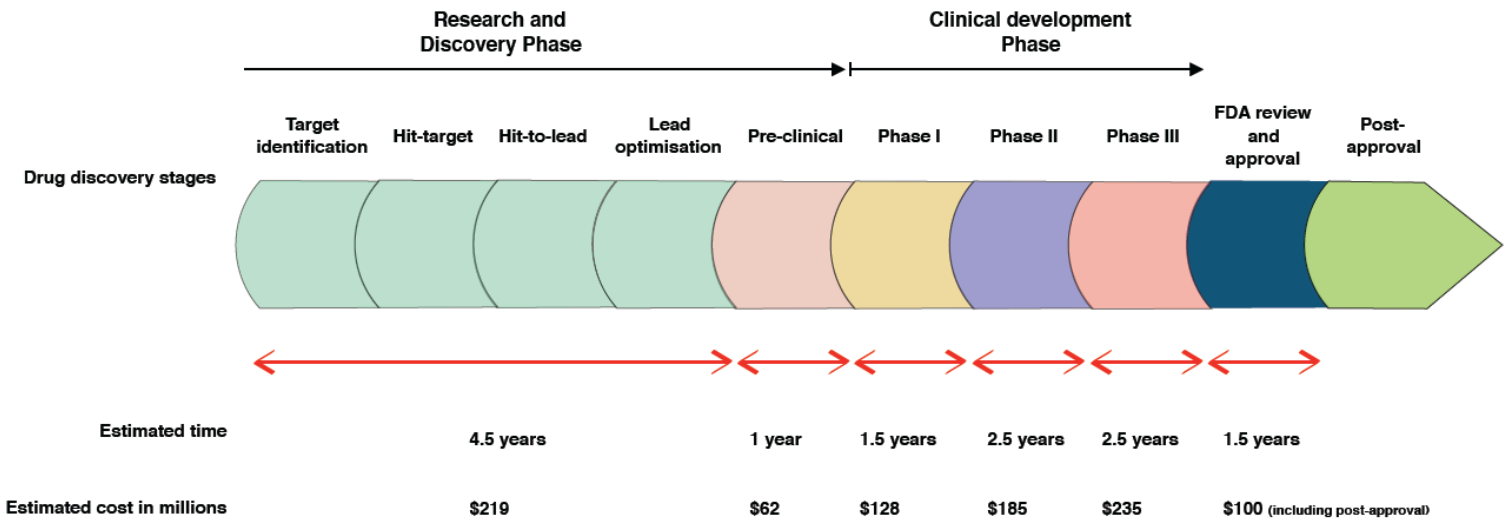


Figure 1.8: Drug discovery and development pipeline. For each stage the average cost and time are provided. Post-approval times not included in the time-line. Data extracted from [125].

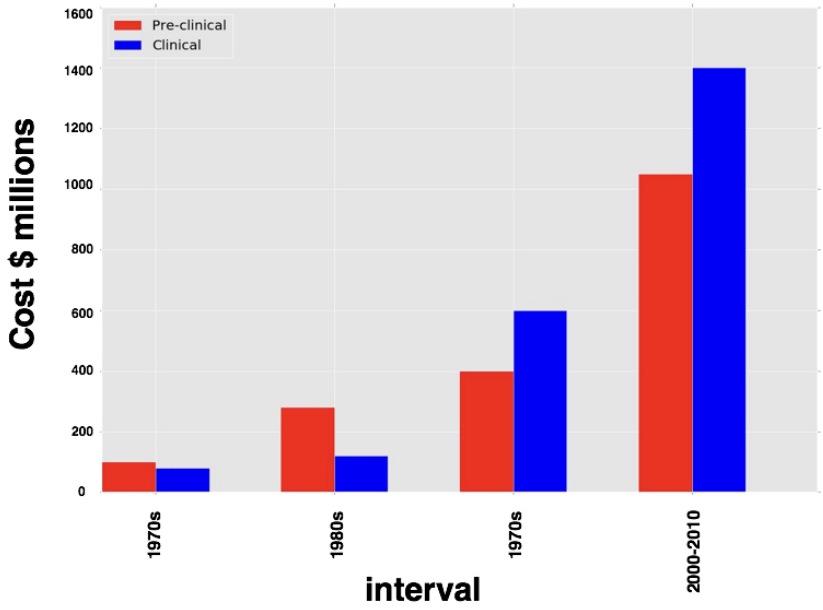


Figure 1.9: Cost of developing a new drug. Blue bars indicate expenses in clinical phases while red represents expenses in pre-clinical stages. Costs are shown in $ millions. Data extracted form: Tufts Center for the Study of Drug Development (`http://csdd.tufts.edu/news/complete_story/pr_tufts_csdd_2014_cost_study`).

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 assisting 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 Next-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 has become very popular [129, 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 been applied to 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 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 to 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 Subsection 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 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, indeep 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.

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

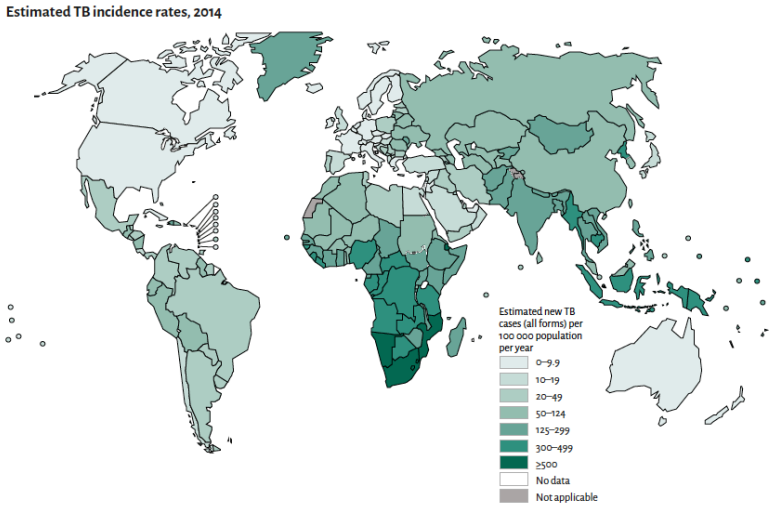


Figure 1.10: Estimated worldwide TB incidence rates in 2014. Figure extracted from [153].

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 challenging 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 an efficient 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 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, Subsection 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 protein 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 antitubercular 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. Drug discovery in cancer

Over the previous sections, we discussed how tuberculosis, among other infectious diseases, has significantly benefited from the application of *in-silico* methods. The importance of computational methods is also observed in other diseases research and discovery programs. Interestingly, bioinformatics-assisted diseases include cancer. Cancer research has significantly improved due to development of large-scale genomics techniques alongside computational methods to deal with the massive amount of generated data. The next section discusses about advances in cancer treatment, focusing on targeted cancer therapy, a particular type of cancer treatment. The emergence of targeted cancer therapies significantly changed the landscape of cancer treatment. Unlike classic chemotherapy agents, targeted therapies perform their activity by specifically attacking proteins involved in the growth, progression, and spread of cancer. However, clinical benefits associated to targeted therapies are often temporal due to the emergence of drug resistance. Next sections will also discuss about this problem, explaining the molecular mechanisms leading to the emergence of drug resistance. Finally, Subsection 3.3 presents a computational model aiming to assist in the choice of non-resistant targeted cancer therapies.

1.5. 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 nor-

mal 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 efficient 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.

1.5.1. 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 becoming 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]. Tumor 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 tumors 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 have 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.

1.5.2. 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 mainly 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 tumor 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 kinase targets are those highly expressed in the tumor stroma and that are required for different stages of tumor 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 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 mimic the ATP binding mode. However, depending of their specific binding mode small molecule protein kinase inhibitors can be classified into multiple classes:

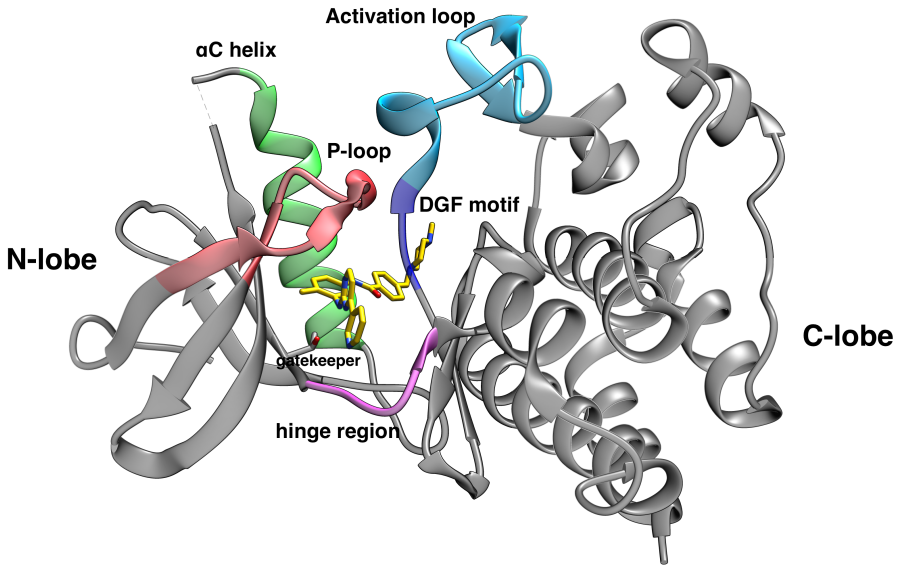


Figure 1.11: 3D structure of ABL1 kinase in complex with imatinib displaying the different structural regions of protein kinases. The structure represents the typical kinase inactive DGF-out conformation. The protein is represented as ribbons and the ligand as sticks. The activation loop is coloured 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.

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 I 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 kinase (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 (Figure 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.

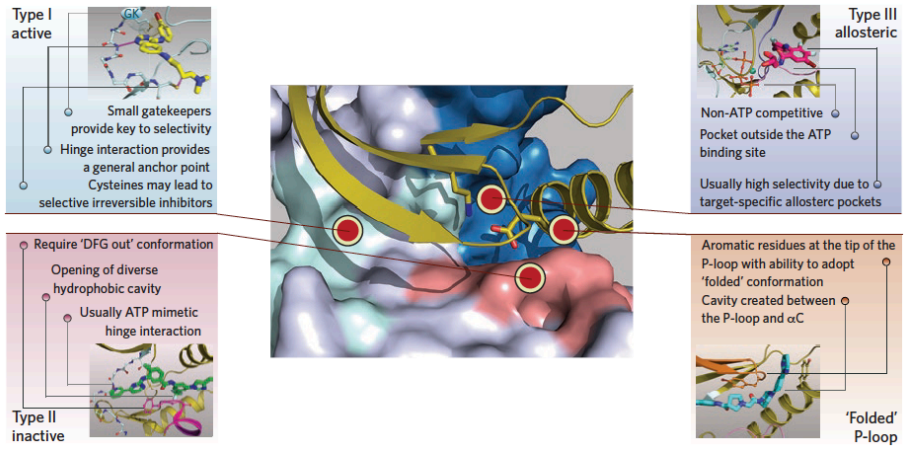


Figure 1.12: Structural features of the canonical classes of small molecule kinase inhibitors. The center panel shows the main interaction sites of different inhibitors types. The side panels show the specific structural features of each of the binding modes. Figure extracted from [199].

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.

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.

1.5.3. 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 underlying 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 reaching the tumor. 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 tumor, 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 tumor 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 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 tumor proliferation and/or development (Subsection 1.5.2). Mutations of these oncogenic kinases can alter the binding of the small molecule kinase inhibitor giving rise to a reestablishment of the tumor proliferation activity. Moreover, evidence continues to emerge that cancers are characterized by extensive intratumor 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 1.5.2). 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 has 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 1.5.2). 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 responsible 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 im-

munotherapy with targeted therapies [234, 235]. 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 [236]. In fact, computational methods are being applied to explain and predict therapeutic resistance [237, 238], tumor clonal evolution [239, 240] and potential drug combinations [241, 242]. 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.

1.6. 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 1.5). 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, 243], which enables computational models to increase their predictive power by learning from known 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 1.5). 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 1.5.2). However, small molecule targeted cancer therapies suffer from a major limitation, the clinical benefit of patients receiving this therapies is often temporal (Subsection 1.5.3). 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 tumor mutational landscape with the potential to confer resistance and 2) providing alternative treatments in those cases where the drug-resistant mutants are already present in the tumor 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 Subsection 3.1 . 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 `http://nannolyze.cnag.cat/`.

Finally, points iv) and v) are presented in Subsection 1.5.3. Particularly, 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)

suggests alternative compounds for those mutations predicted to confer resistance to a given targeted cancer therapy.

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/`.

Manuscript 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

PLOS | COMPUTATIONAL BIOLOGY

RESEARCH ARTICLE

Ligand-Target Prediction by Structural Network Biology Using nAnnoLyze

Francisco Martínez-Jiménez[1,2], **Marc A. Marti-Renom**[1,2,3]*

1 Genome Biology Group, Centre Nacional d'Aanàlisi Genòmica (CNAG), Barcelona, Spain, **2** Gene Regulation, Stem Cells and Cancer Program, Centre for Genomic Regulation (CRG), Barcelona, Spain, **3** Institució Catalana de Recerca i Estudis Avançats (ICREA), Barcelona, Spain

* mmarti@pcb.ub.cat

OPEN ACCESS

Citation: Martínez-Jiménez F, Marti-Renom MA (2015) Ligand-Target Prediction by Structural Network Biology Using nAnnoLyze. PLoS Comput Biol 11(3): e1004157. doi:10.1371/journal.pcbi.1004157

Editor: Avner Schlessinger, Icahn School of Medicine at Mount Sinai, UNITED STATES

Received: August 26, 2014

Accepted: January 27, 2015

Published: March 27, 2015

Data Availability Statement: All the structural data and all the predictions can be downloaded from the server at http://www.marciuslab.org/services/nAnnoLyze.

Funding: This work was supported by the Spanish MINECO (BFU2010-19310, BFU2013-47736-P and PIM2010EPA-00719 to MAMR.). The funders had no role in study design, data collection and analysis, decision to publish, or preparation of the manuscript.

Competing Interests: The authors have declared that no competing interests exist.

Abstract

Target identification is essential for drug design, drug-drug interaction prediction, dosage adjustment and side effect anticipation. Specifically, the knowledge of structural details is essential for understanding the mode of action of a compound on a target protein. Here, we present nAnnoLyze, a method for target identification that relies on the hypothesis that structurally similar binding sites bind similar ligands. nAnnoLyze integrates structural information into a bipartite network of interactions and similarities to predict structurally detailed compound-protein interactions at proteome scale. The method was benchmarked on a dataset of 6,282 pairs of known interacting ligand-target pairs reaching a 0.96 of area under the Receiver Operating Characteristic curve (AUC) when using the drug names as an input feature for the classifier, and a 0.70 of AUC for "anonymous" compounds or compounds not present in the training set. nAnnoLyze resulted in higher accuracies than its predecessor, AnnoLyze. We applied the method to predict interactions for all the compounds in the DrugBank database with each human protein structure and provide examples of target identification for known drugs against human diseases. The accuracy and applicability of our method to any compound indicate that a comparative docking approach such as nAnnoLyze enables large-scale annotation and analysis of compound–protein interactions and thus may benefit drug development.

Author Summary

Description of the "mode-of-action" of a small chemical compound against a protein target is essential for the drug discovery process. Such description relies on three main steps: i) the identification of the target protein within the thousands of proteins in an organism, ii) the localization of the binding interaction site in the identified target protein, and iii) the molecular characterization of the compound's binding mode in the binding site of the target protein. Here, we introduce a new computational method, called nAnnoLyze, which uses graph theory principles to relate compounds and target proteins based on comparative principles. nAnnoLyze aims at addressing two of the three previous steps, that is, target identification and binding site localization. Our results suggest that the nAnnoLyze

accuracy and proteome-wide applicability enables the large-scale annotation and analysis of compound–protein interaction and thus may benefit drug development.

Introduction

The number of newly approved drugs has been significantly decreasing over the last two decades [1]. To make things worse, the therapeutic dogma that has prevailed over the years aimed at single target-specific 'magic bullets' against each disease. However, proteins act in complex interconnected networks, and thus, this 'one gene, one drug, one disease' paradigm is now clearly challenged [2,3]. The polypharmacology concept, which relies on the fact that a drug can modulate its activity by interacting with multiple targets rather than just one, was proposed to address these limitations [2]. Polypharmacology is especially valid in complex diseases like cancer or central nervous system disorders where the modulation of the activity of one single protein is not sufficient to obtain a therapeutic effect [4–6]. Therefore, identification of all possible targets of a chemical compound is critical in the drug discovery process.

Many *in silico* methods have been published for drug target identification using network approaches [7,8]. Broadly, we can distinguish two different classes of methods, structure-free methods and structure-based methods. Within the first group, there are methods based on ligand features [9] that have been successfully used to identify numerous experimentally validated interactions. However, they have difficulties in identifying interactions for drugs with novel scaffolds [10] or for targets with no bioactivity information. Others, named network-based approaches, exploit network properties to provide the drug target interactions and drug repositioning opportunities [11–18]. Although the accuracy of predictions by these methods has significantly increased, the majority cannot explain the mode of action of the drug over the predicted target due to the lack of three-dimensional (3D) information about the ligand and/or the target. The use of 3D structural data helps addressing such limitation. The most popular structured-based methods rely on molecular docking approaches performing a virtual screening of a compound against a limited number of protein targets or of several compounds against one protein target [19–21]. As a result, they provide structurally detailed information about the likely interaction between the compound and its target/s. However, the computational requirements of such approaches make them not generally applicable at proteomic scales. An exception to this limitation is the recent massive human screening of 600,000 drugs against 7,000 human protein pockets by Cardozo and colleagues whose results are available online [22]. To overcome the computational limitations, new structure-based methods use the so-called "comparative docking" approaches that solely rely on structural comparisons, both of compounds and protein targets, to infer new interactions [23,24]. Other methods use local structural comparisons of small molecule binding sites to infer the localization and specificity of binding pockets [25,26] as well as to infer new ligand interactions in known binding pockets [27]. Finally, several other methods that rely on 3D structure comparisons that aim at functionally annotating structures [23,24,28].

Here we introduce nAnnoLyze, a network-based version of the comparative docking method AnnoLyze [23]. Our new method predicts interactions for any query compound against an entire 3D proteome by relying on a bi-partite network of interactions and similarities. Unlike Annolyze, nAnnolyze can predict interactions for any compound regardless if they have been previously co-crystallized with a protein. We have benchmarked nAnnoLyze against a dataset composed by all the interactions for approved drugs present in the Protein Data Bank (PDB) [29]. The method outperforms AnnoLyze precision by 27 folds. Both Annolyze and nAnnolyze

have been already successfully applied. Annolyze was used in an open source drug discovery initiative against neglected tropical diseases [30] while nAnnoLyze has been applied to a set of anti-tubercular drugs against the *Mycobacterium tuberculosis* proteome [31]. Here, we describe the method alongside the predictions for all the small molecule drugs present in DrugBank [32] against the human 3D proteome. To our knowledge, this is the first screening of almost 6,000 drugs against the entire human structure proteome predicted by comparative approaches. The nAnnoLyze network, method and predictions are available online at http://www.marciuslab.org/services/nAnnoLyze.

Results

Benchmark dataset creation

The correct selection of a benchmark dataset is one of the most important steps in assessing the accuracy of a newly developed method. Unfortunately, there were no available and adequate datasets for benchmarking structure-based network methods for ligand-target prediction. The "Yamanishi-2008" dataset [11], which has widely been used previously, could not be used here due to the limited structural coverage of its targets, which added to the increasing concern on biases of the current drug-target interaction datasets [33]. To address these issues, we have generated a benchmark set consisting of a "positive" and a "negative" set. The "positive" set contains all drug-protein annotated pairs between any structure in the PDB and any compound approved by the FDA. The "positive" benchmark set resulted in a total of 6,282 interactions and is considered the "true" set of interactions. The "negative" set was generated by randomly selecting pairs of compounds and targets that have never been annotated in the DrugBank or PDB databases. To assess how many of these drug-protein negative pairs could result as a potentially miss-annotated negative interactions we looked for similar compounds interacting with the "negative" target of each compound. The search resulted in 118 ($\sim$2%) out of the 5,981 pairs that could result in a miss-annotated negative interaction. However, the removal of these pairs of putative miss-annotated "negative pairs" from the set had no effect on the assessment of the nAnnoLyze accuracy. Our final benchmark dataset included thus a total of 6,282 drug-target in the "positive" interactions and 5,981 negative pairs.

nAnnoLyze benchmarking

The nAnnoLyze precision varies at different Z-score thresholds (Fig. 1A) with an optimal threshold at −2.5 local Z-score resulting in a precision of 0.63 and coverage of 0.19 corresponding to 1,148 true positive predictions (Fig. 1B). It is important to note that both the precision and coverage of our method depend dramatically on the definition of false positives for our predictions. Given that our benchmark set relies only on deposited data in the PDB, many of the predictions by nAnnoLyze are likely to be correct despite not being present in our benchmark. For example, the drug Enalapril (DB00584 DrugBank identifier) has been co-solved in only two PDB entries (*i.e.*, 2X90 and 1UZE). However, nAnnoLyze predicts interactions between Enalapril and three other targets in the PDB (*i.e.*, 2X91, 1J36 and 2X8Z). Those structures actually correspond to the same target sequence (Q10714 UniProt id) being solved with no ligands.

To further increase the accuracy of our predictions, we implemented a Random Forest Classifier (RFC) that classifies pairs of compound-protein as binders or not by combining several of the nAnnoLyze scores (that is, the raw score, the Local Z-score, and the Global Z-score). The RFC correctly recalled 66% of the pairs with a precision of 0.73 and an AUC of 0.71 using a 10-fold cross validation (Table 1). The tested RFC did not include the DrugBank ID as input feature to simulate a situation where a completely new compound not deposited in the

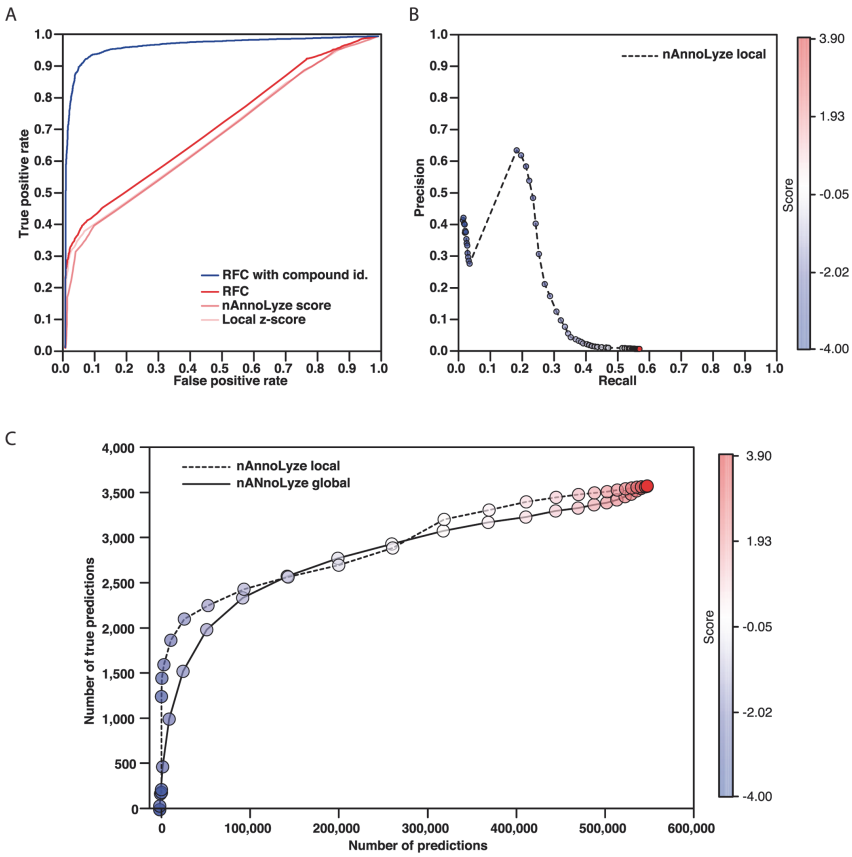


Fig 1. nAnnoLyze benchmarking. A) ROC plots for predictions in the benchmark dataset using 10-fold cross-validation. Blue line for predictions based on our RFC trained using the compound ID, red line for predictions based on our RFC trained with anonymous compounds. Light red lines correspond to the predictions based on individual nAnnoLyze scores. B) Precision/Recall curves for nAnnoLyze local Z-score. C) Enrichment plots for nAnnoLyze local Z-score (dashed black line) nAnnoLyze global Z-Score (solid black line) predictions.

doi:10.1371/journal.pcbi.1004157.g001

databases was tested. However, by using the DrugBank ID as an input feature, the accuracy of nAnnoLyze dramatically improves to a 0.93 precision, 0.93 recall and a 0.97 of AUC (Fig. 1A and Table 1). These results suggest that predictions for known drugs already in our dataset are

Table 1. RFC benchmark.

Type of classification	Precision	Recall	AUC
RFC (DrugID, SCORE, Global Z-score, Local Z-Score)	0.93±0.01	0.93±0.01	0.97±0.01
RFC (SCORE, Global Z-score, Local Z-Score)	0.73±0.01	0.66±0.01	0.71±0.01
Score	0.70±0.02	0.64±0.02	0.68±0.02
Global Z-score	0.70±0.02	0.64±0.02	0.68±0.02
Local Z-score	0.73±0.02	0.64±0.02	0.67±0.01

Mean values and standard deviation after 10-fold cross-validation.

doi:10.1371/journal.pcbi.1004157.t001

much more precise than those for unknown or anonymous compounds. The RFC outperformed the use of any of the single scores from nAnnoLyze (Table 1).

Comparatively, nAnnoLyze reached a 0.61 increase in precision at the optimal cut-off (from 0.02 to 0.63) at the expenses of a decrease in recall by 0.38 with respect to AnnoLyze (Table 2). Finally, It is important to note that the benchmark set used for this test resulted more difficult for AnnoLyze than the original test-set used to benchmark it [23] (Table 2).

nAnnoLyze prediction examples

The human Cyclooxygenase-1 is targeted by NSAID drugs. Cyclooxygenase (COX) is the enzyme responsible for the formation of prostanoids, which are classified in 3 different groups: prostaglandins, prostacyclins, and thromboxanes, each of them is involved in the inflammatory response, among other processes. There are two COX isoenzymes. COX-1 promotes the production of the natural mucus that protects the inner stomach lining while COX-2, is primarily present at sites of inflammation [34]. Traditional non-steroidal anti-inflammatory drugs (NSAIDs) such as Aspirin, Ibuprofen or Flurbiprofen are considered non-selective because they inhibit both COX-1 and COX-2. The inhibition of COX-2 by NSAIDs results in the anti-inflammatory effect, while the inhibition of COX-1 can lead the undesired side effects such as damage to the gastrointestinal tract [35]. nAnnoLyze predicted interactions for several NSAIDs with the 3D model of the human COX-1. Specifically, nAnnoLyze predicted 21 (out of the 44 approved drugs against COX-1) as binders of the COX-1 target (Table 3). In particular, nAnnoLyze predicted the binding of Flurbiprofen (DB00712) and Ibuprofen (DB01050) to COX-1, which are known inhibitory drugs of the human COX-1 (Fig. 2A). The nAnnoLyze path between Flubiprofen and COX-1 starts from a ligand node composed by tripotassium (1R)-4-biphenyl-4-yl-1-phosphonatobutane-1-sulfonate (B70) and two stereoisomers of Flubiprofen (FLR and FLP). Thorough the binding site of FLP to ovine COX-1 (1QEH), nAnnoLyze predicts its binding site of the COX-1 human 3D model. Conversely, the path between Ibuprofen and COX-1 starts in the ligand node composed by 1-(4-ethylphenyl)propan-1-one (I3E) and two stereoisomers of Ibuprofen (IBP and IZP). Those ligands are predicted to bind the same predicted binding site of the human COX-1 thanks to its similarity to the crystal structure of

Table 2. nAnnoLyze benchmark.

		optimal cut-off (max value)
nAnnoLyze	Precision	0.63 (1.00)
	Recall	0.19 (0.59)
AnnoLyze	Precision	0.02 (0.06)
	Recall	0.57 (0.67)

doi:10.1371/journal.pcbi.1004157.t002

Table 3. COX-1 interactions.

Drug ID	Drug name	nAnnoLyze score
DB00712	Flurbiprofen	0.97
DB00328	Indomethacin	0.97
DB01600	Tiaprofenicacid	0.96
DB00870	Suprofen	0.96
DB00821	Carprofen	0.96
DB00788	Naproxen	0.96
DB00500	Tolmetin	0.94
DB00465	Ketorolac	0.94
DB00963	Bromfenac	0.92
DB00586	Diclofenac	0.91
DB06802	Nepafenac	0.90
DB01283	Lumiracoxib	0.90
DB00784	Mefenamicacid	0.89
DB00861	Diflunisal	0.88
DB04552	NiflumicAcid	0.88
DB00991	Oxaprozin	0.88
DB01050	Ibuprofen	0.87
DB00939	Meclofenamicacid	0.86
DB01399	Salsalate	0.86
DB01009	Ketoprofen	0.86
DB00605	Sulindac	0.85

doi:10.1371/journal.pcbi.1004157.t003

the ovine COX-1 (1EQG). Remarkably, the human COX-1 predicted binding site includes the tyrosine 385, which is known to be responsible of the catalytic reaction with the NSAID drugs (Fig. 2A). However, not all the NSAIDs performed with the same accuracy. Aspirin (DB00945), also a known inhibitor of the human COX-1 and COX-2, results in false positive predictions (Table 4 and Fig. 2B). The nAnnoLyze search with Aspirin as input molecule results in many proteases predicted targets. This false-positive pathway starts from the ligand node composed by two Benzoic Acids, the 4-Guanidinobenzoic Acid (GBS) and the Acetylsalicylic acid (AIN). GBS has been crystallized with different trypsin proteins so the pathway goes thorough the GSB binding site of the guanidinobenzoyl-trypsin acyl-enzyme (2AH4) reaching eventually the predicted binding site for the human Trypsin-2 (P07478). The same pathway is used to find other proteases like the Airway trypsin-like protease 4 (Q6ZWK6) or the Trypsin-3 (P35030) resulting in several false positive predictions. Conversely, the Aspirin-COX1 network pathway starts from the ligand node composed by 3,6-dichloro-2-methoxy-benzoic acid (D3M) and Salicylic acid (SAL) (Fig. 2B). The RFC classifier identified a network link between Aspirin and the SAL compound with a similarity score of 0.86. This SAL mediated pathway guides the nAnnoLyze search towards its binding site in the ovine COX-1 (3N8Y), which is homologous to the human COX-1 binding site. This pathway is also the responsible of the link between Aspirin and the human COX-2 with a score of 0.77. However, the lower similarity between the predicted human COX-2 binding site and the ovine COX-1 (3N8Y) introduces a penalty that significantly decreases the score of the link.

Sorafenib pathway targeting through binding of several proteins. Sorafenib, which is marketed as Nexavar, is an approved drug for the treatment of advanced renal cell carcinoma. It is also in Phase III trials for Hepatocellular carcinoma, Non-small-cell lung carcinoma (NSCLC) and melanoma and in Phase II trials for Myelodysplastic syndrome, Acute Myeloid Leukemia

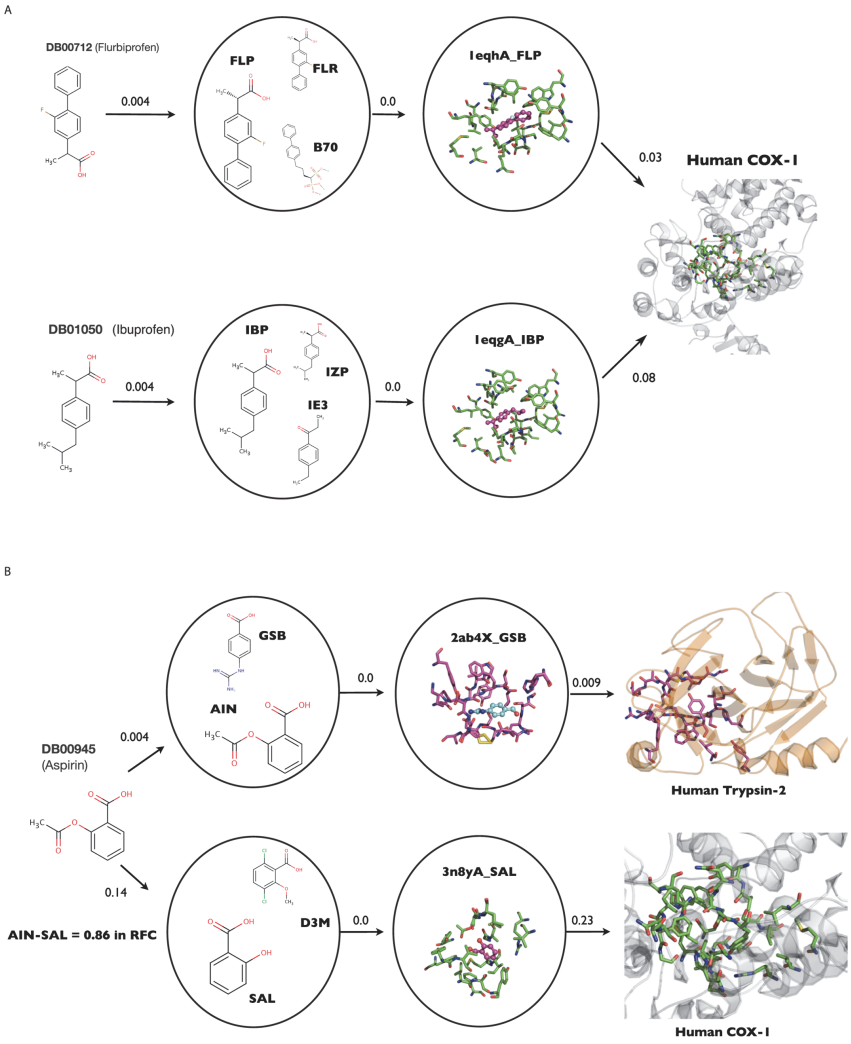

A
DB00712 (Flurbiprofen)
0.004
FLP
FLR
B70
0.0
1eqhA_FLP
0.03
Human COX-1
DB01050 (Ibuprofen)
0.004
IBP
IZP
IE3
0.0
1eqgA_IBP
0.08
B
DB00945
(Aspirin)
0.004
GSB
AIN
0.0
2ab4X_GSB
0.009
Human Trypsin-2
0.14
AIN-SAL = 0.86 in RFC
D3M
SAL
0.0
3n8yA_SAL
0.23
Human COX-1

Fig 2. nAnnoLyze network pathways. A) Network pathways for the predicted interactions between Flurbiprofen and Ibuprofen with the generated three-dimensional model of the Human Cox-1 (PDB Template: 2AYL with 94% sequence identity). B) Aspirin network pathway for miss predicted Aspirin-Trypsin2 (PDB Template: 3P95 with 98% Sequence Identity) and correctly predicted Aspirin-COX1 hit found by the method. Link between Aspirin and SAL was made by the RFC classifier with a score of 0.86 (Tanimoto value of 0.76). In both panels, the arrows represent edges with their weights representing the distance (i.e: the inverse of the similarity). The higher the distance value the lower the similarity between the compounds or binding sites. Ligand network nodes are encircled with the ligand responsible of the predicted pathway represented in larger size. For clarity, only one binding site per node has been plotted.

doi:10.1371/journal.pcbi.1004157.g002

(AML), head and neck, breast, colon, ovarian and pancreatic cancers. Arising as one of the most promising anticancer drugs, Nexavar is known to perform its activity by targeting the Raf/Mek/Erk pathways [36,37]. Specifically it is known to inhibit Raf kinases, Receptor-type tyrosine-protein kinase (FLT3), platelet-derived growth factor (PDGF), Vascular endothelial growth factor receptor 2 & 3 (VEGF2/3) and the Mast/stem cell growth factor receptor Kit. Within our predictions, we found 4 of these links alongside other interesting links for targets involved in the same pathways (Table 5 and Fig. 3A). Interestingly, most of the links have been previously annotated either in DrugBank, PubChem or in the PDB as a crystal structure. However, there are two links not annotated within the predictions, the serine/threonine-protein kinase A-Raf (ARAF) and the Cyclin-dependent kinase 10 (CDK10). ARAF is involved in several pathways, including AML and FoxO signaling and together with FLT3, BRAF, MAPK14 could be a good opportunity to exploit the polyphamarcological profile of Sorafenib against AML. In fact, Phase II trials are showing very promising results in AML combining Sorafenib with other marketed drugs [38,39]. Of the ten predicted targets, only 3 have been co-crystallized with Sorafenib (BRAF, MAPK14 and CDK8), while in the other seven nAnnoLyze proposes the binding site localization of the drug providing insights into the mode of action of the compound. nAnnoLyze predicted the correct binding site for the three targets (Fig. 3B). The predicted binding sites were 75%, 62%, and 86% correct (*i.e.*, % of predicted residues defined as binding site in LigBase) for CDK8, BRAF, and MAPK14, respectively.

Since structurally similar binding sites are more likely to bind the same small molecule. We wanted to assess if the 7 predicted binding sites (*i.e.*, FLT3, CDK10, ARAF, MAPK15, FLT1, RAF1, and CDK19) have similarity with the 3 Sorafenib known binding sites (*i.e.*, BRAF, MAPK14, and CDK8). All of the 7 predicted binding sites are similar to at least one of the already known (Fig. 3C). Within the annotated interactions with non-crystallized structure, FLT3 is the one with lowest similarity to a known structure (ProBiS Z-score of 1.04 with the CDK8 binding site). Unlike FLT3, FLT1 binding site has MAPK14 as the most similar binding

Table 4. Aspirin top 10 predicted targets as well as COX-1 and COX-2 scores.

UniProt ID	nAnnoLyze Score	Protein Name
P07478	0.94	Trypsin-2
Q6ZWK6	0.93	Airway trypsin-like protease 4
E7ESG9	0.93	Transmembrane protease serine 4
A6NL71	0.92	Transmembrane protease serine 11E
Q8IXD7	0.92	Kallikrein-11
Q0WXX5	0.92	Kallikrein 11 isoform 1
A8CED1	0.92	Trypsin-3
O60235	0.92	Transmembrane protease serine 11D protease
A8CED3	0.92	Protease serine 3 (mesotrypsin) isoform CRA_c
A9Z1Y4	0.92	Protease serine 3
Q5T7T7	0.81	COX-1
A8K802	0.77	COX-2

doi:10.1371/journal.pcbi.1004157.t004

Table 5. Sorafenib targets.

Target	Prediction Score	Annotated	Target Structure	KEGG OV Pathways
MAPK 14 (Q13083)	0.99	PDB	Yes	MAPK signaling pathway
		PubChem		FoxO signaling pathway
				VEGF signaling pathway
				Rap1 signaling pathway
				RIG-I-like receptor signaling pathway
				Acute myeloid leukemia
CDK19 (Q9BWU1)	0.97	PubChem	-	-
FLT1 (P17948)	0.90	PubChem	Yes	Ras signaling pathway
				Rap1 signaling pathway
RAF1 (P04049)	0.89	DrugBank	Yes	MAPK signaling pathway
		PubChem		Ras signaling pathway
				Rap1 signaling pathway
				VEGF signaling pathway
				FoxO signaling pathway
				Acute myeloid leukemia
ARAF (Q5H9B3)	0.88	-	Yes (partially)	FoxO signaling pathway
				Acute myeloid leukemia
CDK10 (Q15131)	0.88	-	-	-
BRAF (Q9Y6T3)	0.88	DrugBank	Yes	MAPK signaling pathway
		PDB		Rap1 signaling pathway
		PubChem		FoxO signaling pathway
				Acute myeloid leukemia
CDK8 (P49336)	0.87	Pubchem	Yes	-
		PDB		
FLT3 (Q5VTU6)	0.86	PubChem	Yes	Acute myeloid leukemia
		DrugBank		
MAPK 15 (Q8TD08)	0.86	Pubchem	-	-

doi:10.1371/journal.pcbi.1004157.t005

site with a higher score (2.25 ProBiS Z-score). Regarding the Cyclin dependent kinases CDK10 and CDK19 proposed binding sites, CDK10 binding site has a high similarity (ProBiS Z-score of 2.09) with the MAPK14´s one while the CDK19 binding site is almost identical to that of CDK8 (ProBiS 2.9). As expected, RAF predicted protein binding sites ARAF and RAF1 have BRAF binding site as the most similar (3.5 and 3.94 ProBiS Z-scores, respectively). Following the same trend, the MAPK14 binding site is the most similar to MAPK15 (2.51 ProBiS Z-score). Although small changes in the catalytic site could have a dramatic impact on the binding-affinity of a small molecule, the overall high similarity among the Sorafenib predicted binding sites shows a clear trend towards binding site conservation within this set of proteins. This example shows not only the capability of the method to find drug targets but also the possibility to explore pathways rather than individual proteins as targets.

Discussion

The increase of compound phenotypic screenings over the last years has dramatically increased the number of small molecules with non-annotated protein targets [40–42]. Because target annotation is a crucial step when developing a drug, and specifically the elucidation of the amino acids involved in the interactions is key to understand the mode of action of the compound, many methods have been developed to annotate drug protein targets. However, most of them

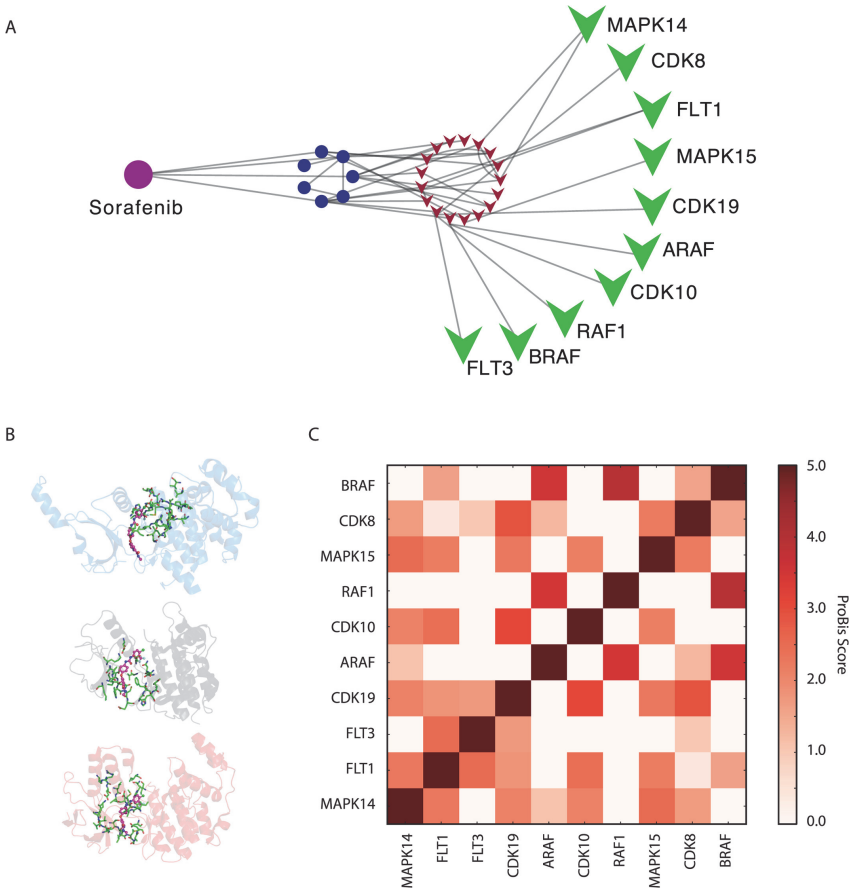


Fig 3. nAnnoLyze multiple-target example. A) Extraction of the Sorafenib sub network. B) Drug and the predicted binding site in CDK8 (PDB 3RGF, blue), BRAF (PDB 1UWH, grey) and MAPK14 (PDB 3GCS, red). C) ProBiS comparison of the binding site of predicted targets for DrugBank molecule Sorafenib (DB00398).

doi:10.1371/journal.pcbi.1004157.g003

do not provide any structural information about the link, and for those providing it, the application at proteome scale for any query compound is unfeasible. Here we introduced nAnnoLyze a method for drug target interaction prediction that provides structural details at

proteome scale. nAnnoLyze relies on a pre-built network of structural similarities to perform its prediction for any query molecule providing not only the connection between the molecule and its predicted target but also the binding site of the ligand in the protein. It is important to note that nAnnoLyze has been specifically tested for drug-target interaction prediction. The accuracy of our method on less studied compounds, such as non-drug like molecules, could lead to a reduction of the precision and the coverage.

The lack of crystal structure for several proteins in other datasets prompted us to build a new dataset of approved drugs. The reduction of the precision by our previous method [23] with this dataset is indicative of the complexity of the new benchmark. The new dataset includes real set of interactions that better simulates a scenario where the different molecules have different affinities to one or many targets. This addressed a current concern about the possible bias of artificial datasets [33]. Unfortunately, the lack of a real "negative" set of drug-protein pairs (*i.e.*, pairs of molecules known not to interact) hampered the creation of the complete dataset. To overcome this issue, we generated a set of drug-protein pairs that, so far, are not annotated as interactions. The nAnnoLyze benchmark using these newly created datasets resulted in satisfactory accuracies, especially in light of the fact that the dataset is bound to produce an overestimation of the false-positive rate (*i.e.*, a drug and a protein are not interacting if they have not been crystallized together) [43]. The limitation of the maximum distance in the search for the shortest pathway can explain some of the missed drug-protein pairs and, consequently, limits the recall reached by the method. Analysis of the precision and recall of specific compounds in the benchmark dataset indicate that nAnnoLyze results in higher accuracy for moderate promiscuous compounds compared to highly promiscuous compounds. Indeed, promiscuous compounds have high-degrees of connectivity in our network, which makes it very difficult to identify specific targets. A similar analysis to identify trends in the accuracy of nAnnoLyze for targets for different protein Pfam families did not result in any clear trend. The usage of binding site to represent a family of targets instead of whole protein domain structures may explain the homogeneity in the performance for different protein families.

Several scores for each prediction permits to explore the effect of the selection of different thresholds values depending of the user needs. For instance, when extracting only the most confident targets for a drug, very low values of Global Z-score will be suitable; while when retrieving the most specific targets for a compound filtering by low values of Local Z-score will be the best option. This, of course, makes it difficult to provide a specific score threshold for the predictions. Despite this, we studied the variation of the performance at different thresholds measured by a ROC curve. The AUC was excellent when using drug names and scores as input feature for the predictions. When only the scores of the predictions were used (that is, treating the compound as anonymous), there was a clear decrease in the AUC suggesting that the method performs better for already known chemical entities rather than for new unseen compounds. This fact makes sense since the method is based upon comparative approaches relating compounds by their structural similarities.

The comparison of the nAnnoLyze method against the original AnnoLyze indicates that our network-based approach predicts drug-protein complexes with higher precision. Importantly, nAnnolyze is a clear progress over Annolyze by improving not only the performance (27-fold higher precision) but also the applicability, since it can be applied to any compound regardless whether it has been previously deposited in the PDB. Moreover, the network-based paradigm implemented in nAnnoLyze allows for the integration of other types of additional information such as the diseases linked to the protein targets, which may eventually allow for drug indication predictions. A successful example of a method for predicting drug-like targets using the modelable human proteome with medical data integration is the Computational Analysis of Novel Drug Opportunities (CANDO) platform [43]. While the aim of our work is

accurately predicting drug-protein interactions, future developments of nAnnoLyze could include medical indications of drugs.

To demonstrate the applicability of the method, we screened all the drugs in the DrugBank database against the entire human 3D proteome that could be modeled by comparative protein structure prediction. We not only provided the drug-protein predictions but also the structural binding localization of the interaction. We carefully described two examples of this screening. The first example illustrates the nAnnoLyze ability to correctly (or incorrectly) predict the binding of a NSAIDs set of drugs to the COX-1 human protein. Within the correctly predicted interactions (*i.e.*, true positives), we included Flurbiprofen and Ibuprofen detailed information about the network routes. In the case of the incorrectly predicted interaction between Aspirin and proteases proteins, the analysis indicates that the clustering in a ligand node of two similar Benzenoids compounds lead to the undesired drug-target association. It is thus likely that adding extra information beyond the chemical similarity during the clustering of the core-network may result in more functionally homogeneous clusters of compounds. Even though, nAnnoLyze was able to reach the two main targets of aspirin (*i.e.*, COX-1 and COX-2) through alternatives network pathways. However, the lower similarity of the human predicted COX-2 binding site with the ovine COX-1 included in the core network penalized the score of the hit. This example also illustrates the nAnnoLyze capacity of predicting interactions when no crystal structure is available for the target.

The second example studied the polypharmacological profile of the anticancer drug Sorafenib. The method correctly retrieved most of the known targets and proposed some others with structural similarities in the binding site and that are involved in the same metabolic pathways as the known ones. This example shows the possibility of studying pathways rather than individual proteins as drug targets, which could be even more interesting in complex diseases such as cancer or Alzheimer where multiple factors play a role in the progress of the disease.

The major limitation of the method is the restricted applicability because is based on structural data, which is still scarce compared to sequence data. In spite of it, we were able to cover 42% of the human proteome with either a crystal structure or a reliable model. Moreover, the amount of crystal structures in the PDB has significantly increased over the past years [44] and the percentage of a proteome that can be modeled by homology has increased thanks to initiatives like the Protein Structure Initiative [45,46]. The more structural information we have, the more information can be extracted and therefore applied in nAnnoLyze. Indeed, the underlying network in nAnnoLyze can continue growing with the integration of new molecules or sets of biomolecules (both compounds and protein targets). To this end, we have developed a Web server that allows everyone to submit their own sets of compounds and check the predictions against pre-built networks for the human and *Mycobacterium* proteomes. So far, we have applied the method in an open source drug discovery initiative against *Mycobacterium tuberculosis* [31] and are currently working in other projects and initiatives. Our goal is to encourage open source drug discovery by releasing the method with all the predictions expecting that other researchers can benefit from our work. Finally, the scientific community could experimentally validate the predictions providing us a feedback to improve the quality of this tool and of future ones.

Materials and Methods

Next, we describe the different steps (Fig. 4A) performed to build a bi-partite network of structural similarities and interactions (Fig. 4B). We continue by describing the methods used to assess the accuracy of nAnnoLyze.

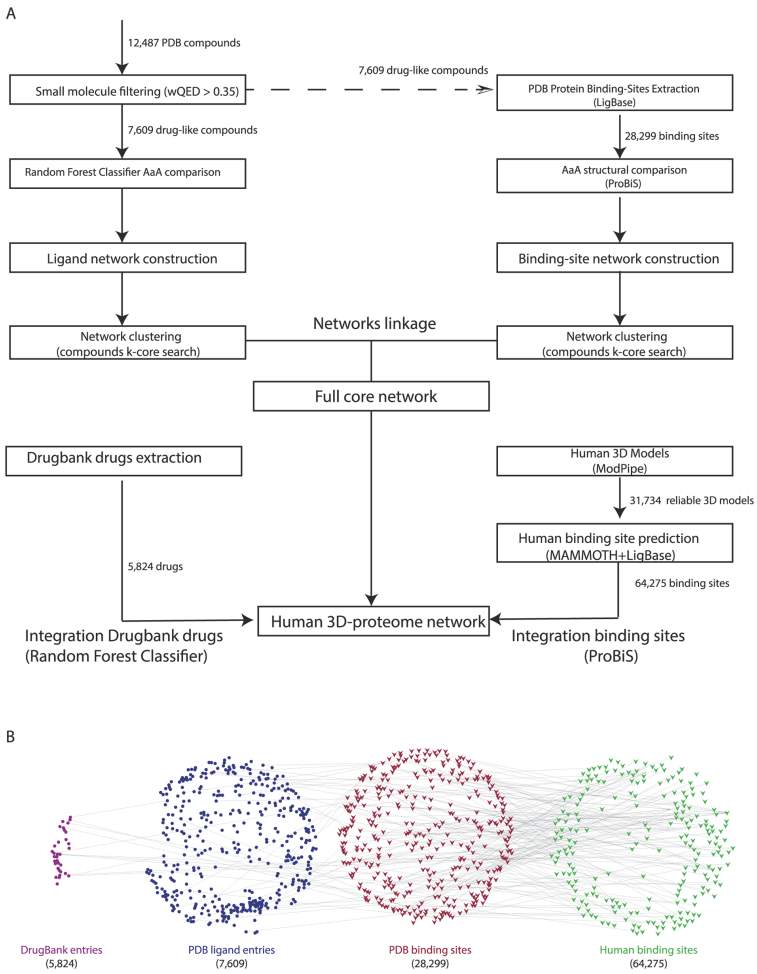


Fig 4. nAnnoLyze network building. A) nAnnoLyze flow chart for building the network of structural similarities between ligands and targets. B) nAnnoLyze underlying sub-networks of drugs (purple circles), compounds in PDB (blue circles), targets in PDB (red triangles), and human target 3D models (green triangles). For easy visualization, the panel shows only part of the final nAnnoLyze network.

doi:10.1371/journal.pcbi.1004157.g004

Ligand sub-network

To build the ligand sub-network, only compounds with a pharmaceutical or a biological function on their co-crystallized proteins were retrieved from the PDB. To perform the filtering, we calculated, for each compound in the PDB, the weighted quantitative estimate of drug-likeness (wQED). Briefly, the wQED is calculated by combining a set of the chemical features of the compound (*i.e.*, molecular weight, octanol-water partition coefficient as LogP, polar surface area, number hydrogen bond donors, number of hydrogen bond acceptors, number of rotatable bonds, number of aromatic rings, and number of possible toxic scaffolds) to quantify its drug-likeness given the pre-calculated value for that chemical features in a gold standard set of drugs [47]. Compounds with good drug-like properties (*i.e.*, wQED $\geq$0.35) were selected resulting in 7,609 PDB compounds. Each selected compound was then represented as a vertex in the ligand network. Links between vertices of the network (*i.e.*, edges) were obtained by structurally comparing all compounds. The weights of the edges were obtained using a Random Forest Classifier (RFC) developed to identify compound similarities [31]. Briefly, the RFC classifier predicts whether two small molecules are likely to bind the same target-binding site by comparing their structural and chemical properties. The usage of a classifier allows for an automatically determination of optimal thresholds after the RFC has been trained with the training-set. Therefore, the all-against-all comparison performed by the RFC resulted in 134,493 pairs of similar compounds. To reduce redundancy in the network we created groups of connected compounds by identifying k-cores in the network. A k-core in a network N, is a maximal connected sub-graph of N in which all vertices have degree at least k. Thus, every k-core in the non-redundant network represents a vertex and edges between vertices indicate the existence of at least one similar compound between the two k-cores. In the ligand network, a k-core would be a set of ligands such every two ligands within the set are similar to each other (*i.e.*, they have an edge in the network). An edge between two k-cores vertices was given the maximum weight of all possible edges between their constitutive compounds. The resulting non-redundant ligand sub-network had 4,101 vertices connected by 24,856 edges.

Protein binding sites sub-network

We first downloaded from the LigBase database (February 19th, 2013) [48], a database containing all ligand-binding sites of known protein structures, all unique protein binding sites composed of at least seven residues within a radius of 5 Å, binding any of the selected 7,609 highly drug-like compounds in the ligand sub-network. We defined "highly drug-like" compounds as those compounds with very good absorption, distribution, metabolism, and excretion properties (*i.e.*, with an wQED $\geq$0.35). This initial protein binding site sub-network resulted in 28,299 binding sites from 22,959 different proteins in the PDB. Next, we populated the network with links (edges) between two proteins by structurally comparing their binding sites. The structural comparison of the binding sites was performed using ProBiS [49], a tool for local structural alignment of binding sites based on geometry as well as physicochemical properties. We defined two binding sites as similar if their similarity Z-score is higher than 2.0. An all-against-all structural comparison of the selected binding sites was performed resulting in 579,155 pairs of similar binding sites. Next, we removed redundancy from the sub-network by applying a similar filtering that is used for the ligand sub-network. The final non-redundant sub-network for binding sites contained 19,487 vertices and 29,811 edges.

Final bi-partite network

Finally, we joined the two sub-networks by creating edges between protein binding sites and ligands. A binding site was linked to a ligand if both have been experimentally observed to

interact (*i.e.*, a solved structure with the target and the ligand exists in the PDB). The two sub-networks were linked by 22,832 edges and the final nAnnoLyze bi-partite network contained 23,588 vertices and 54,667 edges.

Integration of the human structural proteome

To populate the nAnnoLyze network with structures for human targets, we downloaded all human 3D models deposited in ModBase (November 11th, 2013) [50–52] with at least a 1.1 ModPipe Protein quality score [53]. ModBase is a database of comparative protein structure models calculated by the automatic modeling pipeline ModPipe [53]. The likely accuracy of the ModPipe models is predicted by the ModPipe Protein Quality score defined as a composite score that includes sequence identity to the template, coverage, and the three individual scores: the alignment e-value, z-dope [54], and GA341 [55]. This resulted in a total of 31,734 reliable 3D models from 16,694 unique human target sequences. Next, we structurally compared this set of selected models to any non-redundant (90% sequence identity) set of 29,772 structures from the PDB solved with at least one ligand compound. Structural comparisons between two proteins were performed using the MAMMOTH algorithm, which is based on a fast and accurate heuristic method to find, in a sequence-independent mode, the maximal structural subset between two proteins structures [56]. Four different scores were stored for each structural superposition: percentage of sequence and structure identity for the entire protein and percentage of sequence and structure identity for the residues involved in the binding site of the known structure as defined by LigBase. The structure identity between two structures was defined as the percentage of residues with their Cα atoms within 4 Å after optimal superposition. A binding site in a model was considered then similar to a binding site in a known PDB structure if at least the binding site sequence and structure identity were higher than 40%. This identity cut-off was previously validated in a large-scale comparison of known ligand-protein pairs [23]. A total of 576,675 binding sites were predicted for the human proteins (that is, $\sim$18 binding sites per model). Due to the high redundancy in the predicted binding sites, we excluded binding sites fulfilling the following requisites: redundant binding sites (*i.e.*, more than 80% sequence identity to any other binding site) or small binding sites (*i.e.*, with less than 6 residues). A total of 64,275 binding sites ($\sim$2 binding sites per model) remained after the redundancy and size filtering. Next, we compared all human predicted binding sites against all binding sites in our network using ProBiS resulting in 459,356 similarity links (Z-score $>$ 1.0) between any of the human 64,275 binding sites and the 28,299 binding sites in the network. Every significant pair became an edge with a weight equal to the normalized Z-score of the comparison. The final human network included the 7,609 compounds, the 28,299 known binding sites and the 64,275 human predicted binding sites.

Integration of the DrugBank compounds

A total of 6,540 small compounds were downloaded from the DrugBank database (May 15th, 2013). We then looked for similarity with the compounds present in the PDB ligand sub-network by using our trained RFC classifier as described above. Next, all the drugs were integrated in our network by making an edge between every DrugBank compound and their similar PDB compounds retaining the link with higher RFC when more than one link between a DrugBank compound and one network vertex (*i.e.*, a k-core of PDB compounds) was found. A total of 5,824 drugs were integrated into the network through 149,538 edges.

Network-based prediction of DrugBank ligand and human target pairs

Once the network was completed, to predict all possible interactions between DrugBank compounds and any of the modeled targets of the human proteome, we simply calculated the shortest path in the network from every queried DrugBank compound to any human binding sites. We implemented a version of the Dijkstra algorithm that limits the maximum reachable distance in order to speed up the computational time of the search [57]. Each hit was then scored by using the inverse of the sum of all edge weights of the path between the compound and the human target. Such score was then normalized and Z-scored. Specifically, two different Z-scores were calculated for each prediction.

$$Gz = \frac{s - \mu_G}{\sigma_G}$$

The "Global Z-score" (Gz) is obtained by running the predictions of all drugs present in DrugBank against all targets, obtaining a global mean (μ_G) and a global standard deviation (σ_G) to Z-score a specific predicted pair. The "Global Z-score" represents how good is a prediction given its score in the constructed network.

$$Lz = \frac{s - \mu_L}{\sigma_L}$$

The "Local Z-score" (Lz), is similarly calculated by running the predictions of all drugs present in DrugBank retrieving the mean (μ_L) and the standard deviation (σ_L) of the score for a specific target. The "Local Z-score" represents how good is a prediction for a specific binding site or target. For example, highly promiscuous binding sites tend to have higher local Z-scores.

Finally, we combined the three scores (that is, the inverse of the sum of all edge weights, the global Z-score and the local Z-score through a Random Forest Classifier that aims at predicting the interaction of a compound and a target. Two RFCs were trained with and without the DrugBank ID as an input feature of the compound. The RFC classifier, thus, results in a single Boolean score indicating interaction or non-interaction between the compound and the target. To train the RFC, we used the Weka software for data mining tasks [58].

nAnnoLyze benchmark

To benchmark nAnnoLyze, we retrieved all the compound-protein complexes for DrugBank approved drugs from the PDB. A total of 213 approved drugs were uniquely mapped into compounds bound to a protein deposited in the PDB. Next, we retrieved all the proteins binding to those compounds resulting in a protein-compound set of 6,282 entries. To test the method, we first created the benchmark network: the 213 compounds were integrated in the clustered network by using the RFC classifier. To avoid overestimation in the benchmark, we did not create any edge between a ligand in the benchmark and any identical (*i.e.*, RFC score of 1.0) ligand in the network. Next, we extracted from LigBase the 7,074 protein binding sites of the 213 aforementioned compounds and integrated them in the network following the procedure used for the human binding sites. Similarly, we did not create links between identical binding sites in the benchmark and any protein in the network. We then selected all interactions between the 213 compounds and any of the 7,074 binding sites. To assess the accuracy of our method in finding real interactions, we then calculated two different statistics. First, the precision defined as the ratio between the true positives (TP; true drug-protein interactions found by nAnnoLyze) and the sum of TP and false positives (FP, a link between a drug and a protein not in the PDB). Second, the sensitivity (or recall) defined as the ratio of TP and the TP+ false negatives (FN, a link between a compound and protein not found by nAnnoLyze).

nAnnoLyze Web site implementation

We have implemented a Web server where an end user can retrieve all pre-calculated predictions for the DrugBank and human protein as well as submit its own set of compounds. The server takes as input a compound ID and its SMILE in case of a new compound or only the DrugBank ID in case of a DrugBank drug. Then the user needs to select which organism proteome should be searched against. Currently nAnnoLyze has pre-calculated networks for the human and three *Mycobacterium* proteomes. The server search results in a list of all the predicted compound-protein pairs presented as a sortable table for easy filtering depending on the Global Z-score cut-off. A graphical enrichment of the Gene Ontology Terms [59] and KEGG pathways [60] of the predicted targets is also shown above the result table. Each prediction is further detailed by providing a GLMol based visualization (http://webglmol.sourceforge.jp) of the compound and the protein structure alongside the predicted binding site. All the structural data and all the predictions can be downloaded from the nAnnoLyze Web server at http://www.marciuslab.org/services/nAnnoLyze.

Acknowledgments

We thank Stefania Bosi and Prof. Anna Tramontano for their valuable input. Finally, we thank the Sali Lab for maintaining up-to-date their ModBase database.

Author Contributions

Conceived and designed the experiments: FMJ MAMR. Performed the experiments: FMJ. Analyzed the data: FMJ. Contributed reagents/materials/analysis tools: FMJ MAMR. Wrote the paper: FMJ MAMR.

References

1. Hopkins AL, Mason JS, Overington JP (2006) Can we rationally design promiscuous drugs? Curr Opin Struct Biol 16: 127–136. PMID: 16442279
2. Hopkins AL (2008) Network pharmacology: the next paradigm in drug discovery. Nat Chem Biol 4: 682–690. doi: 10.1038/nchembio.118 PMID: 18936753
3. Narayan VA, Mohwinckel M, Pisano G, Yang M, Manji HK (2013) Beyond magic bullets: true innovation in health care. Nat Rev Drug Discov 12: 85–86. doi: 10.1038/nrd3944 PMID: 23370233
4. Dar AC, Das TK, Shokat KM, Cagan RL (2012) Chemical genetic discovery of targets and anti-targets for cancer polypharmacology. Nature 486: 80–84. doi: 10.1038/nature11127 PMID: 22678283
5. Roth BL, Sheffler DJ, Kroeze WK (2004) Magic shotguns versus magic bullets: selectively non-selective drugs for mood disorders and schizophrenia. Nat Rev Drug Discov 3: 353–359. PMID: 15060530
6. Knight ZA, Lin H, Shokat KM (2010) Targeting the cancer kinome through polypharmacology. Nat Rev Cancer 10: 130–137. doi: 10.1038/nrc2787 PMID: 20094047
7. Lecca P, Priami C (2013) Biological network inference for drug discovery. Drug Discov Today 18: 256–264. doi: 10.1016/j.drudis.2012.11.001 PMID: 23147668
8. Harrold JM, Ramanathan M, Mager DE (2013) Network-based approaches in drug discovery and early development. Clin Pharmacol Ther 94: 651–658. doi: 10.1038/clpt.2013.176 PMID: 24025802
9. Keiser MJ, Setola V, Irwin JJ, Laggner C, Abbas AI, et al. (2009) Predicting new molecular targets for known drugs. Nature 462: 175–181. doi: 10.1038/nature08506 PMID: 19881490
10. Yabuuchi H, Niijima S, Takematsu H, Ida T, Hirokawa T, et al. (2011) Analysis of multiple compound-protein interactions reveals novel bioactive molecules. Mol Syst Biol 7: 472. doi: 10.1038/msb.2011.5 PMID: 21364574
11. Yamanishi Y, Araki M, Gutteridge A, Honda W, Kanehisa M (2008) Prediction of drug-target interaction networks from the integration of chemical and genomic spaces. Bioinformatics 24: i232–240. doi: 10.1093/bioinformatics/btn162 PMID: 18586719
12. Yamanishi Y, Kotera M, Kanehisa M, Goto S (2010) Drug-target interaction prediction from chemical, genomic and pharmacological data in an integrated framework. Bioinformatics 26: i246–254. doi: 10.1093/bioinformatics/btq176 PMID: 20529913

13. Chen X, Liu MX, Yan GY (2012) Drug-target interaction prediction by random walk on the heterogeneous network. Mol Biosyst 8: 1970–1978. doi: 10.1039/c2mb00002d PMID: 22538619

14. van Laarhoven T, Marchiori E (2013) Predicting Drug-Target Interactions for New Drug Compounds Using a Weighted Nearest Neighbor Profile. PLoS One 8: e66952. PMID: 23840562

15. van Laarhoven T, Nabuurs SB, Marchiori E (2011) Gaussian interaction profile kernels for predicting drug-target interaction. Bioinformatics 27: 3036–3043. doi: 10.1093/bioinformatics/btr500 PMID: 21893517

16. Alaimo S, Pulvirenti A, Giugno R, Ferro A (2013) Drug-target interaction prediction through domain-tuned network-based inference. Bioinformatics 29: 2004–2008. doi: 10.1093/bioinformatics/btt307 PMID: 23720490

17. Emig D, Ivliev A, Pustovalova O, Lancashire L, Bureeva S, et al. (2013) Drug target prediction and repositioning using an integrated network-based approach. PLoS One 8: e60618. doi: 10.1371/journal.pone.0060618 PMID: 23593264

18. Wang Y, Zeng J (2013) Predicting drug-target interactions using restricted Boltzmann machines. Bioinformatics 29: i126–134. doi: 10.1093/bioinformatics/btt234 PMID: 23812976

19. Cheng AC, Coleman RG, Smyth KT, Cao Q, Soulard P, et al. (2007) Structure-based maximal affinity model predicts small-molecule druggability. Nat Biotechnol 25: 71–75. PMID: 17211405

20. Morris GM, Huey R, Lindstrom W, Sanner MF, Belew RK, et al. (2009) AutoDock4 and AutoDockTools4: Automated docking with selective receptor flexibility. J Comput Chem 30: 2785–2791. doi: 10.1002/jcc.21256 PMID: 19399780

21. Li H, Gao Z, Kang L, Zhang H, Yang K, et al. (2006) TarFisDock: a web server for identifying drug targets with docking approach. Nucleic Acids Res 34: W219–224. PMID: 16844997

22. Reardon S (2013) Project ranks billions of drug interactions. Nature 503: 449–450. doi: 10.1038/503449a PMID: 24284710

23. Marti-Renom MA, Rossi A, Al-Shahrour F, Davis FP, Pieper U, et al. (2007) The AnnoLite and AnnoLyze programs for comparative annotation of protein structures. BMC Bioinformatics 8 Suppl 4: S4. PMID: 17570147

24. Kalinina OV, Wichmann O, Apic G, Russell RB (2011) Combinations of protein-chemical complex structures reveal new targets for established drugs. PLoS Comput Biol 7: e1002043. doi: 10.1371/journal.pcbi.1002043 PMID: 21573205

25. Morris RJ, Najmanovich RJ, Kahraman A, Thornton JM (2005) Real spherical harmonic expansion coefficients as 3D shape descriptors for protein binding pocket and ligand comparisons. Bioinformatics 21: 2347–2355. PMID: 15728116

26. Najmanovich R, Kurbatova N, Thornton J (2008) Detection of 3D atomic similarities and their use in the discrimination of small molecule protein-binding sites. Bioinformatics 24: i105–111. doi: 10.1093/bioinformatics/btn263 PMID: 18689810

27. Hoffmann B, Zaslavskiy M, Vert JP, Stoven V (2010) A new protein binding pocket similarity measure based on comparison of clouds of atoms in 3D: application to ligand prediction. BMC Bioinformatics 11: 99. doi: 10.1186/1471-2105-11-99 PMID: 20175916

28. Roy A, Yang J, Zhang Y (2012) COFACTOR: an accurate comparative algorithm for structure-based protein function annotation. Nucleic Acids Res 40: W471–477. doi: 10.1093/nar/gks372 PMID: 22570420

29. Berman HM, Battistuz T, Bhat TN, Bluhm WF, Bourne PE, et al. (2002) The Protein Data Bank. Acta Crystallogr D Biol Crystallogr 58: 899–907. PMID: 12037327

30. Orti L, Carbajo RJ, Pieper U, Eswar N, Maurer SM, et al. (2009) A kernel for open source drug discovery in tropical diseases. PLoS Negl Trop Dis 3: e418. doi: 10.1371/journal.pntd.0000418 PMID: 19381286

31. Martinez-Jimenez F, Papadatos G, Yang L, Wallace IM, Kumar V, et al. (2013) Target prediction for an open access set of compounds active against Mycobacterium tuberculosis. PLoS Comput Biol 9: e1003253. doi: 10.1371/journal.pcbi.1003253 PMID: 24098102

32. Wishart DS, Knox C, Guo AC, Cheng D, Shrivastava S, et al. (2008) DrugBank: a knowledgebase for drugs, drug actions and drug targets. Nucleic Acids Res 36: D901–906. PMID: 18048412

33. van Laarhoven T, Marchiori E (2014) Biases of drug–target interaction network data.

34. Fitzpatrick FA (2004) Cyclooxygenase enzymes: regulation and function. Curr Pharm Des 10: 577–588. PMID: 14965321

35. Hawkey CJ (2001) COX-1 and COX-2 inhibitors. Best Pract Res Clin Gastroenterol 15: 801–820. PMID: 11566042

36. Peng CL, Guo W, Ji T, Ren T, Yang Y, et al. (2009) Sorafenib induces growth inhibition and apoptosis in human synovial sarcoma cells via inhibiting the RAF/MEK/ERK signaling pathway. Cancer Biol Ther 8: 1729–1736. PMID: 19633425

37. Liu L, Cao Y, Chen C, Zhang X, McNabola A, et al. (2006) Sorafenib blocks the RAF/MEK/ERK pathway, inhibits tumor angiogenesis, and induces tumor cell apoptosis in hepatocellular carcinoma model PLC/PRF/5. Cancer Res 66: 11851–11858. PMID: 17178882

38. Ravandi F, Cortes JE, Jones D, Faderl S, Garcia-Manero G, et al. (2010) Phase I/II study of combination therapy with sorafenib, idarubicin, and cytarabine in younger patients with acute myeloid leukemia. J Clin Oncol 28: 1856–1862. doi: 10.1200/JCO.2009.25.4888 PMID: 20212254

39. Ravandi F, Alattar ML, Grunwald MR, Rudek MA, Rajkhowa T, et al. (2013) Phase 2 study of azacytidine plus sorafenib in patients with acute myeloid leukemia and FLT-3 internal tandem duplication mutation. Blood 121: 4655–4662. doi: 10.1182/blood-2013-01-480228 PMID: 23613521

40. Eggert US (2013) The why and how of phenotypic small-molecule screens. Nat Chem Biol 9: 206–209. doi: 10.1038/nchembio.1206 PMID: 23508174

41. Gamo FJ, Sanz LM, Vidal J, de Cozar C, Alvarez E, et al. (2010) Thousands of chemical starting points for antimalarial lead identification. Nature 465: 305–310. doi: 10.1038/nature09107 PMID: 20485427

42. Ballell L, Bates RH, Young RJ, Alvarez-Gomez D, Alvarez-Ruiz E, et al. (2013) Fueling Open-Source Drug Discovery: 177 Small-Molecule Leads against Tuberculosis. ChemMedChem.

43. Minie M, Chopra G, Sethi G, Horst J, White G, et al. (2014) CANDO and the infinite drug discovery frontier. Drug Discov Today 19: 1353–1363. doi: 10.1016/j.drudis.2014.06.018 PMID: 24980786

44. Berman HM, Coimbatore Narayanan B, Di Costanzo L, Dutta S, Ghosh S, et al. (2013) Trendspotting in the Protein Data Bank. FEBS Lett 587: 1036–1045. doi: 10.1016/j.febslet.2012.12.029 PMID: 23337870

45. Norvell JC, Berg JM (2007) Update on the protein structure initiative. Structure 15: 1519–1522. PMID: 18073099

46. Khafizov K, Madrid-Aliste C, Almo SC, Fiser A (2014) Trends in structural coverage of the protein universe and the impact of the Protein Structure Initiative. Proc Natl Acad Sci U S A 111: 3733–3738. doi: 10.1073/pnas.1321614111 PMID: 24567391

47. Bickerton GR, Paolini GV, Besnard J, Muresan S, Hopkins AL (2012) Quantifying the chemical beauty of drugs. Nat Chem 4: 90–98. doi: 10.1038/nchem.1243 PMID: 22270643

48. Stuart AC, Ilyin VA, Sali A (2002) LigBase: a database of families of aligned ligand binding sites in known protein sequences and structures. Bioinformatics 18: 200–201. PMID: 11836232

49. Konc J, Janezic D (2010) ProBiS: a web server for detection of structurally similar protein binding sites. Nucleic Acids Res 38: W436–440. doi: 10.1093/nar/gkq479 PMID: 20504855

50. Pieper U, Webb BM, Dong GQ, Schneidman-Duhovny D, Fan H, et al. (2014) ModBase, a database of annotated comparative protein structure models and associated resources. Nucleic Acids Res 42: D336–346. doi: 10.1093/nar/gkt1144 PMID: 24271400

51. Pieper U, Webb BM, Barkan DT, Schneidman-Duhovny D, Schlessinger A, et al. (2011) ModBase, a database of annotated comparative protein structure models, and associated resources. Nucleic Acids Res 39: D465–474. doi: 10.1093/nar/gkq1091 PMID: 21097780

52. Pieper U, Eswar N, Stuart AC, Ilyin VA, Sali A (2002) MODBASE, a database of annotated comparative protein structure models. Nucleic Acids Res 30: 255–259. PMID: 11752309

53. Eswar N, John B, Mirkovic N, Fiser A, Ilyin VA, et al. (2003) Tools for comparative protein structure modeling and analysis. Nucleic Acids Res 31: 3375–3380. PMID: 12824331

54. Shen MY, Sali A (2006) Statistical potential for assessment and prediction of protein structures. Protein Sci 15: 2507–2524. PMID: 17075131

55. Eramian D, Shen MY, Devos D, Melo F, Sali A, et al. (2006) A composite score for predicting errors in protein structure models. Protein Sci 15: 1653–1666. PMID: 16751606

56. Ortiz AR, Strauss CE, Olmea O (2002) MAMMOTH (matching molecular models obtained from theory): an automated method for model comparison. Protein Sci 11: 2606–2621. PMID: 12381844

57. Dijkstra EW (1959) A note on two problems in connexion with graphs. Numerische Mathematik 1: 269–271.

58. Hall M, Frank E, Holmes G, Pfahringer B, Reutemann P, et al. (2009) The WEKA Data Mining Software: An Update. SIGKDD Explorations 11.

59. Ashburner M, Ball CA, Blake JA, Botstein D, Butler H, et al. (2000) Gene ontology: tool for the unification of biology. The Gene Ontology Consortium. Nat Genet 25: 25–29. PMID: 10802651

60. Ogata H, Goto S, Sato K, Fujibuchi W, Bono H, et al. (1999) KEGG: Kyoto Encyclopedia of Genes and Genomes. Nucleic Acids Res 27: 29–34. PMID: 9847135

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 `http://www.tropicaldisease.org/TCAMSTB`.

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

OPEN ACCESS Freely available online

PLOS | COMPUTATIONAL BIOLOGY

Target Prediction for an Open Access Set of Compounds Active against *Mycobacterium tuberculosis*

Francisco Martínez-Jiménez[1,2], George Papadatos[3], Lun Yang[4], Iain M. Wallace[3], Vinod Kumar[4], Ursula Pieper[5], Andrej Sali[5], James R. Brown[4]*, John P. Overington[3]*, Marc A. Marti-Renom[1,2]*

1 Genome Biology Group, Centre Nacional d'Anàlisi Genòmica (CNAG), Barcelona, Spain, **2** Gene Regulation Stem Cells and Cancer Program, Centre for Genomic Regulation (CRG), Barcelona, Spain, **3** European Molecular Biology Laboratory – European Bioinformatics Institute (EMBL-EBI), Wellcome Trust Genome Campus, Hinxton, Cambridge, United Kingdom, **4** Computational Biology, Quantitative Sciences, GlaxoSmithKline, Collegeville, Pennsylvania, United States of America, **5** Department of Bioengineering and Therapeutic Sciences, University of California, San Francisco, San Francisco, California, United States of America

Abstract

Mycobacterium tuberculosis, the causative agent of tuberculosis (TB), infects an estimated two billion people worldwide and is the leading cause of mortality due to infectious disease. The development of new anti-TB therapeutics is required, because of the emergence of multi-drug resistance strains as well as co-infection with other pathogens, especially HIV. Recently, the pharmaceutical company GlaxoSmithKline published the results of a high-throughput screen (HTS) of their two million compound library for anti-mycobacterial phenotypes. The screen revealed 776 compounds with significant activity against the *M. tuberculosis* H37Rv strain, including a subset of 177 prioritized compounds with high potency and low *in vitro* cytotoxicity. The next major challenge is the identification of the target proteins. Here, we use a computational approach that integrates historical bioassay data, chemical properties and structural comparisons of selected compounds to propose their potential targets in *M. tuberculosis*. We predicted 139 target - compound links, providing a necessary basis for further studies to characterize the mode of action of these compounds. The results from our analysis, including the predicted structural models, are available to the wider scientific community in the open source mode, to encourage further development of novel TB therapeutics.

Citation: Martínez-Jiménez F, Papadatos G, Yang L, Wallace IM, Kumar V, et al. (2013) Target Prediction for an Open Access Set of Compounds Active against *Mycobacterium tuberculosis*. PLoS Comput Biol 9(10): e1003253. doi:10.1371/journal.pcbi.1003253

Editor: Alexander Donald MacKerell, University of Maryland, Baltimore, United States of America

Received May 2, 2013; **Accepted** August 11, 2013; **Published** October 3, 2013

Funding: MAMR acknowledges the support from the Spanish Government (BFU2010-19310) and the Era-Net Pathogenomics (PIM2010EPA-00719). JPO and GP were supported by funding from the EMBL Member States. JRB acknowledges the support of GlaxoSmithKline R&D and AS acknowledges support from NIH (U54 GM094662 and P01 GM71790). GlaxoSmithKline participated in this research through the co-authorship of VK, LY, and JRB. The funders had no role in study design, data collection and analysis, decision to publish, or preparation of the manuscript.

Competing Interests: VK, LY, and JRB are paid employees of GlaxoSmithKline (GSK). This does not alter our adherence to the PLOS policies on sharing data and materials. All other authors have declared that no competing interests exist.

* E-mail: James.R.Brown@gsk.com (JRB); jpo@ebi.ac.uk (JPO); mmarti@pcb.ub.cat (MAMR)

Introduction

One third of the world's population is infected with *Mycobacterium tuberculosis* (MTB), the causative agent of tuberculosis [1]. Approximately 95% of infected individuals are thought to have persistent, latent MTB infections that remain dormant until activated by specific environmental and host response events. Approximately 10% of latent infections eventually progress to active disease, which, if left untreated, kills more than half of the infected patients [2]. Moreover, there is an increasing clinical occurrence of MTB strains with extensive multi-drug-resistance (*eg*, MTB MDR and MTB XDR), where mortality rates can approach 100% [3]. In some countries, the MTB MDR and XDR strains may account for up to 22% of infections [1]. In addition, current TB therapeutic regimes involve a combination of antibiotics, administered at regular intervals over a 6-month period, which makes patient compliance an issue, especially in developing countries [1,2].

The discovery and development of new antibiotics is widely recognized as one of the major global health emergencies, yet it is also a major pharmaceutical challenge. Most currently used antibiotics were discovered during the golden era from the 1940s to 1960s through large scale screening of compound collections for anti-bacterial activity – the so-called whole cell or phenotypic screens [4]. The emergence of bacterial molecular genomics technologies and the availability of whole genome sequences in the 1990s led to dramatic changes in anti-bacterial drug discovery, where the emphasis was placed on screening essential targets for inhibitory compounds. However, despite intensive efforts, target-based screening has been largely unsuccessful in producing clinical candidate molecules [5]. As a result, a return to whole cell screening has been widely advocated, in combination with novel technologies and bioinformatics to rapid identify targets associated with a compound's mechanism of action (MOA) [4,6].

Recently, the pharmaceutical company GlaxoSmithKline (GSK) completed an anti-mycobacterial phenotypic screening campaign against *M. bovis* BCG, a non-virulent, vaccine *Mycobacterium* strain, with a subsequent secondary screening in *M. tuberculosis* H37Rv (MTB H37Rv) for hit confirmation [7]. A total of 776 potent compound hits (including 177 MTB H37RV hits with limited human cell line toxicity) were made openly available to the wider scientific community through the ChEMBL database (http://dx.doi.org/10.6019/CHEMBL2095176). The aim of this release was to stimulate mechanism of action analyses using

Author Summary

Mycobacterium tuberculosis is a major worldwide pathogen infecting millions individuals every year. Additionally, the number of antibiotic resistant strains has dramatically increased over the last decades. Trying to address this challenge, the pharmaceutical company GlaxoSmithKline has recently published the results of a large-scale high-throughput screen (HTS) that resulted in the release of 776 chemical compound structures active against tuberculosis. We have used this dataset of compounds as input to our computational approach that integrates historical bioassay data, chemical properties and structural comparisons. We propose 139 targets alongside their respective hit compounds and made them open to the wider scientific community. Our hope is that the availability of the experimental data from GSK and our computational analysis will encourage further research providing validated therapeutically targets against this devastating disease.

chemical genetics/proteomics approaches, as well as to provide many potential new starting points for synthetic lead generation activities. To attain these goals, it is essential to identify the likely protein targets of these active compounds. Here, we introduce an integrative computational analysis towards the genome-wide characterization of targets for selected compounds against tuberculosis. Our approach is in contrast to the classical target-based experiments, widely used in drug discovery, that suffer from very high attrition rates in anti-infective molecules [8]. This study should also serve the wider anti-tuberculosis research community by providing a list of genes and pathways that are more likely to be validated as TB targets for drug discovery and development.

We applied computational approaches using three domains of knowledge, namely the "assay space", "chemogenomics space" and "structural space", to identify new targets that are likely to interact with the active compounds from the GSK collection. We characterized the structural and chemical spaces of the recently released set of 776 compounds active against tuberculosis [7] and grouped the compounds into a total of 551 structural families. Subsequently, we predicted their likely targets using three orthogonal and complementary computational approaches. Jointly, we identified several amino-acid biosynthesis proteins as possible targets of several compounds in the dataset. A total of 207 unique pairs of compounds and potential MTB targets have been predicted. These compounds constitute a basis for further hypothesis-led exploration of their mode of action. We briefly outline the possible impact and contribution of our findings to Open Drug Discovery Initiatives [9,10,11], in particular against tuberculosis.

Results/Discussion

The TCAMS-TB compound dataset

GSK recently released the data from a phenotypic screen against tuberculosis (available at ChEMBL http://dx.doi.org/10.6019/CHEMBL2095176) [7]. This open access dataset contains a total of 776 compounds active against *M. bovis* BCG, a non-virulent *Mycobacterium* species widely used in experimental studies as a vaccine component, and a subset of 177 confirmed compounds active against MTB strain H37Rv. The compound collection had been pre-filtered to remove known anti-bacterial compounds to maximize the discovery of novel compounds with anti-Mycobacterium activities. About 90% of the compounds have a quantitative estimate of drug-likeness (QED) value above 0.35 [12], herein called optimal drug-like compounds (Figure 1). The remaining 10% of compounds, which are highlighted by red bars in Figure 1, have higher molecular weights (>400 KD) and slightly higher hydrophobicity, expressed as the calculated logarithm of the 1-octanol/water partition coefficient (ALogP) [13]. For the subset of 177 compounds active against H37Rv, the average molecular weight is statistically smaller than for the entire dataset (Figure 1), consistent with known trends of lipophilicity and cytotoxicity/polypharmacology. The molecular PSA (polar surface area), ALogP (octanol–water partition coefficient) and wQED (weighted QED) scores result in statistically indistinguishable average values and distributions for both datasets. To assess the diversity of the dataset, we applied our Random Forest Score (RFS) to identify pairs of similar compounds (Methods). An all-against-all comparison was performed by nAnnolyze [14] and any pair of compounds with an RFS higher than 0.9 were considered similar. The resulting network of compound similarities was layered using Cytoscape [15] (Figure 1E). The entire dataset of 776 compounds was clustered into a total of 551 compound families, primarily composed of two large compound families and 481 singleton families. The two large families of compounds (GSKFAM_1 and GSKFAM_2) included 38 compounds each connected by 156 and 80 links, respectively (Figure 1F). In summary, the active compound set released by GSK is composed of drug-like molecules with non-redundant and diverse scaffolds.

Integrative computational analysis

The 776 compounds released by GSK were used as input to our integrative computational analysis approach that combines the results from a chemogenomics space search (CHEM), a structural space search (STR) and a historical assay space search (HIST). First, the exploration of the chemical space allowed us to identify likely targets for the input compounds based on their structural similarity to compounds with experimentally validated targets deposited in the ChEMBL database [16]. The approach employed a multi-category Naïve Bayesian classifier, which has been successfully used in ligand-based target prediction efforts [17,18,19]. Second, the exploration of the structural space allowed for the identification of likely targets based on the structural similarity of compounds and protein targets with known three-dimensional structures. The method was based on an improved version of the AnnoLyze program [14]. Finally, the exploration of the historical data on screening assays resulted in testable hypotheses for the anti-*Mycobacterium* mode of action of the selected compounds, based on the historical data from internal GSK screening experiments. This integrative approach allowed us to predict targets for the set of released compounds in the absence of known structural data (CHEM and HIST) or the absence of knowledge of the binding site (STR). When the three-dimensional structure of the target and the localization of the binding site are known or predicted, it is often helpful to follow up with molecular docking (see [20] and examples below). However, such an approach would be prohibitive for large numbers of compounds against a large number of targets, because molecular docking results still need to be interpreted manually for best impact. The three methods used in our integrative approach are further detailed in the Methods section of this manuscript.

Chemogenomics space (CHEM)

We applied a multi-category Naïve Bayesian classifier (MCNBC) that was built and trained using structural and bioactivity information from the ChEMBL database [16]. Given a new compound, the model calculates a likelihood score based on the molecule's individual sub-structural/fingerprint features and

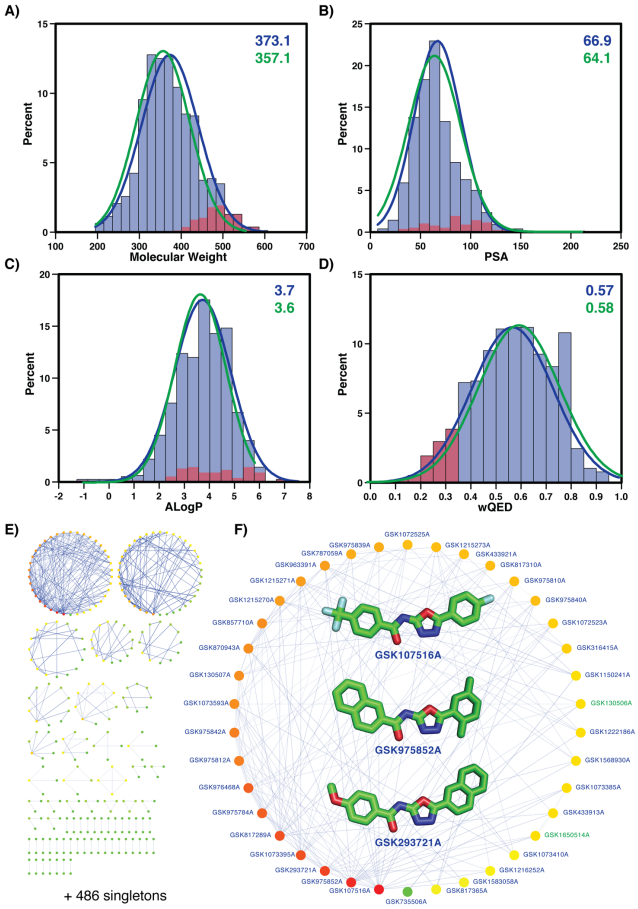


Figure 1. GSK dataset of 776 compounds. Panels A to D describe the drug-like properties of the compounds, including the subset of 177 compounds active against MTB (green color). Red colored subsets correspond to compounds with weighted QED score smaller than 0.35 [12]. The distribution's mean values are shown in the top-right corner of each plot. **A**) Molecular weight distribution. **B**) PSA distribution. **C**) ALogP distribution. **D**) Weighted QED distribution. Panels E and F show the structural clusters of the compounds. Links between compounds indicate 0.9 or higher RFS similarity. **E**) Entire network of 776 compounds resulting in 551 structural families (486 singletons). **F**) Highlight of family number 1 with 38 compounds (inner images for the three most connected compounds in the family).
doi:10.1371/journal.pcbi.1003253.g001

produces a ranked list of likely targets. In total, the 776 compounds in the *M. bovis* BCG dataset resulted in 2,179 statistically significant target associations (at a Z-score >2.0) to proteins in the ChEMBL database from 62 different organisms (63% of hits are to human proteins). A simple orthology search against the MTB proteins from this set resulted in 1,401 compound-target relationships for 84 MTB proteins, with detectable orthology to 34 organisms. The specific predictions from the chemical space search are available at http://www.tropicaldisease.org/TCAMSTB (CHEM type).

Structural space (STR)

We applied a Random Forest Score that identified structural similarities between any compound in the dataset and ligands from

the Protein Data Bank (PDB) [14]. Each compound in the *M. bovis* BCG dataset is compared to ~2,500 ligands for which there are known complex structures in the PDB, identifying structural similarities to be included in a pre-built network of structural relationships between ligands and targets. In total, the 776 compounds resulted in 207 significant target associations (RFS score >0.4) to proteins in a set of modeled three-dimensional structures from the MTB proteome. The specific predictions from the structural space search are available at http://www.tropicaldisease.org/TCAMSTB (STR type).

Historical assay space (HIST)

We used the historical GSK bioassay data to develop hypotheses for the anti-*Mycobacterium* mode of action for the active compounds. Using conservative activity thresholds, we found among the compounds active against MTB H37Rv unambiguous annotations for 49 compounds and their previously measured activity in 120 biochemical assays against 63 human targets (*i.e.*, sub-micromolar IC50 or EC50). Overall, the *M. bovis* BCG screens resulted in a considerably larger number of active compounds and thus have a correspondingly greater amount of historical assay information. A total of 240 compounds were found to have activity recorded in 642 assays involving 209 human targets, with the largest human target classes being GPCRs and protein kinases, as expected. We then searched for orthologous sequences of the human assayed proteins in the MTB H37Rv and *M. bovis* BCG genomes using conservative criteria for assigning human-*Mycobacterium* homology (BLAST E-value $\leq 1.0e^{-10}$). Although there are significant evolutionary differences between bacterial and mammalian genomes, we still found 19 *M. bovis* BCG homologous genes (Table S1) in different target classes (Figure S1), including kinases (8 genes), cytochrome P450s (2 genes) and nine other enzymes such as a putative D-amino acid oxidase, an amidase, a putative flavin-containing monoamine oxidase, a NAD-dependent deacetylase, a putative catechol-O-methyltransferase, a protease, a putative epoxide hydrolase, a 3-ketoacyl-(acyl-carrier-protein) reductase, and a dihydroorotate dehydrogenase 2. While these *M. bovis* BCG genes had orthologous sequences in MTB H37Rv, fewer compounds were associated with putative targets in the latter species. For example, two *Mycobacterium* kinases and five enzymes were exclusively associated with *M. bovis* BCG positive compounds. Two kinases (pknA and pknB) and one enzyme (fabG) were experimentally characterized as essential for the survival of MTB [21,22]. A total of 20 and 94 compounds were indirectly mapped by human protein target homology to 12 MTB H37Rv and 19 *M. bovis* BCG genes, respectively. The specific predictions from the historical assay space search are detailed in Supporting Information and are available at http://www.tropicaldisease.org/TCAMSTB (HIST type).

Subset of compounds with predicted targets

Of the 776 compounds in the GSK dataset, only one compound (GSK445886A) was predicted to hit diverse targets from different pathways by the three independent methods (Figure 2A). A total of 25 and 9 compounds were jointly predicted to hit a target by CHEM/STR and CHEM/HIST searches, respectively. The majority of predictions were obtained by the CHEM approach (404 compounds with predicted targets), followed by the STR approach (38 compounds with a predicted target) and the HIST approach (20 compounds with predicted targets). Such results were expected because the available information on biological activity shrinks as we move from the general "chemical" to the more specific "structural" and "historical" spaces. Interestingly, as an indication of the orthogonality of the three approaches, most of the redundancy of compounds with a predicted target was specific to each approach. In other words, each of the three approaches covered different parts of the space of compound-target predictions. For example, the CHEM approach predicted a target for 300 compound families (compared to a total of 404 unique compounds), of which it still shared 34 with either the STR or the HIST approaches (Figure 1B). A similar trend was observed for the other two approaches, indicating that the common compounds mostly occurred in small compound families or even singletons. Indeed, the GSK445886A compound, which was predicted to have a target by all three approaches, corresponded to a singleton compound family (GSKFAM_293).

To identify whether the three different approaches predicted targets for specific families in the dataset, we calculated the log odds probability (LogOdd) of a given compound family to appear in the list of selected compounds, given their different distributions in the original dataset (Figure 2C). This analysis aimed at identifying possible biases or artifacts specific to each of the three independent methods used in our integrative approach. Eleven compound families were under-represented in the selected dataset and 18 families were over-represented (with LogOdd values smaller than −0.5 and greater than 0.5, respectively). Interestingly, GSKFAM_551, which is a singleton with the SKF-67461 compound, was over-represented in the subset of selected compounds. Such predictions were based mostly on the STR and CHEM searches and may correspond to the chemical properties of the compound, resulting in a high false-positive rate for those two approaches. Conversely, the GSKFAM_4, which contains 15 compounds, is under-represented in the final subset of selected compounds, with only 1 hit identified by the CHEM approach.

Predicted targets

There are a total of 1,044 unique MTB targets associated with a total of 112 pathways annotated in the KEGG database [23] (the mtu identifiers below refer to the relevant KEGG pathway id). Of those, the three orthogonal approaches identified targets for the selected set of compounds in a total of 84 pathways (Figure 3A). The STR search resulted in hits to 71 unique pathways, while the CHEM and the HIST searches resulted in hits to 35 and 16 pathways, respectively. These results were expected, because the target information is reduced from the STR space to the HIST space. A total of 11 unique pathways were predicted by the three approaches (Figure 3A and Table 1); these include many pathways associated with amino acid and nucleotide metabolism, such as arginine and proline metabolism (mtu00330), tryptophan metabolism (mtu00380), phenylalanine metabolism (mtu00360), tyrosine metabolism (mtu00350), histidine metabolism (mtu00340), glycine/serine/threonine metabolism (mtu00260) and pyrimidine metabolism (mtu00240). The results indicate that the GSK compounds potentially target proteins associated with primary metabolism. Interestingly, another seven pathways, not identified by the HIST approach, were found over-represented in the final set of predicted targets (Figure 3B). Those include some further primary and secondary metabolism systems, including streptomycin biosynthesis (mtu00521), folate biosynthesis (mtu00790), nitrogen metabolism (mtu00910), aminoacyl-tRNA biosynthesis (mtu00970), purine metabolism (mtu00230), penicillin and cephalosporin biosynthesis (mtu00311), D-arginine and D-ornithine

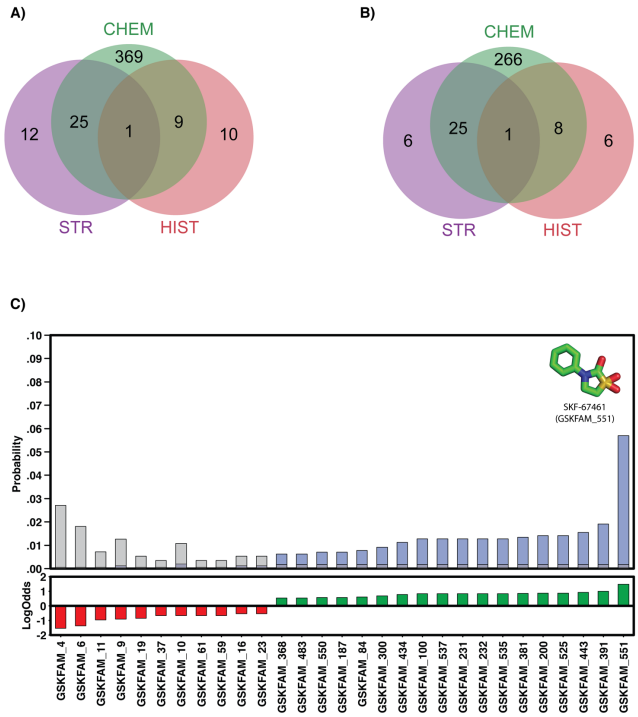


Figure 2. Subset of GSK compounds with predicted targets. A) Venn diagram with common compounds with predictions from the three different approaches (that is, in green from the search of the chemogenomics space, in purple from the search of the structural space, and in red from the historical data). **B**) Venn diagram with common compound families with predictions from the three different approaches. **C**) Most under and over-represented chemical families in our predictions. Upper plot shows the probability of finding a given family in the original dataset (grey bars) compared to the probability of finding it in the dataset with predicted targets (blue bars). Lower plot shows the log odds per selected family (*i.e.,* absolute log odds larger than 0.5).
doi:10.1371/journal.pcbi.1003253.g002

metabolism (mtu00472), and one carbon pool by folate (mtu00670).

Predicted pairs of compound-target

To assess the significance of our predictions using the three different approaches, we calculated a t-statistics p-value of any compound family - KEGG pathway pair (Methods). The search identified 8 different compound families with significant links (p-value $<1\times10^{-5}$) to 14 different KEGG pathways (Table 2). The GSK compound family 1, through its compounds GSK975784A, GSK975810A, GSK975839A, GSK975840A and GSK975842A, was predicted to target the glycerolipid (mtu00561) and glycerophospholipid metabolisms (mtu00564), with significant over-representation through 6 different targets including Rv2182c and Rv2483c, both acyltransferases essential for the survival of the bacteria [21]. The GSK compound family 3 was predicted to target the ABC transporters (mtu02010) through its compounds GSK547481A, GSK547490A, GSK547491A, GSK547499A, GSK547500A, GSK547511A, GSK547512A, GSK547527A, GSK547528A and GSK547543A. Similarly, it was also predicted to target the aminoacyl-tRNA biosynthesis (mtu00970) pathways, through 3 different targets including Rv1640c, a lysyl-tRNA synthetase essential for the survival of the bacteria [21]. The GSK compound family 7, was predicted to target several pathways through 2 different targets Rv0053 (30S ribosomal protein S6) and Rv0650 (a glucokinase), none considered essential for the survival of the bacteria [21]. The GSK compound family 9 through its compounds GSK1188379A and GSK1188380A, was predicted to target the ABC transporters (mtu02010) pathway through the Rv0194 target (ATP-binding cassette, subfamily C) considered non-essential for the survival of the bacteria [21]. Identical results were obtained with the GSK compound family 16 through its compounds GSK1825940A and GSK1825944A. The GSK compound family 35 through its compounds BRL-10143SA and

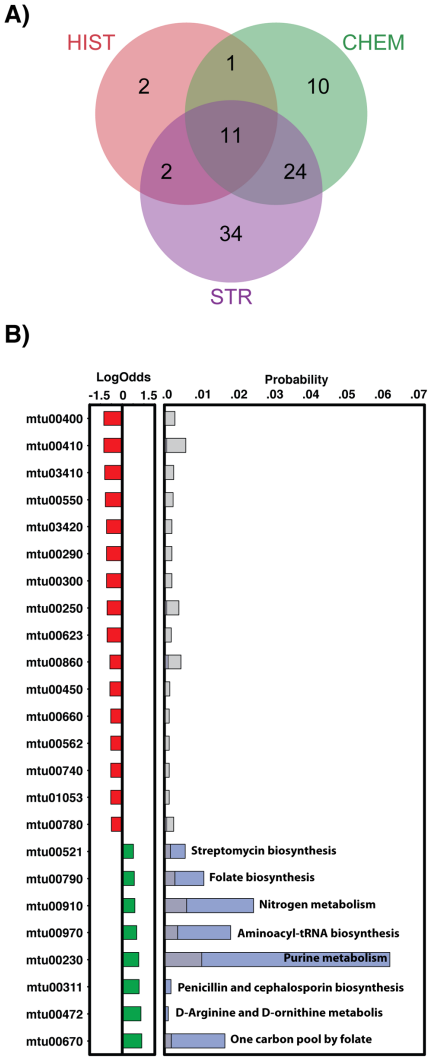


Figure 3. Predicted KEGG pathways targeted by the GSK compounds. A) Venn diagram with common pathways from the three different approaches. **B**) Most under and over-represented pathways in our predictions. Panels A) and B) with the same representation as in Figure 2.
doi:10.1371/journal.pcbi.1003253.g003

BRL-51093AA was predicted to target the one carbon pool by folate (mtu00670) pathway through the Rv2763c and Rv2764c targets (a dihydrofolate reductase and a thymidylate synthase, respectively) considered non-essential for the survival of the bacteria [21]. The GSK compound family 173 through its compound GSK14022909A was predicted to target the aminoacyl-tRNA biosynthesis (mtu00970) pathway through three essential targets [21], Rv1640c, Rv3598c and Rv3834c (a lysyl-tRNA synthetase, a lysyl-tRNA ligase, and a seryl-tRNA ligase, respectively), which are essential for the survival of the organism [21]. Interestingly, this family is also predicted to target Rv3105c and Rv3135 genes (a peptide chain release factor 2 and a PPE family protein), which are also essential for the survival of the organism [21]. Finally, the GSK compound family 334 through compound GSK270671A was predicted to target the nitrogen metabolism (mtu00910) pathway through the Rv1284 and Rv3588 targets (carbonic anhydrases) considered essential for the survival of the bacteria [21].

An example of a serine/threonine-protein kinase (pknB) target

Even though target Rv0014c, a serine/threonine-protein kinase, was not identified as belonging to an enriched pathway (it is not annotated in the KEGG database), it was predicted by the HIST approach to be a target for the GSK1365028A, GSK1598164A, GSK275628A and GW664700A (all singleton families in our compound clustering). Kinases are the most prominent human target class having identifiable orthologs in both *M. tuberculosis* H37Rv and *M. bovis* BCG genomes (Figure 4A). The human genome encodes over 450 kinases, while *Mycobacterium* contains between 4 and 24 serine/threonine kinases, depending on the exact species (*M. tuberculosis* and *M. bovis* have 11 conserved kinases each). At least two of these kinases, pknA and pknB, have been determined to be essential for *in vitro* viability of *M. tuberculosis* [21]. To further evaluate potential MoA of kinase inhibitors, we computationally docked several compounds into the adenine-binding portion of the ATP binding pockets of the two available experimental structures for the essential kinase pknB. The criteria for choosing the compounds were whole cell screening activity of MIC90 less than 10 μM and IC50 less than 8 μM. Two structures (PDB IDs: 2PZI and 3F69) were selected because both were co-crystallized with an inhibitor, clearly detailing their ATP binding pockets.

An empirical docking score threshold of −8.5 kJ/mol was chosen to identify putative positive bindings of the active compounds across the two pknB PDB models (Table S2). GSK1598164A, an inhibitor of several human serine/threonine protein kinases, was positive in both H37RV and BCG whole cell screens, based on favorable docking scores (−9.19 and −8.96 kJ/mol against 2PZI and 3F69, respectively). Both GSK1598164A and the enzymatic product ADP in the crystal structure were found to interact with the Glu93 of pknB, where the nitrogen atoms on the 'head' unit form the hydrogen bond with Glu93 (Figure 4B). Glu93 is conserved across both human and TB kinases (Figure 4A). Several residues in the putative hydrophobic binding pocket (Leu17, Gly18, Phe19, Val25, Ala38, Val72, Met92, Glu93 and Val95) were also found to be within 4 Å of both GSK1598164A and ADP. In conclusion, our analysis suggests that several bactericidal compounds in the published phenotypic screen act by inhibiting essential *M. tuberculosis* kinases.

Table 1. List of seven common hit pathways identified by the three independent approaches.

Pathway	Approach	Targets	Compound families
mtu00240	STR	*Rv1381*	**255**
Pyrimidine metabolism		Rv3048c	**86**
		Rv3314c	**255**
	CHEM	Rv2139	**Several**
		Rv2764c	**Several**
		Rv3247c	**497**
	HIST	Rv2139	**2**
mtu00260	STR	Rv0489	**551**
Glycine, serine and threonine metabolism		*Rv1296*	**551**
		Rv3708c	**551**
	CHEM	Rv1905c	**5,252,497**
		Rv3170	**Several**
	HIST	Rv3170	**5**
mtu00330	STR	*Rv1652*	**476,488**
Arginine and proline metabolism	CHEM	Rv0458	**60**
		Rv1905c	**5,252,497**
		Rv3170	**Several**
	HIST	Rv1263	**5,272**
		Rv3170	**5**
mtu00340	STR	Rv0187	**551**
Histidine metabolism		Rv0520	**551**
		Rv1498c	**300**
		Rv1603	**551**
		Rv1605	**551**
	CHEM	Rv0458	**60**
		Rv3170	**Several**
	HIST	Rv3170	**5**
mtu00350	STR	Rv0187	**551**
Tyrosine metabolism		Rv0520	**551**
		Rv1498c	**300**
		Rv1703c	**551**
	CHEM	Rv3170	**Several**
	HIST	Rv3170	**5**
mtu00360	STR	Rv1908c	**551**
Phenylalanine metabolism		Rv3469c	**551**
	CHEM	Rv3170	**Several**
	HIST	Rv1263	**5,272**
		Rv3170	**5**
mtu00380	STR	Rv0859	**551**
Tryptophan metabolism		Rv1908c	**551**
	CHEM	Rv0458	**60**

Table 1. Cont.

Pathway	Approach	Targets	Compound families
		Rv1323	**Several**
		Rv3170	**Several**
	HIST	Rv1263	**5,272**
		Rv3170	**5**

The additional four common pathways identified not shown correspond to general pathway descriptions (i.e., mtu01100 "Metabolic pathways", mtu01110 "Biosynthesis of secondary metabolites", mtu01120 "Microbial metabolism in diverse environments", and mtu00000 "No Pathway"). Target genes in italics are either *in vivo* or *in vitro* essential in the TraSH Essentiality database [21].
doi:10.1371/journal.pcbi.1003253.t001

An example of a compound targeting the aminoacyl-tRNA biosynthesis pathway

The CHEM and STR methods identified Rv3598c (lysS1 lysine-tRNA ligase 1) and Rv3834c (serS serine-tRNA ligase) as possible targets for the GSK1402290A compound, respectively. Both enzymes are part of the aminoacyl-tRNA biosynthesis pathway (mtu00970) and are essential in *in vitro* experiments [21]. Moreover, the mtu00970 pathway was selected in our analysis as being significantly associated with GSKFAM_173 (GSK1402290A compound).

The CHEM approach predicted that the human lysyl-tRNA synthetase (UniProt ID Q15046) was a likely target of GSK1402290A, with a likelihood score of 11.3 and a Z-score of 2.4. Furthermore, the model indicated that the individual fragments contributing to this prediction were derived by its fused triazole ring (*e.g.*, pyrazole and imidazole features), as well as by its aniline group. In fact, the model for this target was trained using 47 active compounds from ChEMBL and almost all of them contained the aforementioned fragments (Figure 5A). Moreover, the predicted human target shared in OrthoMCL [24] the ortholog group (OG5_126972) with MTB's lysine-tRNA ligase 1 (UniProt ID P67607).

The STR method predicted a link between the compound and the target through a 3D model of the Rv3834c protein built based on the known structure of a seryl-tRNA synthetase from *Aquifex aeolicus*. The Rv3834c target and the seryl-tRNA synthetase template aligned with 43% sequence identity and resulted in good quality models (MPQS>1.5) [25]. To further evaluate potential MoA of the GSK1402290A compound, we computationally docked it into the nAnnoLyze predicted binding site for Rv3834c (Figure 5). The AutoDock run resulted in a best pose with −8,4 kJ/mol, indicating interactions between the GSK1402290A compound and the Rv3834c target (Figure 5B). In support of this model, the interactions occur with conserved protein residues, given the curated multiple sequence alignment for PFAM family PF00587 (tRNA synthetase class II core domain).

In summary, our CHEM and STR predictions suggest that GSK1402290A could act as an inhibitor of the aminoacyl-tRNA biosynthesis pathway and provide the basis for further chemical optimization of this compound.

Open targets against tuberculosis

The recent publication of a large-scale screening effort for identifying drug-like small molecule compounds active against tuberculosis has been used as starting point for our research. Here, we predicted the likely mode of action of a selected set of compounds active against tuberculosis, based on a computational approach that integrates data from historical assay results, chemical features and their relationship to activity, and structural comparisons. Our integrated approach resulted in prediction of several compound-target pairs, which can be further tested using genomics, genetics and biochemical assays. More broadly, our approach can be applied to whole cell screens for any pathogen, provided sufficient datasets are available.

We have predicted a wide range of MTB specific as well as more evolutionary conserved targets. While compounds with known activity against a human protein could be compromised by toxicity, and therefore should be eliminated from further study, empirical evidence suggests that existence of a human orthologous sequences is not a strong filter for selecting pathogen targets. Many clinically used antibiotics have targets with human orthologs, such quinolones (DNA gyrase and topoisomerases), rifampicin (RNA polymerase), mupirocin (isoleucyl-tRNA synthetase) and the latest anti-TB drug now in Phase II testing, bedaquiline (F_1F_0 ATPase) [4,6]. The associated side effects of antibiotics are mostly due to high doses treatments affecting off-target proteins (including human ortologs) and not specifically to on-target effects. The billion plus years of evolutionary distance between prokaryotes and mammals has lead to significant divergence between orthologous proteins such that there is sufficient structure activity relationship or SAR bandwidth to develop specific inhibitors of the pathogen target, in our case MTB.

It is important to note that we also had a subset of compounds with historical data indicating activity against human protein targets with no known homologs in MTB, such as the GPCRs. Thus, their mechanism of action against MTB must be due to non-human target related interactions. These compounds must be pursued with caution as drug candidates given their known *in vitro* interaction with a human protein. Nevertheless, such compounds could be valuable tools for understanding MTB viability. In general, knowledge of potential human protein interactions adds to the design of effective counter-screens to drive compound SAR specificity and potency towards the pathogen.

The public availability of the data and compounds [7] as well as our predictions (http://www.tropicaldisaes.org/TCAMSTB/ or ftp://ftp.ebi.ac.uk/pub/databases/chembl/tb) will facilitate further research on drug discovery against tuberculosis. A major goal of our work is to encourage other researchers to experimentally validate the described targets and make their findings publicly available as soon as possible, thus optimizing the process of developing a safe and well tolerated novel therapy for tuberculosis.

Methods

Compound dataset

All compound datasets used in this study (that is, BCG dataset of 776 GSK compounds including the H37Rv sub-dataset of 177

Table 2. Significant links between GSK compound families and KEGG pathways.

GSK Family	Compound	Target	Pathways
1	GSK975784A	*Rv2182c*	**Glycerolipid metabolism (mtu00561)**
			Glycerophospholipid metabolism (mtu00564)
		Rv2483c	No Pathway
	GSK975810A	*Rv2182c*	**Glycerolipid metabolism (mtu00561)**
			Glycerophospholipid metabolism (mtu00564)
		Rv2483c	No Pathway
	GSK975839A	*Rv2182c*	**Glycerolipid metabolism (mtu00561)**
			Glycerophospholipid metabolism (mtu00564)
		Rv2483c	No Pathway
		Rv2299c	No Pathway
	GSK975840A	*Rv2182c*	**Glycerolipid metabolism (mtu00561)**
			Glycerophospholipid metabolism (mtu00564)
		Rv2483c	No Pathway
	GSK975842A	*Rv2182c*	**Glycerolipid metabolism (mtu00561)**
			Glycerophospholipid metabolism (mtu00564)
		Rv2483c	No Pathway
		Rv2045c	No Pathway
		Rv2139	Pyrimidine metabolism (mtu00240)
		Rv2299c	No Pathway
		Rv2483c	No Pathway
3	GSK547481A	Rv0194	**ABC transporters (mtu02010)**
	GSK547490A	Rv0194	**ABC transporters (mtu02010)**
	GSK547491A	Rv0194	**ABC transporters (mtu02010)**
	GSK547499A	Rv0194	**ABC transporters (mtu02010)**
	GSK547500A	Rv0194	**ABC transporters (mtu02010)**
	GSK547511A	Rv0194	**ABC transporters (mtu02010)**
	GSK547512A	Rv0194	**ABC transporters (mtu02010)**
	GSK547527A	*Rv1640c*	**Aminoacyl-tRNA biosynthesis (mtu00970)**
		Rv3598c	**Aminoacyl-tRNA biosynthesis (mtu00970)**
		Rv0194	**ABC transporters (mtu02010)**
	GSK547528A	*Rv1640c*	**Aminoacyl-tRNA biosynthesis (mtu00970)**
		Rv3598c	**Aminoacyl-tRNA biosynthesis (mtu00970)**
		Rv0194	**ABC transporters (mtu02010)**
	GSK547543A	Rv0194	**ABC transporters (mtu02010)**
7	GSK1829727A	Rv0053	**Ribosome (mtu03010)**
		Rv0379	No Pathway
		Rv0650	Glycolysis/Gluconeogenesis (mtu00010)
			Galactose metabolism (mtu00052)
			Starch and sucrose metabolism (mtu00500)
			Amino sugar & nucl. sugar metab. (mtu00520)
			Streptomycin biosynthesis (mtu00521)

Table 2. Cont.

GSK Family	Compound	Target	Pathways
	GSK1829729A	*Rv3855*	No Pathway
		Rv0053	**Ribosome (mtu03010)**
		Rv0379	No Pathway
		Rv0650	Glycolysis/Gluconeogenesis (mtu00010)
			Galactose metabolism (mtu00052)
			Starch and sucrose metabolism (mtu00500)
			Amino sugar & nucl. sugar metab. (mtu00520)
			Streptomycin biosynthesis (mtu00521)
	GSK1829816A	Rv0053	**Ribosome (mtu03010)**
		Rv0379	No Pathway
		Rv0650	Glycolysis/Gluconeogenesis (mtu00010)
			Galactose metabolism (mtu00052)
			Starch and sucrose metabolism (mtu00500)
			Amino sugar & nucl. sugar metab. (mtu00520)
			Streptomycin biosynthesis (mtu00521)
	GSK479031A	Rv0053	**Ribosome (mtu03010)**
		Rv0379	NoPathway (mtu00000)
		Rv0650	Glycolysis/Gluconeogenesis (mtu00010)
			Galactose metabolism (mtu00052)
			Starch and sucrose metabolism (mtu00500)
			Amino sugar & nucl. sugar metab. (mtu00520)
			Streptomycin biosynthesis (mtu00521)
	GSK957094A	Rv3170	Gly, Ser and Thr metabolism (mtu00260)
			Arginine and proline metabolism (mtu00330)
			Histidine metabolism (mtu00340)
			Tyrosine metabolism (mtu00350)
			Phenylalanine metabolism (mtu00360)
			Tryptophan metabolism (mtu00380)
		Rv0053	**Ribosome (mtu03010)**
		Rv0379	No Pathway
		Rv0650	Glycolysis/Gluconeogenesis (mtu00010)
			Galactose metabolism (mtu00052)
			Starch and sucrose metabolism (mtu00500)
			Amino sugar & nucl. sugar metab. (mtu00520)
			Streptomycin biosynthesis (mtu00521)
9	GSK1188379A	Rv0194	**ABC transporters (mtu02010)**
	GSK1188380A	Rv0194	**ABC transporters (mtu02010)**
16	GSK1825940A	Rv0194	**ABC transporters (mtu02010)**
	GSK1825944A	Rv0194	**ABC transporters (mtu02010)**
35	BRL-10143SA	*Rv1649*	Aminoacyl-tRNA biosynthesis (mtu00970)
		Rv2763c	**One carbon pool by folate (mtu00670)**
			Folate biosynthesis (mtu00790)
			One carbon pool by folate (mtu00670)
		Rv2764c	Pyrimidine metabolism (mtu00240)
	BRL-51093AM	Rv2763c	**One carbon pool by folate (mtu00670)**
		Rv2764c	Folate biosynthesis (mtu00790)
			One carbon pool by folate (mtu00670)

Table 2. Cont.

GSK Family	Compound	Target	Pathways
			Pyrimidine metabolism (mtu00240)
173	GSK1402290A	*Rv1640c*	**Aminoacyl-tRNA biosynthesis (mtu00970)**
		Rv3598c	**Aminoacyl-tRNA biosynthesis (mtu00970)**
		Rv3834c	**Aminoacyl-tRNA biosynthesis (mtu00970)**
		Rv3105c	No Pathway
		Rv3135	No Pathway
334	GSK270671A	*Rv1284*	**Nitrogen metabolism (mtu00910)**
		Rv3588c	**Nitrogen metabolism (mtu00910)**
		Rv3273	**Nitrogen metabolism (mtu00910)**
		Rv1707	No Pathway

Target genes in italics are either *in vivo* or *in vitro* essential in the TraSH Essentiality database [21]. Pathways highlighted in bold are responsible of the significant link to the GSK family.

doi:10.1371/journal.pcbi.1003253.t002

compounds) were obtained directly from the ChEMBL database (as deposition set http://dx.doi.org/10.6019/CHEMBL2095176). Chemical properties of the compounds (Figure 1) were calculated as previously described [12].

Exploring the chemogenomics space

A multi-category Naïve Bayesian classifier (MCNBC) was built using structural and bioactivity information from the ChEMBL database (version 14) [16]. In brief, the classifier learns the various classes (in this case protein targets) by considering the frequency of occurrence of certain sub-structural features for the different chemical compounds. Given a new, unseen compound, the model calculates a Bayesian probability score based on the molecule's individual features and produces a ranked list of likely targets. The model was built in Accelrys Pipeline Pilot (version 8.5). The structure and bioactivity data were extracted from the ChEMBL database and conformed the following filters: (i) the activity value was better than 10 uM (pIC50>5), (ii) the target type was a protein, (iii) the activity type was IC50, Ki or EC50, and (iv) the target confidence score was above 7.0. The last filter ensured that there was a reported direct interaction between the ligand and the protein target. The script resulted in 489,056 distinct compound-target pairs. To increase the robustness of the model, only targets with 40 or more active compounds were considered further, thus reducing the number of unique compound-target pairs to 466,686, spanning 1,258 distinct targets and 271,918 distinct compounds.

Two multiple-category models were subsequently built. Firstly, a model was created by choosing at random 85% of the compound records as the training set, so that the remaining 15% could be used as a test set for model validation, ensuring no overlapping structures in the 85-15 partition [17]. The MCNBC trained on 85% of the 271,918 ChEMBL compounds and associated targets was then used to predict the targets for the remaining 15% of the ChEMBL subset, containing 40,788 distinct compounds, unseen by the model. Standard ECFP_6 fingerprints were employed as molecular descriptors for the classifier [26]. These fingerprints encode a molecular structure as a series of overlapping features/fragments of a diameter of up to three bond lengths.

For each compound in the test set, the Pipeline Pilot model generated a likelihood score P_{total} for all possible targets. This is derived by the Laplacian-corrected Bayes rule of conditional probability $P(A|F_i)$ for each fingerprint feature i of the compound.

$$P_i(A|F_i) = (A_{Fi} + 1)/[T_{Fi}(A/T) + 1]$$

$$P_{total} = \log \Pi(P_i(A|F_i)) = \sum \log P_i(A|F_i)$$

where F_i is the i^{th} fingerprint feature; A is the number of active molecules for a target; T is the total number of molecules; A_{Fi} is the number of *active* molecules containing feature i; and T_{Fi} is the number of all molecules containing feature i.

For the purposes of this validation, only the top five target predictions were considered (*i.e.*, the ones with the highest positive likelihood score). This reflects a real-life situation where only a small number of target predictions can be practically and economically tested experimentally. To test the accuracy of the method, the five target predictions were then compared to the actual target reported for that particular compound.

The model derived by the training set ranked the correct target highest among all 1,258 possible targets for 82% of the compounds in the test set (Figure 6A). The target is correctly predicted on the second guess for 6% of the compounds and correctly predicted on the third guess for 2% of the compounds. In total, 92% of the compounds in the test set are correctly assigned to their known targets within the top five predicted targets. The ChEMBL database groups most of the individual protein targets into a hierarchical classification of target family names. Given this information, further analysis was done to examine the accuracy of the target classification predictions. Individual targets were replaced by their respective protein classification annotation using a lookup dictionary. In total, 568 unique protein classification labels were considered. The model's predictive power improves, returning the correct protein family as the top ranked prediction in 88% of the compounds and within the top five predictions in 94% of the compounds (Figure 6A). After the successful validation of the method, a second model was created utilizing 100% of the data and keeping the rest of the parameters intact. The derived model was then used for predicting the targets of all GSK compounds.

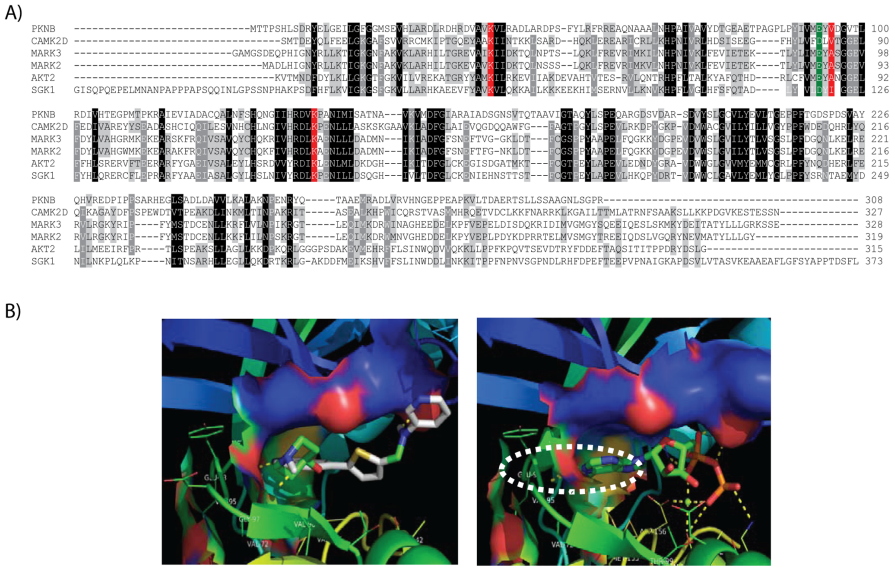


Figure 4. PknB kinase docking to GSK1598164A. A) Multiple sequence alignment of *Mycobacterium* PknB kinase with selected human kinases. Human kinases were selected on the criteria of having available PDB structures and top Psi-BLAST scores to *M. bovis* transmembrane serine/threonine-protein kinase B (pknB). First sequence in the alignment (gene name; PDB identifier) is *M. tuberculosis* transmembrane serine/threonine-protein kinase B (PknB; 3F69), which is 99% identical to *M. bovis* PknB and was used in compound docking models. Other sequences are CAMK2D (2EWL), MARK3 (2QNJ), MARK2 (3IEC), AKT2 (1GZK) and SGK1 (2R5T). Residues known to interact with ADP in pknB are highlighted in red. The amino acids aligned with Glu93, which may be essential for the binding of the GSK1132084A, are highlighted in green. **B**) Binding models of the GSK1598164A and ADP within pknB binding site (left and right panels, respectively).
doi:10.1371/journal.pcbi.1003253.g004

Exploring the structural space

A network of structural similarities between compounds and targets was built to identify the most likely target of a given compounds in our GSK dataset. To explore the structural space we used an improved version of our previously published AnnoLyze algorithm [14], which was based on homology detection through structural superimposition of targets and their interaction networks to small compounds similarly to previously published approaches [27,28]. Briefly, the new nAnnoLyze algorithm relies in four pre-built layers of interconnected networks, First, the "GSK Ligand" network where nodes are GSK compounds and edges correspond to their similarity as measured by a Random Forest classifier score (RFS) (see below). Second, the "PDB Ligand" network where nodes are ligands in the Protein Data Bank (PDB) [29] and edges correspond to their similarity also measured by the RFS. The "GSK Ligand" network is linked to the "PDB ligand" network by edges corresponding to the compound similarity measure by the RFS. Third, the "PDB Protein" network where nodes are proteins in PDB and edges corresponds to their structural similarity as measured with the MAMMOTH structural superimposition [30]. Fourth, the "MTB Models" network where nodes are structure models of MTB targets and edges corresponds to their structural similarity after superimposition by the MAMMOTH program. The two central networks (that is, "PDB Ligand" and "PDB Protein" networks) are connected by co-appearance in any solved structure in the PDB and the "PDB Protein" and the "MTB Models" networks are also linked by the structural comparison between any protein in the PDB and all models from MTB. Finally, once all the networks are constructed, we identified the closest path between any GSK compounds and a MTB target and scored their relationship as the sum of all similarities scores in the network. Such score was then normalized between 0 (non-similar) and 1 (similar) and only pairs of GSK compounds and their MTB targets with scores higher than 0.4 were kept.

To identify whether two compounds could be considered similar, we developed a new Random Forest classifier (RFS), which was trained with a dataset of "similar" and "non-similar" ligands. Two ligands were similar if they bind the same binding site as defined by the LigASite database, a gold-standard dataset of biologically relevant binding sites in protein structures [31]. To avoid overestimation in the validation of our approach, all ligands in the database that were included in a testing set of 2,380 ligands from the PDB were removed. Our training set of similar ligands included 197 pairwise comparisons considered as "true similar" and a set of randomly paired ligands as "true non-similar" comparisons. The SMSD program [32] was then used to compare all pair of selected ligands to obtain their Tanimoto score, bond

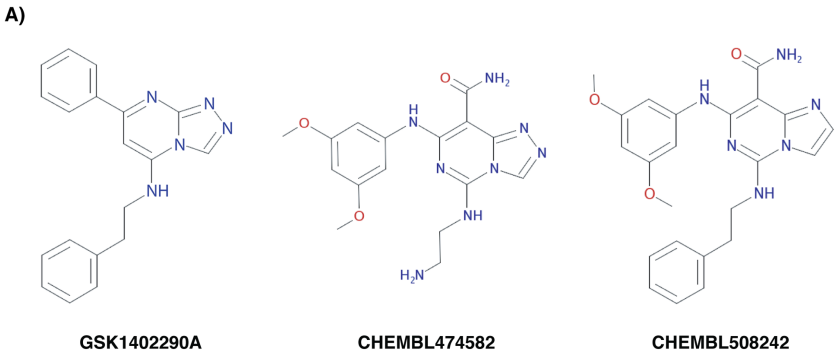


Figure 5. Targeting the aminoacyl-tRNA biosynthesis pathway. A) CHEM results show that GSK1402290A shared several substructural features with compounds reported as potent lysyl-tRNA synthetase inhibitors in the ChEMBL database (*e.g.,* CHEMBL474582 and CHEMBL508242). **B**) STR results predicted the serS as a target of GSK1402290A with its binding site including residues F205, H209, G225, T226, E228, R257, F276, K278, and E280, which are conserved in the PFAM family PF00587 (tRNA synthetase class II core domain). Zoomed image shows the pose for GSK1402290A predicted by AutoDock and the binding site residues (i.e., within 6 Å from the compound) coloured from low sequence conservation (blue) to high sequence conservation (red).
doi:10.1371/journal.pcbi.1003253.g005

breaking energy, Euclidian distance for equivalent atoms, stereochemical match, substructure fragment size, and finally the molecular weight difference. Such scores were then normalized and constituted a vector defining the similarity between any two compared ligands, which was then used as input for the Random Forest classifier. The aim of the classifier was thus to identify hidden relationships between the six scores to maximize its capacity to identify true pairs of similar ligands and discern them from non-similar ligands. The classifier was then tested with a 10-fold cross validation procedure and resulted in an area under the ROC curve of 0.97 and a very small false positive rate of 1.6% (Figure 6B).

To populate the "MTB Model" network with structures of MTB targets, we built all possible comparative structure models for any protein in the *M. tuberculosis* H37Rv, *M. bovis* BCG, and *M. smegmatis* genomes using the ModPipe program [25]. All sequences were obtained from the Genomes Web site of the NCBI database. Such modeling resulted in a total of 34,894 comparative models for which 5,008 were predicted to be reliable models (that is, 1.1 or higher ModPipe quality score and ga341 higher than 0.7). Next, we structurally compared this set of selected models to any non-redundant (90% sequence identity) structure in the PDB that contained at least one known ligand. Structural comparisons between two proteins were performed using the MAMMOTH algorithm [30]. Four different scores were stored for each structural superimposition: percentage of sequence and structure identity for the entire protein and percentage of sequence and structure identity for the residues involved in the binding site defined as any residue in the PDB template structure within 6 Ångstroms of any atom in the ligand. A binding site in a model was considered then similar to a binding site in a known PDB structure if at least the binding site sequence and structure

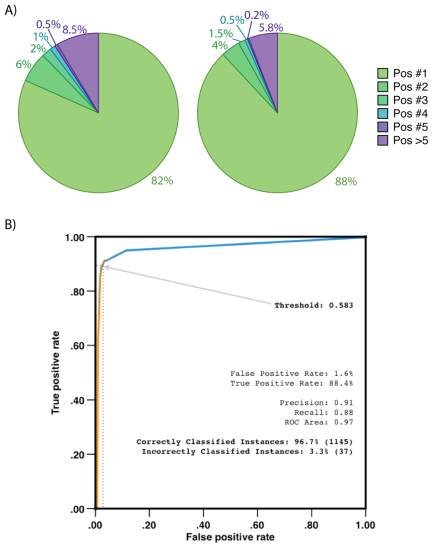


Figure 6. Predictive accuracy of the CHEM and STR methods. A) Predictive power of the MCNBC model using individual targets (left) or target classification information (right). **B**) Accuracy of the RFS differentiating similar from non-similar pairs of ligands. ROC curve indicates the optimal threshold for the RFS score of 0.58, which results in an area under the curve of 0.97 and a false positive rate of only 1.6%.
doi:10.1371/journal.pcbi.1003253.g006

similarity were higher than 40%. This similarity cut-off was previously validated in a large-scale comparison of known ligand-protein pairs [14].

The final entire network of comparisons included the 776 compounds from the GSK dataset, ~2.500 unique ligands from the PDB, ~16,000 unique protein structures from the PDB and a total of ~5,000 structure models from MTB. Such network resulted in 207 pairs of GSK compound to MTB target short paths (*i.e.*, score >0.4).

Exploring historical assay data

GSK proprietary compound screening databases were queried for any historical assay data associated with both *Mycobacterium* species active compounds. The majority of these screens were against human protein targets. The threshold above which compound efficacy against specific human targets was considered significant was defined as pIC50≥5.0 for inhibition or antagonist assays, and pEC50≥5.0 for agonist, activation or modulator assays. Activities at more than 600 target-result type combinations (some targets are assayed in both an antagonist and agonist mode) were analyzed amongst the BCG and H37Rv active compounds, representing potential modes of action. The target activities for the screened compounds were analyzed to identify targets over-represented amongst the anti-malarial actives vs. inactives.

Using BLASTP [33] we queried the protein complement of published MTB H37Rv and *M. bovis* BCG genomes with RefSeq proteins [34] for all human targets accepting a homology cut-off of an E-value ≤1.0e-10 and visual inspection of the alignments. Putative homologous relationships were confirmed by reciprocal BLASTP searches of identified *Mycobacterium* homologues against the human RefSeq protein databases. Initial multiple sequence alignments were performed using the program CLUSTALW v1.8 [35] with default settings and subsequently refined manually using the program SEQLAB of the GCG Wisconsin Package v11.0 software package (Accelrys, San Diego, CA, USA).

Statistical assessment of predicted links between compounds and targets

We measured two different statistics to assess the significance of a particular link between a chemical compound and a target pathway. Firstly, we calculated the LogOdds (that is, the odds of an observation given its probability). A feature i (in our case, a compound in Figure 2C or a pathway in Figure 3B) has a probability ($p_{i,c}$) in the entire dataset and a probability ($p_{i,r}$) of being at the subset of selected compounds/pathways. Their LogOdds are defined as the logarithm of its Odds (O_i):

$$O_i = \frac{p_{i,c}}{(1-p_{i,c})} \Big/ \frac{p_{i,r}}{(1-p_{i,r})}$$

Therefore, Odds higher than 1 (or positive LogOdds) indicate over-occurrence of the compound/pathway in the selected subset. Odds smaller than 1 (or negative LogOdds) indicate under-representation of the compound/pathway in the selected subset. Secondly, a *p-value* score was calculated for each predicted link between a compound and a target pathway using a Fisher's exact test for 2×2 contingency tables comparing two groups of annotations (*i.e.*, the group of compounds in a given pathway and the group of compounds in the entire dataset) [36].

Computational docking of compound in the structure of selected targets

Autodock 4.2 was used for docking studies [37]. The *ga_num_evals* were set at 250,000 to balance docking performance and CPU consumption. Thirty replicates were run for each chemical-protein pair and the binding conformation with the lowest docking score was chosen for visualization using PyMOL.

Supporting Information

Figure S1 Target class space. A) For positive hits in *M. tuberculosis* H37Rv screens, the distribution of human target classes affected by compounds based on known human protein potency and selectivity criteria as described in the text. The number of human targets is indicated for each class as well as the potential number of homologous genes (in parentheses). B) Distribution of 49 compounds screened against 1 or more targets having pIC50 or pEC50 values >5.5 in 120 assays by human target classes. Some compounds have historical assay information and potency against multiple target classes. Also indicated is the number of assays against targets with putative homologues in *M. tuberculosis* (in parentheses). C) Similar analysis of human target classes and D) 240 compounds in 642 assays for *M. bovis* BCG screens.
(DOCX)

Table S1 Predicted M. tuberculosis H37Rv and M. bovis BCG gene targets based on homology to human target assays.
(DOCX)

Table S2 Docking scores for the active compounds across two pknB structure models.
(DOCX)

References

1. WHOTuber2012 (Last accessed January 2013) Global Tuberculosis Report 2012.
2. Connell DW, Berry M, Cooke G, Kon OM (2011) Update on tuberculosis: TB in the early 21st century. Eur Respir Rev 20: 71–84.
3. Berry M, Kon OM (2009) Multidrug- and extensively drug-resistant tuberculosis: an emerging threat. Eur Respir Rev 18: 195–197.
4. Lewis K (2013) Platforms for antibiotic discovery. Nat Rev Drug Discov 12: 371–387.
5. Payne DJ, Gwynn MN, Holmes DJ, Pompliano DL (2007) Drugs for bad bugs: confronting the challenges of antibacterial discovery. Nat Rev Drug Discov 6: 29–40.
6. Roemer T, Boone C (2013) Systems-level antimicrobial drug and drug synergy discovery. Nat Chem Biol 9: 222–231.
7. Ballell L, Bates RH, Young RJ, Alvarez-Gomez D, Alvarez-Ruiz E, et al. (2013) Fueling Open-Source Drug Discovery: 177 Small-Molecule Leads against Tuberculosis. ChemMedChem 8(2):313–21.
8. Ballell L, Nathan CF, Marti-Renom MA, Davies G, Green J, et al. (2013) Opening Doors for Drug Discovery Against Tuberculosis. Science Translational Medicine in press.
9. Orti L, Carbajo RJ, Pieper U, Eswar N, Maurer SM, et al. (2009) A kernel for open source drug discovery in tropical diseases. PLoS Negl Trop Dis 3: e418.
10. Kepler T, Marti-Renom MA, Maurer SM, Rai AK, Taylor G, et al. (2006) Open Source Research - The power of Us. Aust J Chem 59: 291–294.
11. Munos B (2006) Can open-source R&D reinvigorate drug research? Nat Rev Drug Discov 5: 723–729.
12. Bickerton GR, Paolini GV, Besnard J, Muresan S, Hopkins AL (2012) Quantifying the chemical beauty of drugs. Nat Chem 4: 90–98.
13. Ghose AK, Crippen GM (1987) Atomic physicochemical parameters for three-dimensional-structure-directed quantitative structure-activity relationships. 2. Modeling dispersive and hydrophobic interactions. J Chem Inf Comput Sci 27: 21–35.
14. Marti-Renom MA, Rossi A, Al-Shahrour F, Davis FP, Pieper U, et al. (2007) The AnnoLite and AnnoLyze programs for comparative annotation of protein structures. BMC Bioinformatics 8 Suppl 4: S4.
15. Shannon P, Markiel A, Ozier O, Baliga NS, Wang JT, et al. (2003) Cytoscape: a software environment for integrated models of biomolecular interaction networks. Genome Res 13: 2498–2504.
16. Gaulton A, Bellis LJ, Bento AP, Chambers J, Davies M, et al. (2012) ChEMBL: a large-scale bioactivity database for drug discovery. Nucleic Acids Res 40: D1100–1107.
17. Nidhi, Glick M, Davies JW, Jenkins JL (2006) Prediction of biological targets for compounds using multiple-category Bayesian models trained on chemogenomics databases. J Chem Inf Model 46: 1124–1133.
18. Koutsoukas A, Simms B, Kirchmair J, Bond PJ, Whitmore AV, et al. (2011) From in silico target prediction to multi-target drug design: current databases, methods and applications. J Proteomics 74: 2554–2574.
19. Chen B, Harrison RF, Papadatos G, Willett P, Wood DJ, et al. (2007) Evaluation of machine-learning methods for ligand-based virtual screening. J Comput Aided Mol Des 21: 53–62.
20. Kolb P, Ferreira RS, Irwin JJ, Shoichet BK (2009) Docking and chemoinformatic screens for new ligands and targets. Curr Opin Biotechnol 20: 429–436.
21. Sassetti CM, Boyd DH, Rubin EJ (2003) Genes required for mycobacterial growth defined by high density mutagenesis. Mol Microbiol 48: 77–84.
22. Murphy DJ, Brown JR (2007) Identification of gene targets against dormant phase Mycobacterium tuberculosis infections. BMC Infect Dis 7: 84.
23. Ogata H, Goto S, Sato K, Fujibuchi W, Bono H, et al. (1999) KEGG: Kyoto Encyclopedia of Genes and Genomes. Nucleic Acids Res 27: 29–34.
24. Li L, Stoeckert CJ, Jr., Roos DS (2003) OrthoMCL: identification of ortholog groups for eukaryotic genomes. Genome Res 13: 2178–2189.
25. Eswar N, John B, Mirkovic N, Fiser A, Ilyin VA, et al. (2003) Tools for comparative protein structure modeling and analysis. Nucleic Acids Res 31: 3375–3380.
26. Rogers D, Hahn M (2010) Extended-connectivity fingerprints. J Chem Inf Model 50: 742–754.
27. Kalinina OV, Wichmann O, Apic G, Russell RB (2011) Combinations of protein-chemical complex structures reveal new targets for established drugs. PLoS Comput Biol 7: e1002043.
28. Kalinina OV, Wichmann O, Apic G, Russell RB (2012) ProtChemSI: a network of protein-chemical structural interactions. Nucleic Acids Res 40: D549–553.
29. Berman H, Henrick K, Nakamura H, Markley JL (2007) The worldwide Protein Data Bank (wwPDB): ensuring a single, uniform archive of PDB data. Nucleic Acids Res 35: D301–303.
30. Ortiz AR, Strauss CE, Olmea O (2002) MAMMOTH (matching molecular models obtained from theory): an automated method for model comparison. Protein Sci 11: 2606–2621.
31. Dessailly BH, Lensink MF, Orengo CA, Wodak SJ (2008) LigASite–a database of biologically relevant binding sites in proteins with known apo-structures. Nucleic Acids Res 36: D667–673.
32. Rahman SA, Bashton M, Holliday GL, Schrader R, Thornton JM (2009) Small Molecule Subgraph Detector (SMSD) toolkit. J Cheminform 1: 12.
33. Altschul SF, Madden TL, Schaffer AA, Zhang J, Zhang Z, et al. (1997) Gapped BLAST and PSI-BLAST: a new generation of protein database search programs. Nucleic Acids Res 25: 3389–3402.
34. Pruitt KD, Katz KS, Sicotte H, Maglott DR (2000) Introducing RefSeq and LocusLink: curated human genome resources at the NCBI. Trends Genet 16: 44–47.
35. Thompson JD, Higgins DG, Gibson TJ (1994) CLUSTAL W: improving the sensitivity of progressive multiple sequence alignment through sequence weighting, position-specific gap penalties and weight matrix choice. Nucleic Acids Res 22: 4673–4680.
36. Fisher RA (1922) On the interpretation of $\chi 2$ from contingency tables, and the calculation of P. Journal of the Royal Statistical Society 85: 87–94.
37. Morris GM, Huey R, Lindstrom W, Sanner MF, Belew RK, et al. (2009) AutoDock4 and AutoDockTools4: Automated docking with selective receptor flexibility. J Comput Chem 30: 2785–2791.

Acknowledgments

We thank LLuís Ballell and David Barros for their support in the initial stages of this work.

Author Contributions

Conceived and designed the experiments: JRB JPO MAMR. Performed the experiments: FMJ GP LY IMW VK UP. Analyzed the data: FMJ GP LY MAMR. Contributed reagents/materials/analysis tools: IMW VK UP. Wrote the paper: FMJ GP AS JRB JPO MAMR.

PLOS ONE

RESEARCH ARTICLE

Release of 50 new, drug-like compounds and their computational target predictions for open source anti-tubercular drug discovery

María Jose Rebollo-Lopez[1], Joël Lelièvre[1]*, Daniel Alvarez-Gomez[1], Julia Castro-Pichel[1], Francisco Martínez-Jiménez[2,3], George Papadatos[4], Vinod Kumar[5], Gonzalo Colmenarejo[6], Grace Mugumbate[4], Mark Hurle[5], Vanessa Barroso[9], Rob J. Young[7], María Martinez-Hoyos[1], Rubén González del Río[1], Robert H. Bates[1], Eva Maria Lopez-Roman[1], Alfonso Mendoza-Losana[1], James R. Brown[5], Emilio Alvarez-Ruiz[9], Marc A. Marti-Renom[2,3,8]*, John P. Overington[4], Nicholas Cammack[1], Lluís Ballell[1], David Barros-Aguire[1]

1 Diseases of the Developing World, GlaxoSmithKline, Tres Cantos, Madrid, Spain, **2** Genome Biology Group, Centre Nacional d'Anàlisi Genòmica (CNAG), Barcelona, Spain, **3** Gene Regulation Stem Cells and Cancer Program, Centre for Genomic Regulation (CRG), Barcelona, Spain, **4** European Molecular Biology Laboratory–European Bioinformatics Institute (EMBL-EBI), Hinxton, Cambridge, United Kingdom, **5** Computational Biology, Quantitative Sciences, GlaxoSmithKline, Collegeville, Pennsylvania, United States of America, **6** Centro de Investigación Básica, CSci Computational Chemistry, GlaxoSmithKline, Tres Cantos, Madrid, Spain, **7** CSC Medicinal Chemistry, Medicines Research Centre, GlaxoSmithKline, Stevenage, Hertfordshire, United Kingdom, **8** Institució Catalana de Recerca i Estudis Avançats (ICREA), Barcelona, Spain, **9** Centro de Investigación Básica, Platform Technology & Science, GlaxoSmithKline, Tres Cantos, Madrid, Spain

* joel.l.lelievre@gsk.com (JL); mmarti@pcb.ub.cat (MAMR)

Citation: Rebollo-Lopez MJ, Lelièvre J, Alvarez-Gomez D, Castro-Pichel J, Martínez-Jiménez F, Papadatos G, et al. (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

Editor: Anil Kumar Tyagi, University of Delhi, INDIA

Received: July 9, 2015

Accepted: October 19, 2015

Published: December 7, 2015

Data Availability Statement: The data on the 50 antitubercular compounds are available from the Dryad Digital Repository: http://dx.doi.org/10.5061/dryad.8r351.

Funding: The research behind these results received funding from the TB Alliance, the European Union's 7th framework programme (FP7-2007–2013) under grant agreement ORCHID no. 261378 and the ERA-NET Pathogenomics Project GeMoA (PIM2010EPA-00719). Those funders had no role in study design, data collection and analysis, decision to publish, or preparation of the manuscript. The funder

Abstract

As a follow up to the antimycobacterial screening exercise and the release of GSK´s first Tres Cantos Antimycobacterial Set (TCAMS-TB), this paper presents the results of a second antitubercular screening effort of two hundred and fifty thousand compounds recently added to the GSK collection. The compounds were further prioritized based on not only antitubercular potency but also on physicochemical characteristics. The 50 most attractive compounds were then progressed for evaluation in three different predictive computational biology algorithms based on structural similarity or GSK historical biological assay data in order to determine their possible mechanisms of action. This effort has resulted in the identification of novel compounds and their hypothesized targets that will hopefully fuel future TB drug discovery and target validation programs alike.

Introduction

Although the Millennium Development Goal (MDG) target to halt and reverse the Tuberculosis (TB) epidemic by 2015 has been achieved, the global burden of disease remains enormous. The World Health Organization (WHO) estimates that about one third of the world's population could be latently infected with Tuberculosis. Although the vast majority will not go on to

GlaxoSmithKline provided support in the form of salaries for authors [MJR, JL, DAG, JCP, GC, VB, RJY, MMH, RGR, RHB, EMLR, AML, JRB, EAR, NC. LB and DBA], but did not have any additional role in the study design, data collection and analysis, decision to publish, or preparation of the manuscript. The specific roles of these authors are articulated in the 'author contributions' section.

Competing Interests: The authors of this manuscript have the following competing interests: several authors are employed by GlaxoSmithKline. This does not alter their adherence to PLOS ONE policies on sharing data and materials.

develop TB, in 2012 there were 8.6 million new cases of active disease, 1.3 million of which resulted in death attributable to TB [1].

More worryingly, the increasing prevalence of Multi Drug Resistant (MDR) and Extensively Drug Resistant (XDR) TB highlights the shortcomings of the present therapeutic options against the disease [1]. According to WHO, in 2013, there were an estimated 480 000 new cases of MDR-TB worldwide, 9% of which were found to be practically untreatable XDR-TB. In 2014, XDR-TB has already been reported by 100 countries. Therefore, there is still an urgent need for new drugs with novel mechanisms of action, able to treat both MDR/XDR and drug sensitive TB patients in a cost effective way. To not tackle this challenge head on now will be at our own future peril.

To stimulate community-based research efforts towards the discovery of novel TB therapeutics, and as a follow-up to our previous release of 177 compounds into the public domain [2] as promising starting points for new TB medicines, we have recently screened the latest chemical diversity available within GSK compound collections and identified 50 novel, non-cytotoxic, high quality chemical starting points active against replicating *Mycobacterium tuberculosis*. The presentation of this data has been complemented with a multipronged computational analysis to predict the possible biological targets of each one of these molecules.

Materials and Methods

HTS ATP assay

While the resazurin-based method was a reliable way to test the phenotypic activity of antitubercular compounds, it was unfortunately unsuitable for HTS campaigns given the low signal-to-noise ratio and the frequent interference of fluorescent compounds. As an alternative to a resazurin-based readout, we used a commercially available system based on ATP measurement (BacTiter-Glo, Promega). This assay measured the effect of the compounds on bacterial growth by determining the amount of ATP per well, which is proportional to the number of living bacteria. The reagent caused bacterial cell lysis and generated a luminescent signal proportional to the amount of ATP present and thus to the number of viable cells in culture. The assay relied on the activity of a thermostable luciferase and on the properties of a buffer formulation for extracting ATP from bacteria.

Single Shot inhibition assay

***Mycobacterium bovis* BCG str. Pasteur 1173P2 (BCG).** Bacterial inocula were grown for 4–5 days in Middlebrook 7H9 medium (Difco cat. # 271310) with glucose as carbon source. The culture medium contained per liter: 4.7 g Middlebrook 7H9 powder, 5 g albumin, 1 g glucose, 0.85 g NaCl, and 0.25 g Tween 80. The solution was sterilized by filtration through a 0.2 mm filter. The HTS assay was carried out in 1536-well sterile plates (Greiner, 782074). The screening compounds were added to the plates as a 50 nL solution in neat DMSO (Sigma, D8418) prior to addition of the assay components by using an Echo 555 instrument (Labcyte Inc). The assay plates were subsequently filled with 5 μL of the bacterial solution (adjusted to 10^5 bacteria per mL) using a Multidrop Combi NL instrument (Thermo Fischer Scientific Inc.). Inoculated plates were stacked in groups of 7–8 plates, with the top plate covered with a sterile lid. Plates were carefully wrapped with aluminum foil to prevent evaporation and allowed to incubate at 37°C at 80% relative humidity for seven days. After the incubation period, plates were removed from the incubator and allowed to equilibrate at room temperature. Freshly reconstituted BacTiter-Glo (5 μL, Promega) was added to each well using the Multidrop Combi. After standing at room temperature for 7–8 min, the luminescence signal was quantified with an Acquest reader (Molecular Devices) in the focused luminescence mode.

Every assay plate contained two columns of negative controls (control 1) with DMSO, which correspond to 100% activity reactions (maximum luminescence), and two columns of positive controls (control 2) in which 100% inhibition was reached by adding a known inhibitor (2 μM rifampicin as standard; bacterial growth completely inhibited). These controls were used to monitor assay quality through determination of Z´ as well as for normalizing the data on a per-plate basis. The effect of a given compound is calculated as: % Inhib. = 100 x [(data—ctrl 1)/(ctrl2—ctrl 1)].

***Mycobacterium tuberculosis* H37Rv (*M. tuberculosis*).** For *M. tuberculosis*, the HTS assay was carried out in sterile 384-well white microtest plates TC surface (353988 BC Falcon). 250 nL of screening compound were added to the plates as a solution in neat DMSO. The inoculum was standardized to 10^7 CFU/mL by measuring the OD at 600nm (an OD600 = 0.125 is equivalent to 10^7 CFU/mL) and then diluted 1 in 100 (10^5 CFU/mL) in 4.7 g Middlebrook 7H9 powder, 5 g albumin, 1 g glucose, 0.85 g NaCl, and 0.025% Tyloxapol (Sigma T8761). 25 μL of the 10^5 CFU/mL solution were dispensed in all 384w compound plates. Every assay plate contained one column of negative control (control 1, 6^{th} column) with neat DMSO and one column of positive control (control 2, 18^{th} column) in which 100% inhibition was reached by adding a known inhibitor (0.1 mg/mL of rifampicin, Sigma R3501). The incubation was as described previously [2]. This time, 10 μL of reconstituted BacTiter-Glo™ Microbial Cell viability Assay (Promega, G8231) reagent was added to each well and the plate was left 30 min at room temperature. The luminescence was measured using the Spectramax M5 (Molecular Devices) with integration time 250 mseconds (endpoint).

pIC_{50} ATP assay (*M. bovis* BCG Pasteur and *M. tuberculosis* H37Rv)

The assay was performed in 384 well plates for *M. tuberculosis* H37Rv and in 1536 well plates for *M. bovis* BCG Pasteur. For each compound, 11 two-fold dilutions were done in DMSO (final concentration 1%). The controls were as the ones used for the *M. tuberculosis H37Rv* Single Shot ATP assay. The method used (inocula, incubation, measurement) is the same as in the *M. tuberculosis* H37Rv Single Shot ATP assay, maintaining 8 min incubation, once the BacTiter-Glo™ is added, in the case of *M. bovis BCG Pasteur* and 30 min for *M. tuberculosis H37Rv*.

The effect of a given compound was calculated as % inhibition at single shot or pIC_{50} (ActivityBase, ID Business Solutions Limited). Zprime lower limit had been established at 0.4. Plates with Zprime values below this cutoff were rejected.

Statistical analysis for HTS

Statistical cutoffs were obtained as the mean plus 3 standard deviations calculated with a robust algorithm [3] from the population of growth inhibitions; compounds above the statistical cutoffs were deemed to have significant inhibition compared to the majority of the compounds that had inhibitions within the noise. A cutoff was calculated for each batch of plates tested in one day. Previously the plates were corrected for the presence of patterns when necessary by using an in-house developed plate pattern recognition and fixing algorithm [4].

Pattern Correction

The plates that display gradient patterns were fixed with the Pattern Recognition & Fixing Algorithm. The algorithm corrects the responses by calculating a robust 2D running median across the wells and performs a weighted subtraction from the original responses such as it leaves unmodified the "outlier" responses. Plates with VEP (variance explained by pattern) > 0.35 were deemed in need of pattern fixing.

Activity against Non-replicating *Mycobacterium tuberculosis*

Non-Replicating (NR) conditions were induced in Sauton's-based minimal containing 0.05% KH_2PO_4, 0.05% $MgSO_4$, 0.005% ferric ammonium citrate, 0.00001% $ZnSO_4$ and 0.01% NH_4Cl supplemented with 0.05% butyrate, 0.5% BSA, 0.085% NaCl and 0.02% tyloxapol. pH was set to 5.0 with 2N NaOH and $NaNO_2$ was added from freshly prepared 1M stock (in distilled H_2O) to a final concentration of 0.5mM. NR conditions also included 1% O_2 and 5% CO_2.

Assay conditions: 150 nL of each compound at 1mM concentration (in 100% DMSO) dispensed in 384 well plates and stored at -20°C. Bacterial pellets obtained from log-phase *M. tuberculosis* H37Rv grown in roller bottles at 37°C and 20% O_2 were washed twice with phosphate buffer saline (PBS; Difco), which had 0.02% tyloxapol (PBS-Tyloxapol). Bacterial suspension with an OD of 0.1 at 580 nm was then prepared in NR medium and $NaNO_2$ added fresh for a final concentration of 0.5 mM. 15 μL of this suspension was dispensed in to each well of the compound plate. Plates were incubated for 3 days at 37°C in oxygen-controlled incubators at 1% O_2 and 5% CO_2. 60 μL of complete 7H9 medium was added to each well after NR exposure and the plates incubated at 37°C with 21% O2 and 5% CO_2 to allow outgrowth of bacteria. OD was read after 7 days using a microplate reader.

M. tuberculosis inhibition assay (MABA)–H37Rv and resistant strains

The measurement of the minimum inhibitory concentration (MIC) for each tested compound was performed in 96-well flat-bottom polystyrene microtiter plates. Ten twofold drug dilutions in neat DMSO starting at 5 mM were performed. These drug solutions (5 μL) were added to 95 μL Middlebrook 7H9 medium (lines A-H, rows 1–10 of the plate layout). Isoniazid was used as a positive control; eight twofold dilutions of isoniazid starting at 1.2 mM were prepared, and this control curve (5 μL) was added to 95 μL Middlebrook 7H9 medium (row 11, lines A-H). Neat DMSO (5 μL) was added to row 12 (growth and blank controls). The inoculums were standardized to ~$1x10^7$ $CFUmL^{-1}$ and diluted 1:100 in Middlebrook 7H9 broth (Middlebrook ADC enrichment, a dehydrated culture medium which supports growth of mycobacterial species, available from Becton–Dickinson, cat. # 211887), to produce the final inoculum of H37Rv strain (ATCC25618) and resistant clinical isolates to isoniazid and rifampicin. This inoculum (100 mL) was added to the entire plate except G-12 and H-12 wells (blank controls). All plates were placed in a sealed box to prevent drying out of the peripheral wells and were incubated at 37°C without shaking for six days. A resazurin solution was prepared by dissolving one tablet of resazurin (VWR International Ltd., Resazurin Tablets for Milk Testing, cat.# 330884Y') in 30 mL sterile phosphate-buffered saline (PBS). Of this solution, 25 μL were added to each well. Fluorescence was measured (Spectramax M5, Molecular Devices; lex 530 nm, lem 590 nm, cut-off 570 nm) after 48 h to determine the MIC value.

Intracellular assay

M. tuberculosis H37Rv containing the Photinus pyralis luciferase gene (Hygromicin resistant plasmid) was grown in 7H9 suplemented with 10% ADC and 0.05% Tyloxapol until the OD600 is 0.5–0.8. We divided 160 ml of culture in 4 tubes of 50 ml each and pelleted at 2860g for 10 min. 10 glass beads (4mm) were added in order to disperse the bacterial pellet of each tube by shaking for 60 seconds. Then 6 ml of fresh RPMI media were added and leave on the bench for 5 min. Carefully we collected 5 ml of the supernatant and discard the rest. The supernatants of 4 tubes were collected into the same sterile tube and centrifuged at 402g for 5 minutes to avoid any remaining clumps. This dispersed bacterial suspension was diluted into RPMI-0.05% Tyloxapol and we calculated the volume needed to have a multiplicity of infection (MOI) of 1, using the following conversion: OD600 0.1 = $1x10^7$ CFU/ml. THP1 cells (ATCC®

TIB-202 ™) were maintained in complete RPMI1640 (RPMI 1640 HEPES modification, 2 mM L-glutamine, 1 mM sodium pyruvate, 10% fetal bovine serum) and incubated at 37°C with 5% CO2. THP1 phagocytes ($2x10^5$ cell/mL) were infected for 4 h in a roller bottle with a MOI of 1 in RPMI-20nM PMA and extracellular bacteria were discarded by washing 5 times in complete RPMI (5 x 402g, 5 min). We dipensed 50 μL/well (10,000 cells/well) of infected THP1 cells in white 384-well plates with 250nl/ well of compound in DMSO.Plates were incubated for 5 days at 37°C/ 5% CO2. Then, 25μl of reconstituted Bright-Glo™ Luciferase Assay System (Promega) were added to each well and plates were incubated at RT for 30 minutes. Finally, the luminescence was read in an Envision system (Perkin Elmer) using these settings: US LUM 384 (cps) 7000004/ Measurement height 0 mm/ Measurement time 0.1 s. Aperture: 384 Plate US Luminescence aperture.

HepG2 cytotoxicity assay

Actively growing HepG2 cells were removed from a T-175 TC flask using 5 mL Eagle's MEM (containing 10% FBS, 1% NEAA, 1% penicillin/streptomycin) and dispersed in the medium by repeated pipetting. Seeding density was checked to ensure that new monolayers were not >50% confluent at the time of harvesting. Cell suspension was added to 500 mL of the same medium at a final density of $1.2x10^5$ cells.mL^{-1}. This cell suspension (25 μL, typically 3000 cells per well) was dispensed into the wells of 384-well clear-bottom plates (Greiner, cat. # 781091) using a Multidrop instrument. Prior to addition of the cell suspension, the screening compounds (250 nL) were dispensed into the plates with an Echo 555 instrument. Plates were allowed to incubate at 37°C at 80% relative humidity for 48 h under 5% CO_2. After the incubation period, the plates were allowed to equilibrate at room temperature for 30 min before proceeding to develop the luminescent signal. The signal developer, CellTiter-Glo (Promega) was equilibrated at room temperature for 30 min and added to the plates (25 μL per well) using a Multidrop. The plates were left for 10 min at room temperature for stabilization and were subsequently read using a ViewLux instrument (PerkinElmer).

The human biological samples were sourced ethically and their research use was in accord with the terms of the informed consents.

Physicochemical properties

CLND solubility assay. GSK in-house kinetic solubility assay: 5 μL of 10mM DMSO stock solution diluted to 100 uL with pH7.4 phosphate buffered saline, equilibrated for 1 hour at room temperature, filtered through Millipore Multiscreen HTS-PCF filter plates (MSSL BPC). The filtrate is quantified by suitably calibrated flow injection Chemi-Luminescent Nitrogen Detection [5]. The standard error of the CLND solubility determination is ±30 μM, the upper limit of the solubility is 500 μM when working from 10 mM DMSO stock solution.

ChromlogD assay. The Chromatographic Hydrophobicity Index (CHI) [6] values were measured using a reversed phase HPLC column (50 x 2 mm x 3 μM Gemini NX C18, Phenomenex, UK) with fast acetonitrile gradient at starting mobile phase of 100% pH = 7.4 buffer. CHI values are derived directly from the gradient retention times by using a calibration line obtained for standard compounds. The CHI value approximates to the volume % organic concentration when the compound elutes. CHI is linearly transformed into ChromlogD [7] by the formula: ChromlogD = 0.0857CHI-2.00. The average error of the assay is ±3 CHI unit or ±0.25 ChromlogD.

Exploring the 2D chemogenomics space

We applied a multi-category Naïve Bayesian classifier (MCNBC) that was built and trained using 2D structural and experimental bioactivity information from the ChEMBL database

version 16 [8]. In brief, the classifier learns the various categories (in this case protein targets) by considering the enrichment of certain 2D sub-structural features of active compounds across the protein targets. Given a new, unseen compound, the model calculates a Bayesian probability score for each target based on the compound's individual features and produces a ranked list of likely targets. The model was built in Accelrys Pipeline Pilot (version 8.5) using standard ECFP_6 fingerprints [9] to encode the chemical structures. Further information on the model generation and validation can be found in our previous publication [10]. The statistical significance of the probability scores was assessed with Z-scores. These were computed by calculating the background probability score distribution for each protein target using all the compounds in ChEMBL. Lastly, given that the majority of bioactivities in the ChEMBL database are against human, mouse and rat protein targets, the predicted targets were mapped to their orthologous *M. tuberculosis* ones using the OrthoMCL [11] database [12].

Exploring the 3D structural space

A network of 3D structural similarities between compounds and targets was built to identify the most likely targets of a given compound in the GSK dataset. To explore the structural space, we used nAnnolyze, an improved version of our previously published AnnoLyze algorithm [13], which was based on homology detection through structural superimposition of targets and their interaction networks to small compounds, similar to previously published approaches [14, 15]. Briefly, the new algorithm relies in four pre-built layers of interconnected networks. First, the "GSK Ligand" network where nodes are GSK compounds and edges correspond to their similarity as measured by a previously developed Random Forest Classifier (RFS) score [10]. The RFS classifier predicts whether two small molecules are likely to bind the same target-binding site by comparing their structural and chemical properties. Second, the "PDB Ligand" network where nodes are clusters of highly similar ligands of the Protein Data Bank (PDB) [16] and edges correspond to their similarity measured by the RFS. The "GSK Ligand" network is linked to the "PDB ligand" network by edges corresponding to the compound similarity measure by the RFS. Third, the "PDB Protein" network with nodes corresponding to clusters of highly similar small molecule binding-site of proteins in PDB and edges correspond to their structural similarity as measured with the ProBiS structural superimposition method [17]. Fourth, the previously built "*M. tuberculosis* Models" network [10] with nodes corresponding to predicted small molecules binding-sites in three-dimensional models of *M. tuberculosis* targets and edges correspond to their structural similarity after comparison by the ProBiS program. The two central networks (that is, "PDB Ligand" and "PDB Protein Binding-Sites" networks) are connected by co-appearance of the compound and the protein in any solved structure in the PDB. The "PDB Protein" and the "*M. tuberculosis* Models" networks are linked by the structural comparison between any binding-site in the PDB network and all binding-sites in models from *M. tuberculosis*. Finally, once all the networks are constructed, we identified the closest path between any GSK compounds and *M. tuberculosis* targets. To score the hit, we used the inverse of the edges weight of the pathway. Next, the final score is normalized to 1 (being 1 the best score and 0 the worst one) and Z-scored. Specifically, two different Z-scores are calculated for each prediction. The first, called Global Z-score, is obtained by running the predictions of all drugs present in DrugBank against all targets and using the global mean and the global standard deviation to Z-score specific predicted pair. The Global Z-score represents how good a prediction is given its score in the constructed network. The second, called Local Z-score, is calculated by running the predictions of all drugs present in DrugBank and retrieving the mean and the standard deviation of the score for a specific target binding-site. The Local Z-score represent how good a prediction is for a specific

binding-site; highly promiscuous binding-sites tend to have bad local Z-scores. The nAnnolyze approach was recently evaluated using all the FDA approved drugs present in PDB. In such dataset, the nAnnolyze predictions result in an area under the Receiver Operating Characteristic curve (AUC) of 0.70 [18].

The final entire network of comparisons included the 50 compounds from the GSK dataset, 7,609 unique ligands from the PDB, 28,299 unique compound binding-sites in protein structures from the PDB, and a total of 5,008 structure models from *M. tuberculosis*.

Exploring historical assay data

GSK proprietary compound screening databases were queried for any historical assay data associated with *M. tuberculosis* H37Rv (*M. tuberculosis* H37Rv) active compounds. The majority of these screens were against human protein targets. The threshold above which compound efficacy against specific human targets was considered significant was defined as $pIC_{50} \geq 5.0$ for inhibition or antagonist assays, $pEC_{50} \geq 5.0$ for agonist, activation or modulator assays (*i.e.* overall $pXC_{50} \geq 5.0$).

Using BLASTP [19] we queried the protein complement of published *M. tuberculosis* H37Rv for all human targets accepting a homology cutoff of an E-value $\leq$1.0e-10 and visual inspection of the alignments. Putative homologous relationships were confirmed by reciprocal BLASTP searches of identified *M. tuberculosis* H37Rv homologues against the human RefSeq protein databases (April 2014).

Statistical assessment of predicted links between compounds and targets

We measured two different statistics to assess the significance of a particular link between a chemical compound and a target pathway. Firstly, we calculated the LogOdds (that is, the odds of an observation given its probability). A feature *i* (in our case, a compound in or a pathway) has a probability ($p_{i,c}$) in the entire dataset and a probability ($p_{i,r}$) of being at the subset of selected compounds/pathways. Their LogOdds are defined as the logarithm of its Odds (O_i):

$$O_i = \frac{\frac{p_{i,c}}{(1-p_{i,c})}}{\frac{p_{i,r}}{(1-p_{i,r})}}$$

Therefore, Odds higher than 1 (or positive LogOdds) indicate over-occurrence of the compound/pathway in the selected subset. Odds smaller than 1 (or negative LogOdds) indicate under-representation of the compound/pathway in the selected subset. Secondly, a *p-value* score was calculated for each predicted link between a compound and a target pathway using a Fisher's exact test for 2×2 contingency tables comparing two groups of annotations (*i.e.*, the group of compounds in a given pathway and the group of compounds in the entire dataset) [20]

Results

Screening process and drug like properties of hits identified

GSK DDW has very recently added to its Corporate small molecule repository some structurally new 254,000 compounds, known as the "Top-up" library, whose diverse profile reflects the latest intelligence on how specific physicochemical property descriptors (sp^3 character, lipophilicity/ water solubility, molecular size *etc.*) can affect attrition at the different stages of the drug discovery phase. Given the differentiated structural profile of this compound library, we would

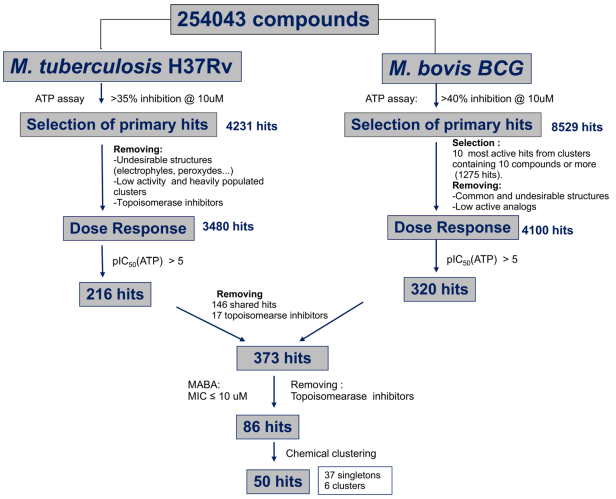


Fig 1. HTS progression cascade leading to 50 confirmed H37Rv-positive compounds.

doi:10.1371/journal.pone.0142293.g001

expect novel hits that engage new targets beyond those identified during our previous studies [2].

We have reported previously how non pathogenic and Biosafety Level 2 friendly *M. bovis* BCG can act as a modest surrogate to predict the antitubercular activity against *M. tuberculosis* H37Rv [2]. It is for this reason that, in this occasion, we decided to undertake parallel screening activities against both strains. Active compounds meeting the pre-established (Fig 1) threshold of activity in the primary ATP antimycobacterial assay, were progressed to evaluation in a resazurin based H37Rv assay. This resulted in the identification of 4,231 compounds that exerted an inhibition of *M. tuberculosis* H37Rv growth superior to 35% and, in the case of BCG, 8,529 compounds with growth inhibition values above 40%. At this stage a set of automatic filters directed towards the identification and elimination of a few remaining undesirable structural features such as electrophiles, peroxides and Michael acceptors, was applied resulting in a first reduction on the number hit structures; this set of filters is an updated version of a previously published one.[32] The fraction of undesirable compounds in the "Top up" library is very small (estimated in 0.08%); however, most of them (ca. 200) showed up in the initial list of hits as they are reactive compounds prone to cytotoxicity effects.

This first selection was further narrowed through the application of a specifically designed in house algorithm that helped prioritize highly active structural clusters and remove analogs showing lower percentages of inhibition. Finally, all known antitubercular classes were manually removed. The resulting 320 (BCG screening stream) and 216 (H37Rv screening stream) hits were checked for structural duplication, resulting in 373 compounds that were progressed to Minimum Inhibitory Concentration (MIC) determination against H37Rv in the MABA resazurin assay. In order to progress only the most promising hits, we applied very strict

selection criteria: we did not select any compound which either did not reach, at least, the 90% inhibition cutoff or required more than 10μM to reach this threshold. Therefore, only 86 of these 373 compounds were moved forward (MICs lower or equal to 10 μM). This relatively low percentage of actives in the H37Rv MABA assay highlights a lack of correlation between the two readouts employed in this screening effort. The corresponding 86 compounds were then clustered in chemotypes resulting in a final hit list of 50 representative highly potent compounds, including 37 singletons, five clusters of two representatives and one cluster of three representatives (Figs 1 and 2 and Table B in S1 File). When tested for MIC determination in the H37Rv MABA assay, these 50 compounds showed MIC values between 0.2 and 10 μM. Amongst those 50 hits, 7 were Mtb specific and the rest were all identified as hits in both screening campaigns (BCG and Mtb). Activity against *M. bovis* BCG (pIC50s) is described in Fig 2 and Table B in S1 File. In order to determine the therapeutic window of the hits, the HepG2 cytotoxicity of each hit was evaluated. From the dose–response results, 24 compounds displayed TOX_{50} between 10 and 100 μM and 26 had no detectable cytotoxic effects ($TOX_{50} \geq$ 100 μM). The library used in these screens is composed of lead-like compounds with very low lipophilicity. In addition, the few remaining compounds with reactive substructures were automatically removed (see above), as well as topoisomerase inhibitors that inhibit also eucariotic cells (Fig 1). As a result, the final 50 compounds have a very high probability of targetting *Mycobacterium* through specific mechanisms that would explain the low cytotoxicity observed in HepG2 cells.

All the 50 hits were also tested for their activity against non-replicating *M. tuberculosis* as described previously [21]. Interestingly, 5 of them were active with pIC_{50} ranging from 4.5 to 5.2 (equivalent to IC_{50}´s between 31 and 6 μM). Finally, the inhibition of the intracellular growth of mycobacterium tuberculosis was determined. Out of the 26 representative compounds tested, all but one (TCMDC-143682) were active, with pIC50s above 5. Interestingly, a set of 10 compounds was also tested against clinical isolates resistant to isoniazid (inh^R) or rifampicin (rif^R) and all the compounds were as active as against the reference strain H37Rv (Fig 2 and Table C).

This new compound set again resides comfortably within the range of properties occupied by marketed drugs and on average has a slightly lower lipophilicity than the first set (see Figs 2 and 3; Table B in S1 File and Figures C-E in S1 File). The compounds identified generally presented a combination of a reasonable level of solubility and anti-mycobacterial activity, indicating their attractiveness as starting points for lead optimisation. No statistically significant difference in the distributions of physicochemical properties was observed between the 7 H37Rv-specific compounds and rest of the compounds, although they are structurally dissimilar.

Target Prediction

The final 50 compounds were computationally analyzed with the goal of identifying their likely target proteins. Our computational approach integrates 2D chemogenomics space (CHEM), structural comparisons (STR) and historical bioassay data (HIST). The results from this analysis were also compared to those from our previous analysis [10].

2D Chemogenomics space (CHEM). The exploration of the chemical space allowed us to identify likely targets (Table 1) for the input compounds based on their structural similarity to compounds with experimentally validated targets deposited in the ChEMBL database [8]. We applied a multi-category naïve Bayesian classifier (MCNBC) that was built and trained using structural and bioactivity information from the ChEMBL database [8]. Given a new

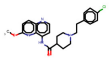
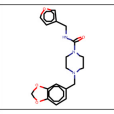
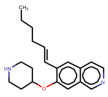
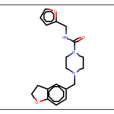
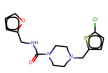
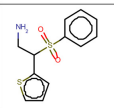
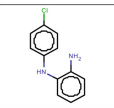
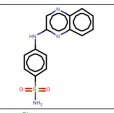
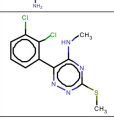
GSKnumber	Structure	H37Rv MIC (μM)	BCG pIC_{50}	ACT In NR % inhibition	ACT In NR pXC50	ACT In Intracellular pX50[b]	TOX_{50} HepG2 (μM)	clogP	Solubility (ug/ml)	Chrom LogD (pH 7.4)	MW
TCMDC-143649		0,2	5,3	**ACTIVE 7 expts out of 10 with inh > 50%**	**5,1**		>100	4,82	4	5,49	424,92
TCMDC-143650		1,25	5,9	INACTIVE			>100	2,80	86	3,48	343,38
TCMDC-143651		1,6	5,6	INACTIVE		5.9	16	4,57	36	3,61	310,43
TCMDC-143652		2	5,6	INACTIVE			>100	2,89	88	3,47	341,4
TCMDC-143653		2*	5,9	INACTIVE		5.6	>100	3,24	98	4,82	339,84
TCMDC-143655		2	5,3	INACTIVE			32	1,10	134	2,5	267,37
TCMDC-143654		2	7	**ACTIVE 8 expts out of 11 with inh > 50%**	**4,5**	7.8	79	3,26	59	5,76	218,68
TCMDC-143656		2	5,2	INACTIVE			>100	2,13	112	1,07	300,34
TCMDC-143661[a]		2,5	4,9	INACTIVE		6.1	>100	4,40	85	4,94	301,2

Fig 2. Complete biological profile of selected hit compounds and corresponding physico chemical properties. [a] Mtb specific. *This compound has been evaluated against a clinical isolate of *M.tuberculosis* resistant to isoniazid and its MIC was in the range of H37Rv (1.6 uM). [b] Compounds being tested in the intracellular assay, data will be available from Dryad Digital Repository: http://dx.doi.org/10.5061/dryad.8r351.

doi:10.1371/journal.pone.0142293.g002

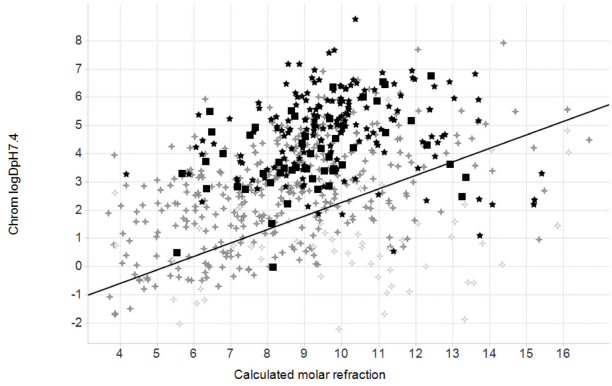


Fig 3. Plot of calculated chromatographic $logD_{7.4}$ versus calculated molar refraction (CMR). All data were generated using the latest version of the GSK calculator. Grey crosses represent marketed drugs with >30% oral bioavailability, white crosses <30% oral bioavailability, and the two disclosed sets by black squares (the current 50 compounds) or black stars (the CMC2013 set of 177). The line represents a discriminator between likely good and bad permeability. The chromatographic logD scale gives values approximately two units higher than the traditional distribution values assessed in octanol/water.

doi:10.1371/journal.pone.0142293.g003

compound, the model calculates a likelihood score based on the molecule's individual sub-structural/fingerprint features and produces a ranked list of likely targets.

In total, the 50 compounds resulted in 262 statistically significant target associations (at a Z-score > 2.0) to 221 different proteins in the ChEMBL database from 24 different organisms (57% of hits are to human proteins). A simple orthology search the OrthoMCL database against the *M. tuberculosis* proteins from this set resulted in 128 compound-target relationships for 61 *M. tuberculosis* proteins, with detectable orthology to 16 organisms (Table C in S1 File).

Historical assay space (HIST). We used the historical GSK bioassay data to develop hypotheses for the anti-mycobacterial mode of action for the active compounds. Using conservative activity thresholds ($pXC50 \geq 5.0$) we found that among the 50 compounds active against *M. tuberculosis* H37Rv, 25 displayed additional activity in 65 different historical biochemical assays against human (50 unique genes), bacterial (1 gene) and viral (1 gene) putative targets (Figure A.A in S1 File). Some compounds were present in multiple historical assays resulting in a total of 93 assay experiments (Figure A.B in S1 File).

The largest human target classes were G protein coupled receptors (GPCRs) and protein kinases, which might partly reflect the relative abundance of different ligand classes in GSK's pharmacological screening collection. We searched for orthologous sequences of the human assayed proteins in the *M. tuberculosis* H37Rv genome using conservative criteria (BLASTP E-value $\leq$1.0e-10) for assigning human-*Mycobacterium* protein homology. Although there are significant evolutionary differences between *Mycobacterium* and human genomes in terms of both gene content and amino acid sequence divergence, we still found 17 *M. tuberculosis* H37Rv gene homologues (Table A in S1 File), which fell into different target class categories (Figure A in S1 File), including kinases (8 genes), cytochromes (6 genes), other enzymes (2 genes) and ion channels (2 genes).

Table 1. Significant links between compound families and targets.

Compound	FamID	Target	Pathway	Essentiality Prediction
TCMDC-143652	1	Rv3569c	Degradation of aromatic compounds (mtu01220) Steroid degradation (mtu00984)	Non
TCMDC-143653	1	Rv3569c	Degradation of aromatic compounds (mtu01220) Steroid degradation (mtu00984)	Non
TCMDC-143657	1	Rv3569c	Degradation of aromatic compounds (mtu01220) Steroid degradation (mtu00984)	Non
TCMDC-143650	1	Rv3569c	Degradation of aromatic compounds (mtu01220) Steroid degradation (mtu00984)	Non
TCMDC-143666	3	Rv2855	Glutathione metabolism (mtu00480)	Yes
TCMDC-143687	3	Rv0427c	Base excision repair (mtu03410)	Non
	3	Rv1629	Base excision repair (mtu03410)	Yes
	3	Rv2855	Glutathione metabolism (mtu00480)	Yes
TCMDC-143688	5	Rv1284	Nitrogen metabolism (mtu00910)	Yes
TCMDC-143670	5	Rv3273	Nitrogen metabolism (mtu00910)	Non
	5	Rv3588c	Nitrogen metabolism (mtu00910)	Non
	5	Rv1284	Nitrogen metabolism (mtu00910)	Yes
	5	Rv3273	Nitrogen metabolism (mtu00910)	Non
	5	Rv3588c	Nitrogen metabolism (mtu00910)	Non
TCMDC-143649	9	Rv0194	ABC transporters (mtu02010)	Non
TCMDC-143690	13	Rv1284	Nitrogen metabolism (mtu00910)	Yes
	13	Rv3588c	Nitrogen metabolism (mtu00910)	Non
TCMDC-143655	29	Rv1151c	Amino sugar and nucleotide sugar metabolism (mtu00520)	Non
TCMDC-143686	36	Rv0233	Purine metabolism (mtu00230)	Non
	36	Rv0733	Purine metabolism (mtu00230)	Non data
	36	Rv1843c	Purine metabolism (mtu00230)	Non
	36	Rv2584c	Purine metabolism (mtu00230)	Non
	36	Rv3275c	Purine metabolism (mtu00230)	Yes
	36	Rv3307	Purine metabolism (mtu00230)	Non
	36	Rv3411c	Purine metabolism (mtu00230)	Yes
TCMDC-143685	38	Rv1905c	D-Arginine and D-ornithine metabolism (mtu00472) Penicillin and cephalosporin biosynthesis (mtu00311)	Non

doi:10.1371/journal.pone.0142293.t001

The specific predictions from the historical assay space search are detailed in S1 File.

3D Structural space (STR). Finally, we applied a Random Forest Score that identified structural similarities between any compound in the dataset and ligands from the PDB [22]. Each compound in the *M. tuberculosis* H37Rv dataset is compared to ~7,600 ligands for which there are known complex structures in the PDB, identifying structural similarities to be included in a pre-built network of structural relationships between ligands and targets. In total, the 50 compounds resulted in 1,890 significant target associations (global Z-score < -1) to proteins in a set of modeled three-dimensional structures from the *M. tuberculosis* proteome (data not shown).

Predicted targets. The similarities and differences of the predictions of the three independent approches are detailed in S1 File.

There were a total of 1,044 unique *M. tuberculosis* targets associated with 112 pathways annotated in the KEGG database [23]. The KEGG being a suite of databases and associated software for understanding and simulating higher-order functional behaviours of the cell or the organism from its genome information. The "mtu" identifiers below refer to the relevant KEGG pathway ids. The three orthogonal approaches identified 66 different pathways (Fig 4A) associated to the 50 hit compounds. The relative increment in the number of putatively affected pathways per compound comparing to the previous TCAMS-TB dataset [10] can be explained by the higher structural diversity of the novel top-up library. Within the commonly identified pathways, there are many associated with amino acid and nucleotide metabolism, e.g. the mtu00260 (Glycine, serine and threonine metabolism), mtu00380 (Tryptophan metabolism), mtu00330 (Arginine and proline metabolism), mtu00270 (Cysteine and methionine metabolism), mtu00240 (Pyrimidine metabolism), mtu00230 (Purine metabolism), mtu00360 (Phenylalanine metabolism), mtu00290 (Valine, leucine and isoleucine biosynthesis). Some of them also appear overrepresented in the predictions, e.g phenylalanine, tyrosine and tryptophan biosynthesis (Fig 3B). Interestingly, there are others overrepresented pathways not directly associated with amino acid metabolism such as mtu05152, mtu01220 (Degradation of aromatic compounds), or mtu00363 (Bisphenol degradation).

To assess the significance of the compound-target predictions using the three different approaches, we calculated a t-statistics p-value of any compound family-KEGG pathway link (Methods). There are 8 compounds families significantly associated (p-value < 0.005) to 10 different KEGG pathways. The threshold used in this study is less restrictive than in the prior study [10] due to the smaller number of compounds. This results in a higher number of associations found between compounds and KEGG pathways. Family 1 is significantly linked with both mtu01220 (Degradation of aromatic compounds) and mtu00984 (Steroid degradation). Specifically, the link found by compounds TCMDC-143652, TCMDC-143653, TCMDC-143657, and TCMDC-143650 targeting Rv3569c (4,9-DHSA Hydrolase) involved both pathways. Family 3 is significantly associated with two different KEGG pathways, mtu00480 (Glutathione metabolism) and mtu003410 (Base excision repair). Specifically, compounds TCMDC-143687 and TCMDC-143666 are predicted to hit Rv2855 (NADPH-dependent mycothiol reductase) involved in Glutathionine metabolism and essential for the survival of the bacteria [24], while TCMDC-143687 is predicted to hit the base excision repair pathway through Rv0427c (Exodeoxyribonuclease III protein XthA) and Rv1629 (DNA polymerase I PolA) being the later essential for the growth of the bacteria [24, 25]. Family 5 has a strong association (p-value 1.0e-08) with mtu00910 (Nitrogen metabolism pathway) an essential pathway for the bacteria survival. Specifically, compounds TCMDC-143688 and TCMDC-143670 are predicted to hit Rv1284 (beta-carbonic anhydrase), Rv3273 (carbonate dehydratase) and Rv3588c (beta-carbonic anhydrase CanB) three proteins involved in the Nitrogen metabolism, where Rv1284 play a key role in the essentiality of this pathway. Moreover, compound TCMDC-143690 belonging to singleton family 13, is also predicted to interact with Rv1284 and Rv3588c targeting the nitrogen metabolism pathway with a completely different chemical scaffold. Another interesting significant link is the compound TCMDC-143648 targeting the mtu02010 (ABC transporters pathway) through the Rv0194 target (transmembrane multidrug efflux pump). Family 29 composed by compound TCMDC-143655 is predicted to interact with Rv1151c (transcriptional regulatory protein), which is involved in transcriptional mechanism and belongs to mtu00520 (amino sugar and nucleotide sugar metabolism). Family 36 with compound TCMDC-143686 is significantly associated with the mtu00230 (purine metabolism) pathway. This compound is predicted to attack the pathway by targeting 7

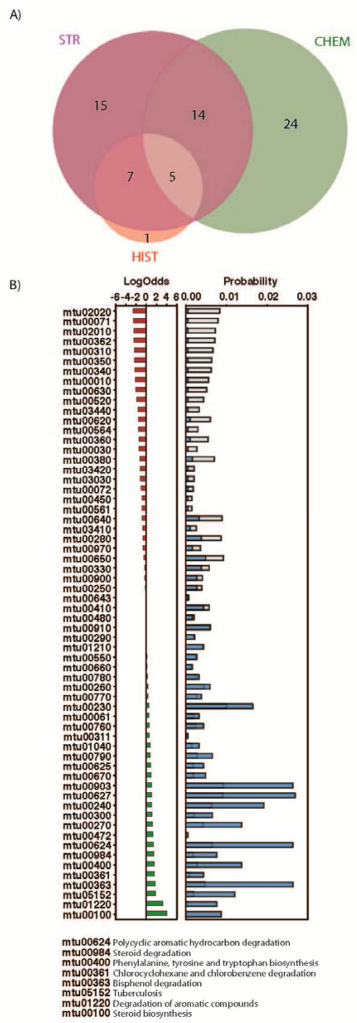


Fig 4. Predicted KEGG pathways targeted by the GSK compounds. A) Venn diagram with common pathways from the three different approaches. B) Most under and over-represented pathways in our predictions. Panels A) and B) with the same representation as in Figure E in S1 File.

doi:10.1371/journal.pone.0142293.g004

different proteins in this pathway (Rv0233, Rv0733, Rv1843c, Rv2584, Rv3275c, Rv3307, and Rv3411c). Among the predicted targets, there are two essential for the bacterial survival [24, 25], the N5-carboxyaminoimidazole ribonucleotide mutase (Rv3275c) and the inosine-5'-monophosphate dehydrogenase (Rv3411c). Finally, a significant link between Family 38 (TCMDC-143685) and pathways mtu00472 (D-arginine and D-ornithine metabolism) and mtu00311 (Penicillin and cephalosporin biosynthesis) was also observed trough the target Rv1905c (a Probable D-amino acid oxidase Aao).

Discussion

Screening for new antitubercular inhibitors in whole cell based assays still sustains a high proportion of the drug discovery pipeline against TB. While this choice of screening strategy is not devoid of its own specific issues [26] the completion of a number of screening efforts and, most importantly, the public release of these datasets, is enabling the in depth validation of novel Mode-of-Actions (MoA) against TB [27–32]. This target elucidation work, in time, is promising to open up new opportunities for TB drug discovery where the limitations associated with the medicinal chemistry optimisation of hits identified by whole cell screening can be addressed through the support provided by technologies typically associated with target based discovery programs, e.g. particular target assays and crystallography. We expect that by accessing these technologies, a more rational understanding of the optimization process and the early identification of potential target related toxicological liabilities could be attained.

It is with this goal in mind that we here present a novel set of antitubercular compounds together with some developability parameters that should provide the TB R&D community with novel chemical starting points for further discovery or, more importantly, future target identification programs. The present release incorporates compounds which, on average, would appear to be in more favourable physical space than those in the previous publication [2, 10]. Given the predominantly intracellular lifestyle of Mtb and the suspected impact of non-replicating bacteria in TB chemotherapy [33, 34], we decided to investigate whether the compounds were capable of inhibiting Mtb growth in THP-1 cells and were active against non-replicating bacteria. 96% of the compounds tested in the intracellular assay were found to be active and 10% of the whole set retained activity in the non-replicating assay. On the basis of the drug-like properties presented in Fig 2 and Table B in S1 File, ten molecules were selected for further characterisation against isoniazid and rifampicin clinical resistant isolates. All compounds were found to be active within the same range as the reference strain H37Rv.

Interestingly, 7 compounds of the set were Mtb specific (inactive against M. bovis BCG). While a number of especulative explanations can be postulated, e.g. differences in permeability, active transport, metabolic state, etc., this lack of correlation highlights the risks associated with the use of non pathogenic surrogate strains in antitubercular research.

To further characterize the activity of the novel antitubercular compounds, we have integrated a series of orthogonal computational approaches for predicting their putative targets. Our analysis found nine chemical families targeting 21 different proteins from 13 biochemical pathways in *M. tuberculosis*. Within the 21 proteins, there are 5 assessed as essential in previous studies. The essentiality of these targets makes them top priority targets for further validation. However, some non-essential targets can have a key role in TB infection in-vivo and therefore we should consider them in the search of new strategies for defeating TB. Our target identification work aims to facilitate further chemical and biochemical experiments to optimize the properties of the compounds against TB. Optimally, additional computational approaches could then interrogate the newly generated compounds to further characterize their mode-of-action. This iterative process is very desirable to maximize the impact of the openly released

new compounds against TB. In particular, we also release the 3D structural models for the significant predictions of targets and compounds identified by the STR approach. Such models and the predicting binding site could be used for computational docking or molecular dynamics analysis to further validate our prediction

Supporting Information

S1 File. Supporting information, Figures and Tables. Figure A. Target class space. A) For positive hits in *M. tuberculosis* H37Rv screens, the distribution of human target classes affected by compounds based on known human protein potency and selectivity criteria as described in the text. The number of human targets is indicated for each class as well as the potential number of *Mtb* homologous genes (in parentheses). B) Distribution of 25 compounds screened against 1 or more targets having pIC50 or pEC50 values > 5.5 in 65 assays by human target classes. Some compounds have historical assay information and potency against multiple target classes. Also indicated is the number of assays against targets with putative homologues in *M. tuberculosis* (in parentheses). **Figure B. Box plot of average PFI** (calculated Chrom Log D7.4 + #Ar) distribution of the 177 compounds released previously [2], the current 50 hits and a representative set of oral drugs. **Figure C**. **Box plot of average calculated Chrom Log D7.4** distribution of the 177 compounds released previously [2], the current 50 hits and a representative set of oral drugs. **Figure D. Box plot of average calculated molar refraction** (CMR) distribution of the 177 compounds released previously [2], the current 50 hits and a representative set of oral drugs. **Figure E. Subset of GSK compounds with predicted targets. A**) Venn diagram with common compounds with predictions from the three different approaches (that is, in green from the search of the chemogenomics space, in purple from the search of the structural space, and in red from the historical data). **B**) Venn diagram with common compound families with predictions from the three different approaches. **C**) Most under and over-represented chemical families in our predictions. Upper plot shows the probability of finding a given family in the original dataset (grey bars) compared to the probability of finding it in the dataset with predicted targets (blue bars). Lower plot shows the log odds per selected family (*i.e*., absolute log odds larger than 0.5). **Table A. Predicted *Mtb*H37Rv gene targets** based on homology to 65 historical human target assays for 25 compounds. Notes: [a] Human target classes are defined in the text. Some compounds were reported active across more than one target class hence the greater number of total than tested compounds. [b] *M. tuberculosis* H37Rv homologs determined by BLASTP searches using human target proteins [19]. [c] Essentiality scoring based on Sassetti et al.[24]. NE = No Evidence from these sources. **Table B.** Complete biological profile of selected hit compounds and corresponding physico chemical properties. **Table C.** Target association based on the structural similarity of the hits to compounds with experimentally validated targets deposited in the ChEMBL database. (PDF)

Acknowledgments

The research behind these results received funding from the TB Alliance, the European Union's 7th framework programme (FP7-2007–2013) under grant agreement ORCHID no. 261378 and the ERA-NET Pathogenomics Project GeMoA (PIM2010EPA-00719). Marc A. Marti-Renom (mmarti@pcb.ub.cat) is the Corresponding Author for computational analysis.

Author Contributions

Conceived and designed the experiments: JRB JPO MAMR AM EAR. Performed the experiments: MJRL DAG FMJ GP VK VB EMLR GM MMH RGR. Analyzed the data: MJRL DAG

FMJ GP MAMR VK MH JCP RHB GC RGR. Wrote the paper: JL MJRL FMJ GP JRB JPO MAMR GC AM RY EAR JCP LB DB RGR. Gave final approval: NC.

References

1. WHO. Global tuberculosis report 2014, 2014.
2. Ballell L, Bates RH, Young RJ, Alvarez-Gomez D, Alvarez-Ruiz E, Barroso V, et al. Fueling open-source drug discovery: 177 small-molecule leads against tuberculosis. ChemMedChem. 2013; 8 (2):313–21. doi: 10.1002/cmdc.201200428 PMID: 23307663
3. Thompson M, Bee HM, Evans WH, Gardner MJ, Greenhow EJ, Howarth R, et al. Robust Statistics—How Not to Reject Outliers Part 1. Basic Concepts Analyst. 1989; 114:1693–7.
4. Coma I, Herranz J, Martin J. Statistics and decision making in high-throughput screening. Methods in molecular biology. 2009; 565:69–106. doi: 10.1007/978-1-60327-258-2_4 PMID: 19551358
5. Hill AP, Young RJ. Getting physical in drug discovery: a contemporary perspective on solubility and hydrophobicity. Drug discovery today. 2010; 15(15–16):648–55. doi: 10.1016/j.drudis.2010.05.016 PMID: 20570751
6. Valko K, Bevan C, Reynolds D. Chromatographic Hydrophobicity Index by Fast-Gradient RP-HPLC: A High-Throughput Alternative to log P/log D. Analytical chemistry. 1997; 69(11):2022–9. doi: 10.1021/ac961242d PMID: 21639241
7. Young RJ, Green DV, Luscombe CN, Hill AP. Getting physical in drug discovery II: the impact of chromatographic hydrophobicity measurements and aromaticity. Drug discovery today. 2011; 16(17–18):822–30. doi: 10.1016/j.drudis.2011.06.001 PMID: 21704184
8. Bento AP, Gaulton A, Hersey A, Bellis LJ, Chambers J, Davies M, et al. The ChEMBL bioactivity database: an update. Nucleic acids research. 2014; 42(Database issue):D1083–90. doi: 10.1093/nar/gkt1031 PMID: 24214965
9. Rogers D, Hahn M. Extended-connectivity fingerprints. Journal of chemical information and modeling. 2010; 50(5):742–54. doi: 10.1021/ci100050t PMID: 20426451
10. Martinez-Jimenez F, Papadatos G, Yang L, Wallace IM, Kumar V, Pieper U, et al. Target prediction for an open access set of compounds active against Mycobacterium tuberculosis. PLOS computational biology. 2013; 9(10):e1003253. doi: 10.1371/journal.pcbi.1003253 PMID: 24098102
11. Available from: http://orthomcl.org/orthomcl/.
12. Chen F, Mackey AJ, Stoeckert CJ Jr., DS Roos. OrthoMCL-DB: querying a comprehensive multi-species collection of ortholog groups. Nucleic acids research. 2006; 34(Database issue):D363–8. PMID: 16381887
13. Marti-Renom MA, Rossi A, Al-Shahrour F, Davis FP, Pieper U, Dopazo J, et al. The AnnoLite and AnnoLyze programs for comparative annotation of protein structures. BMC bioinformatics. 2007; 8 Suppl 4: S4. PMID: 17570147
14. Kalinina OV, Wichmann O, Apic G, Russell RB. Combinations of protein-chemical complex structures reveal new targets for established drugs. PLOS computational biology. 2011; 7(5):e1002043. doi: 10.1371/journal.pcbi.1002043 PMID: 21573205
15. Kalinina OV, Wichmann O, Apic G, Russell RB. ProtChemSI: a network of protein-chemical structural interactions. Nucleic acids research. 2012; 40(Database issue):D549–53. doi: 10.1093/nar/gkr1049 PMID: 22110041
16. Berman H, Henrick K, Nakamura H, Markley JL. The worldwide Protein Data Bank (wwPDB): ensuring a single, uniform archive of PDB data. Nucleic acids research. 2007; 35(Database issue):D301–3. PMID: 17142228
17. Konc J, Janezic D. ProBiS: a web server for detection of structurally similar protein binding sites. Nucleic acids research. 2010; 38(Web Server issue):W436–40. doi: 10.1093/nar/gkq479 PMID: 20504855
18. Martinez-Jimenez F, Marti-Renom MA. Ligand-target prediction by structural network biology using nAnnoLyze. PLOS computational biology. 2015; 11(3):e1004157. doi: 10.1371/journal.pcbi.1004157 PMID: 25816344
19. Altschul SF, Madden TL, Schaffer AA, Zhang J, Zhang Z, Miller W, et al. Gapped BLAST and PSI-BLAST: a new generation of protein database search programs. Nucleic acids research. 1997; 25 (17):3389–402. PMID: 9254694

20. Fisher RA. On the Interpretation of $\chi 2$ from Contingency Tables, and the Calculation of P. Journal of the Royal Statistical Society. 1922; 85(1):87–94.
21. Warrier T, Martinez-Hoyos M, Marin-Amieva M, Colmenarejo G, Porras-De Francisco E, Alvarez-Pedraglio AI, et al. Identification of novel anti-mycobacterial compounds by screening a pharmaceutical small-molecule library against non-replicating Mycobacterium tuberculosis. In preparation2015.
22. Berman HM, Westbrook J, Feng Z, Gilliland G, Bhat TN, Weissig H, et al. The Protein Data Bank. Nucleic acids research. 2000; 28(1):235–42. PMID: 10592235
23. Kanehisa M. The KEGG database. Novartis Found Symp. 2002; 247:91–101; discussion -3, 19–28, 244–52. PMID: 12539951
24. Sassetti CM, Boyd DH, Rubin EJ. Genes required for mycobacterial growth defined by high density mutagenesis. Molecular microbiology. 2003; 48(1):77–84. PMID: 12657046
25. Griffin JE, Gawronski JD, Dejesus MA, Ioerger TR, Akerley BJ, Sassetti CM. High-resolution phenotypic profiling defines genes essential for mycobacterial growth and cholesterol catabolism. PLOS pathogens. 2011; 7(9):e1002251. doi: 10.1371/journal.ppat.1002251 PMID: 21980284
26. Goldman RC. Why are membrane targets discovered by phenotypic screens and genome sequencing in Mycobacterium tuberculosis? Tuberculosis. 2013; 93(6):569–88. doi: 10.1016/j.tube.2013.09.003 PMID: 24119636
27. Abrahams KA, Cox JA, Spivey VL, Loman NJ, Pallen MJ, Constantinidou C, et al. Identification of novel imidazo[1,2-a]pyridine inhibitors targeting M. tuberculosis QcrB. PLOS one. 2012; 7(12):e52951. doi: 10.1371/journal.pone.0052951 PMID: 23300833
28. Pethe K, Bifani P, Jang J, Kang S, Park S, Ahn S, et al. Discovery of Q203, a potent clinical candidate for the treatment of tuberculosis. Nature medicine. 2013; 19(9):1157–60. doi: 10.1038/nm.3262 PMID: 23913123
29. Trefzer C, Skovierova H, Buroni S, Bobovska A, Nenci S, Molteni E, et al. Benzothiazinones are suicide inhibitors of mycobacterial decaprenylphosphoryl-beta-D-ribofuranose 2'-oxidase DprE1. Journal of the American Chemical Society. 2012; 134(2):912–5. doi: 10.1021/ja211042r PMID: 22188377
30. Ioerger TR, O'Malley T, Liao R, Guinn KM, Hickey MJ, Mohaideen N, et al. Identification of new drug targets and resistance mechanisms in Mycobacterium tuberculosis. PLOS one. 2013; 8(9):e75245. doi: 10.1371/journal.pone.0075245 PMID: 24086479
31. Remuinan MJ, Perez-Herran E, Rullas J, Alemparte C, Martinez-Hoyos M, Dow DJ, et al. Tetrahydropyrazolo[1,5-a]pyrimidine-3-carboxamide and N-benzyl-6',7'-dihydrospiro[piperidine-4,4'-thieno[3,2-c]pyran] analogues with bactericidal efficacy against Mycobacterium tuberculosis targeting MmpL3. PLOS one. 2013; 8(4):e60933. doi: 10.1371/journal.pone.0060933 PMID: 23613759
32. Andries K, Villellas C, Coeck N, Thys K, Gevers T, Vranckx L, et al. Acquired Resistance of Mycobacterium tuberculosis to Bedaquiline. PLOS one. 2014; 9(7):e102135. doi: 10.1371/journal.pone.0102135 PMID: 25010492
33. Lee J, Hartman M, Kornfeld H. Macrophage apoptosis in tuberculosis. Yonsei medical journal. 2009; 50 (1):1–11. doi: 10.3349/ymj.2009.50.1.1 PMID: 19259342
34. Nathan C, Barry CE 3rd. TB drug development: immunology at the table. Immunological reviews. 2015; 264(1):308–18. doi: 10.1111/imr.12275 PMID: 25703568

3.3. Rational design of non-resistant targeted cancer therapies

This chapter presents a computational model that predicts cancer associated mutations with the highest chance to confer resistance to a targeted therapy. Furthermore, for those mutations predicted as highly resistance-like it suggests alternative non-resistant compounds. The model exemplified it applicability in two targeted therapies: EGFR-gefitinib for the treatment of Lung adenocarcinoma and Lung Squamous Cell Carcinoma; and the ERK2-VTX11e therapy for the treatment of melanoma and colorectal cancer.

Manuscripts presented in this section:

Martínez-Jiménez, F., Overington J. P., Al-Lazikani B., & Marti-Renom, M. a. (2016). **Rational design of non-resistant targeted cancer therapies**. Nucleic Acids Research. (*Submitted*).

Rational design of non-resistant targeted cancer therapies.

Francisco Martínez-Jiménez[1,2,3,$], John P. Overington[4], Bissan Al-Lazikani[5] and Marc A. Marti-Renom[1,2,3,6,*]

1. CNAG-CRG, Centre for Genomic Regulation (CRG), Barcelona Institute of Science and Technology (BIST), Baldiri i Reixac 4, 08028 Barcelona, Spain.
2. Gene Regulation, Stem Cells and Cancer Program, Centre for Genomic Regulation (CRG), Dr. Aiguader 88, 08003 Barcelona, Spain.
3. Universitat Pompeu Fabra (UPF), Barcelona, Spain.
4. Stratified Medical, 40 Churchway, London NW1 1LW, UK.
5. Institute of Cancer Research, Cotswold Road, Sutton, Surrey, UK.
6. ICREA, Pg. Lluís Companys 23, 08010 Barcelona, Spain.

Keywords: targeted cancer therapies, drug resistance, phenotype prediction, signatures of mutational processes.

*Correspondence should be addressed to martirenom@cnag.crg.eu

ABSTRACT

Drug resistance is one of the major problems in targeted cancer therapy. High mutation rates and selective pressure can efficiently result in drug-resistance to therapy. Although there are many mechanisms for drug resistance, a classic mechanism is due to changes in the amino acids in the drug-target binding site. Despite of the numerous efforts made to individually understand and overcome these mutations, there is a lack of comprehensive analysis of the mutational landscape that can potentially cause resistance.

Herein, we present a framework that computationally predicts the potential of a sequence mutation to confer resistance to targeted therapies in cancer. Our model first quantifies the likelihood of mutations in the drug target using the probabilities from the mutational signatures associated to the cancer class. Next, it uses structural information of the drug-protein interaction to predict the *resistance-likeness* of the aforementioned mutations. The combination of the predicted likelihood with the *resistance-likeness* allows the detection of mutations with the highest chance to be responsible of resistance to a particular targeted cancer therapy. Finally, for these treatment-threatening mutations, the classifier proposes alternative therapies overcoming the resistance.

We exemplified the applicability of the model using the EGFR-gefitinib treatment for Lung Adenocarcinoma (LUAD) and Lung Squamous Cell Cancer (LSCC) and the ERK2-VTX11e treatment for melanoma and colorectal cancer. Our model correctly identified the phenotype of some of the known resistance mutations, including the EGFR-T790M or the ERK2-P58L/S/T. Moreover, the model predicted new clinically unseen mutations as potentially responsible of resistance to EGFR-gefitinib and ERK2-VTX11e targeted cancer therapies. Finally, we provided a map of the predicted sensitivity of alternative ERK2 and EGFR inhibitors, with a particular focus in two molecules with a low predicted *resistance-likeness*.

In summary, we introduced a new computational framework aiming at connecting the mutational landscape of tumors with the drug-resistance phenotype generated by spontaneous mutations in drug targets.

INTRODUCTION

Non-selective cytotoxic agents have traditionally dominated cancer treatment. However, the strong side effects and the limited effectiveness associated with drug resistance have led to the search of alternative treatments [1]. In the last decade, rationally designed targeted therapies have been proposed as less damaging and more accurate alternative to treat cancer [2]. In fact, targeted therapies have produced substantial clinical responses in the treatment of chronic myeloid leukemia (CML) [3], non-small cell lung cancer (NSCLC) [4] or melanoma [5]. Unfortunately, after a short period of time, most tumors develop resistance to these treatments causing a cancer relapse with fatal consequences [6, 7].

There are several mechanisms conferring drug resistance to targeted therapies [8]. Mechanism such as activation of survival signaling pathways or the inactivation of downstream death signaling pathways [9, 10], increasing drug efflux or alterations in drug metabolism [11, 12]. Epigenetic changes and their influence of in the tumor microenvironment have also been proposed to play a key role in chemoresistance [12, 13]. Moreover, secondary mutations of drug targets are frequently reported as a mechanism of drug resistance. In NSCLCs, patients initially responding to first generation EGFR inhibitors such as gefitinib and erlotinib, acquire resistance within 1 year. In 50% of such cases, a secondary T790M gatekeeper mutation has been identified [14, 15]. Recently, a third generation of T790M-EGFR selective inhibitors, such as rociletinib [16] or osirmetinib [17] have been designed to overcome resistance in EGFR-T790M positive patients [18].

Unfortunately, EGFR-T790M is a singular example, we still are far from completely overcome the clinical challenge of resistance due to mutations in oncogenic kinases. Many studies have been carried to both systematically analyze resistance to kinase inhibitors [19] and to propose alternatives to standard kinase inhibitor treatments [20]. Nevertheless, these studies do not cover the whole spectrum of mutations of the target, being usually limited to a small, and clinically reported, number of kinase mutations. Moreover, the nature of tumors is complex and very heterogeneous [21]. Estimates of the number of coding mutations in the entire cell population of a tumor are of the order of thousands or even millions of mutations depending of the tumor type and size [22]. Standard NGS sequencing of solid biopsies only enables the detection of mutations present in > 5% of tumor cells [23]. The low sensitivity of standard NGS technologies alongside the heterogeneous nature of solid tumors, may

lead to a significant loss of low-frequency mutations present in small cell populations. Remarkably, low-frequency mutations can confer resistance to targeted therapies and therefore, become clonal drivers once the cancer treatment begins [7, 24, 25].

The invasive nature and the technical limitations associated with sequencing methods of solid biopsies highlight the importance of computational models in cancer evolution and drug resistance. The advent of the massive cancer genomic data has prompted the development of several mathematical and computational models [26]. Some of these models focus on characterizing tumor evolutionary processes [27-29] while others, study tumor response to single targeted treatment [30-33] or combinational therapy [34]. However, none of these models, which are usually applied to known drug-resistant mutations, specifically predict which are the causative mutations leading to drug resistance.

Herein, we present a computational framework for *de-novo* predicting mutations with potential to confer resistance to small molecule targeted therapies. Additionally, the model provides a list of alternative compounds ranked by their predicted sensitivity to these *resistance-like* mutations. The framework connects the mutational landscape of tumors with the drug-resistance phenotype generated by spontaneous mutations in drug targets. We exemplified the applicability of the framework in two protein kinases, EGFR and ERK2 (also known as MAPK1). EGFR is well-studied model in resistance to targeted cancer therapy, and consequently, is a good system to validate the full scope of the framework. We computationally predicted the likelihood and the *resistance-likeness* of the EGFR residues involved in the binding of gefitinib in LUAD and LSCC. Additionally, using the mutational signatures previously defined [35], we also analyzed the possible aetiology (or aetiologies) associated to each of the most critical EGFR mutations. Our model correctly predicted the phenotype of the EGFR-T790M mutation, with the added value of the identification of new unseen mutations that might confer resistance to gefitinib treatment. ERK2, on the other hand, is as a promising target in the treatment of melanoma [36, 37] and colorectal cancer [38]. We predicted the VTX11e-resistance potential of 424 ERK2 mutations. These predictions include the correct identification of 8 mutations alongside new unseen ERK2 mutations predicted to confer resistance to VTX11 treatment in melanoma and colorectal cancer. Moreover, the structural nature of the predictions helped to elucidate the specific mechanism of resistance of each mutation. Finally, for both

EGFR and ERK2 treatment-threatening mutations, the model proposed alternative inhibitors that might overcome resistance.

METHODS

The likelihood model

We developed a model to estimate cancer-associated likelihoods of spontaneous mutation in drug targets (Fig 1). First, using published mutational signatures [35, 39], we annotated the contribution of each of the 30 signatures to the 36 different classes of cancer present in the study. Second, for each signature, we extracted the probabilities of the 96 possible pyrimidine-based mutations (C>A, C>T, C>G, T>A, T>C, T>G) in their 5' and 3' contiguous bases context from the COSMIC database (from http://cancer.sanger.ac.uk/cosmic/signatures). Next, for each signature without described strand-bias we extended the probabilities to the purine-based mutations (G>A, G>C, G>T, A>C, A>T, A>G). Signatures with strong mutational strand-bias towards a specific type of base pair were manually updated depending of their specific type of bias. For instance, signature 7 has a strong transcriptional strand-bias indicating that mutations occurs a pyrimidines base pairs, therefore the mutational probabilities of purines in signature 7 are set to 0. Signatures with strong mutational strand-bias are signatures 4, 7, 11, 22, 24 and 29. This approach resulted in a total of 192 mutational probabilities for each signature.

We compute the likelihood ($L_{m,c}$) of a specific mutation (m) in a particular type of cancer (c) as the sum the probabilities of that mutation in all the signatures involved in that cancer type, weighted by the specific contribution of that signature to the cancer class. Since, several nucleotides mutations can lead to the same amino acid change, all these probabilities are eventually added to measure the amino acid mutation likelihood using the equation:

$$L_{m,c} = \sum_{m=0}^{M} \sum_{c=0}^{S} W_c * P_{m,c}$$

where M is all possible nucleotide changes associated to an amino acid mutation m, S are the signatures associated to the studied cancer class c, W_c is the contribution of signature c to the studied cancer and $P_{m,c}$ is the probability of a given mutation m in the signature c.

EGFR and ERK2 mutants and structural model generation

We applied the likelihood model to predict the probability of mutation of all the amino acids involved in the EGFR binding site to gefitinib (PDB code: 4WKQ), and VTX11e binding to ERK2 (PDB code: 4QTE). We defined a drug binding-site in a protein structure as all the amino acids with at least one atom within 9.5 Å of distance to the co-crystalized drug.

Next, models of all the possible mutations of the drugs binding-sites were generated using the *mutate_model* function of the MODELLER software with default parameters [40, 41]. Models for truncating mutations (*i.e.,* introducing a stop codon) were not generated. The final number of three-dimensional (3D) models was 367 and 424 for EGFR and ERK2, respectively.

Enrichment analysis of the predicted nucleotide mutations likelihood

To measure whether a nucleotide mutation A>B is enriched among the most likely target mutations in a particular cancer class, we calculated the odds ratio of the specific nucleotide mutation A>B for the top 50 likely mutants. More specifically, the odds ratio of a particular nucleotide mutation *A>B* at the i^{th} position in the distribution is given by:

$$odds = \frac{(A > B)_i / (A > B)_{i+}}{\neg(A > B)_i / \neg(A > B)_{i+}}$$

where $(A>B)_i$ denotes the number of *A>B* mutations between the *0* and i^{th} position. $(A>B)_{i+}$ represents the number of A>B mutations between *i+1* and the N^{th} position, being N the total number of amino acid mutations.

Drug-response predictor

We developed two Random Forest Classifiers (RFC). The first classifier, called aa-RFC (amino acid based RFC) predicts the phenotypic effect of an amino acid mutation to the binding affinity between a drug and the target protein. The second classifier, called lig-RFC (ligand based RFC), aims to predict the sensitivity of a group of compounds to a particular mutation in their protein target. Both classifiers use structural and sequential information of the drug-protein interaction to perform the predictions (see below for detailed information about the specific features used

for each classifier). The lig-RFC emphasizes in the ligand-target interaction while omitting some information relative to the amino acid characteristics, which makes it computationally faster to build. Both classifiers were built using the WEKA package [42] with the following parameters: *numTrees* = 1,000; *numFeatures* = 20; *maxDepth* = FALSE. Evaluation of the classifiers performance was done by 10-fold cross validation (CV). Additionally, the relative importance of each variable in the classifiers was calculated by the *randomForest* package of R [43]. Next, we describe all necessary steps to generate and test the classifiers.

Dataset generation

The aa-RFC and lig-RFC were trained using the Platinum database [44]. Briefly, this database contains information about experimentally measured changes in drug binding affinity upon mutations. Moreover, most the entries in the database contain crystal structures of the drug-protein complexes. When no crystal structure was available for either the wild type or the mutated structure, a 3D model was generated using MODELLER with default parameters. The database originally included 1,008 instances. Since the aa-RFC classifier has been developed to individually assess the resistance potential of a single mutation, we removed 208 instances containing double (155), triple (30) or more mutants. The final dataset contained 770 instances of 3D structures (409 models) and their binding compounds. Next, the database was split into four different classes corresponding to four different phenotypes: (i) “strong resistance” (SRES, 293 instances) with a fold change in binding affinity smaller or equal to -5.0, which disrupt the binding of the compound with the target protein; (ii) “resistance” (RES, 227 instances) with a fold change between -5.0 and -1.2, which decreased binding affinity between the compound and the target protein; (iii) “neutral” (NEU, 70 instances) with a fold change between -1.2 and 1.2, which indicates not significant alteration of the binding affinity of the compound; and (iv) “increased sensitivity” (ISEN, 180 instances) with a fold change higher than 1.2 for mutations increasing the affinity of the compound. Finally, the unbalanced nature of the dataset could have introduced bias in the classifier predictions towards SRES and RES classes because of the higher number of instances. Therefore, we randomly removed instances of the SRES and RES classes to 180. The final dataset was therefore composed by 180 instances of the SRES, RES and ISEN classes and 70 of the NEU class.

Sequence and structure features calculated from the 3D models/structures

For each instance in the dataset we calculated a set of features to describe the structural and sequential changes introduced by the mutation. The complete list of features alongside their description and information about their inclusion in the two classifiers are next detailed:

1. **Molecular surface area of the drug binding-site (aa-RFC, lig-RFC).** Total molecular surface area of wild type (WT) and mutated (MT) drug binding-site. Additionally, the absolute numerical difference between the two values was included. The *get_area* function of PyMol 1.8 Version [45] was used for their calculation.
2. **Solvent accessibility of the WT and MT amino acid (aa-RFC, lig-RFC).** Additionally, the absolute numerical difference between the two values was included. The *get_area* function of PyMol 1.8 Version [45] was used for their calculation.
3. **Relative solvent accessibility (RSA) of the WT/MT residue (aa-RFC, lig-RFC).** Ratio between the solvent accessibility area and the general residue surface area calculated using DSSP with default parameters [46]. Additionally, the absolute numerical difference between the two values was included
4. **Half sphere exposure of the WT/MT amino acid (aa-RFC).**[47]. The *HSExposure* class from the *Biopython* library [48] was used for its calculation. Additionally, the absolute numerical difference between the two values was included.
5. **Type of amino-acid change (aa-RFC, lig-RFC)**. A vector of 20 positions representing the 20 amino acids. In the vector, a -1 represents the wild type amino acid, a 1 represents the new residue introduced by the mutation, and 0 represents no change.
6. **Hydrogen bonding (aa-RFC, lig-RFC).** We calculated whether there is a hydrogen bond between the WT/MT residue and the drug bound molecule. Information about the hydrogen bond type and distance were also included. The upper bound to assess an hydrogen bond was 3.2 Å.
7. **Structural environment of the amino acid (aa-RFC, lig-RFC).** We represented the structural environment with concentric spheres surrounding

the mutated amino acid. Each of the spheres has different radius ranging from 0 Å to 6 Å in steps of 1 Å. The spheres were represented using 6 vectors of 20 positions indicating the presence or absence of an amino acid. A number one in a vector implied that the amino acid representing that position was within that radius.

8. **Sequence environment (aa-RFC, lig-RFC).** We defined the amino acid sequence environment as the composition of all 10 contiguous amino acids in sequence (5 amino acids preceding and 5 amino acids following the mutated amino acid). Each position was represented by a vector of 20 amino acids where 1 indicated presence and 0 absence of the given amino acid in the sequence environment.
9. **Secondary structure of the amino acid (aa-RFC, lig-RFC).** We calculated the secondary structure of the WT/MT amino acid using DSSP with default parameters [46].
10. **Protein stability change (aa-RFC, lig-RFC).** We calculated the change in the stability of the protein caused by the mutation using I-Mutant 2.0 [49]. We included two variables, the first one describes the numerical change in stability measured in kcal/mol and the second was a categorical variable representing the sign of stability change: UNSTABLE for negative values, STABLE for positive values and UNKNWON for mutations where I-Mutant 2.0 could not compute a score (that is, in 19% of cases).
11. **Residue conservation (aa-RFC, lig-RFC).** To calculate the conservation score we first performed a BLAST search [50] using as query the target sequence. The resulting multiple alignment was used as input to the *SubsMat* function from Biopython library [48] to obtain a residue conservation score based on the BLOSUM62 matrix [51].
12. **Structural alignment of the MT model to the WT structure (aa-RFC, lig-RFC).** Root Mean Squared Deviation (RMSD) of the structural alignment between the wild type and the mutated protein structures. Two different RMSD were calculated, the first resulted from the original structural alignment and the second from the refined one. The *Super* function from PyMol 1.8 Version [45] was used to perform both structural alignments.
13. **Distance to the ligand (aa-RFC, lig-RFC).** We measured the distances between the alpha carbon of the WT/MT amino acid to all the atoms of the ligand. Next, we calculate the minimum, maximum and average distances to

the ligand. For all of these distances the absolute numerical difference between the WT and MT value was included. PyMol 1.8 Version [45] was used for their calculation.

14. **Charge of the WT and MT amino acids (aa-RFC, lig-RFC).** A vector of 20 positions was generated with -1 for negatively charged amino acids (ASP, GLU), a +1 for positively charged amino acids (LYS, ARG) and 0 for the remainders.
15. **Change in the hydrophobicity (aa-RFC, lig-RFC).** We calculate the difference between WT and the MT amino acids using a pre-calculated hydrophobicity scale [52].
16. **Drug affinity of the ligand with the WT protein (aa-RFC, lig-RFC)**. We retrieved the binding affinity using BindingDB [53]. Depending of the availability on the BindingDB record, the binding affinity was measured by the inhibitory constant (Ki), the dissociation constant (Kd) or the half maximal inhibitory concentration (IC_{50}) measures.
17. **Salt bridge between WT/MT amino acid with other residues (aa-RFC)**. Number of salt bridges between the GLU and ASP amino acids of the WT/MT protein surface were calculated. Additionally, the absolute numerical difference between the two values was included. An upper bound cu-off of 4.0 Å distance between the anionic group of GLU/ASP and the cationic group of LYS/ARG was used.
18. **Salt bridge between WT/MT amino acid with the ligand (aa-RFC, lig-RFC)**. We used the PLIP [54] software with default parameters (v1.2.0) to calculate salt bridges between ASP or GLU residues of the protein and the query drug. Information about the distance measured in Å, type of acceptor and donor groups (Phosphate, Carboxylate, Guanidine, Tertamine or Quartamine) was also included in the lig-RFC.
19. **Disulphide bonds (aa-RFC).** If the mutated residue is a cysteine, we identified putative intra-cysteine disulphide bonds. The expected SG–SG distance for disulfide bond is around 2 Å but more generous definition accounts for inaccuracies in experimental data. Therefore we used disulphide bond distances between 1.8 Å and 2.2 Å.
20. **Halogen bonds (lig-RFC).** The presence of halogen bonds between the WT/MT amino acid and the ligand. It also included information about the type of donor and acceptor atoms. This feature was calculated using PLIP [54]

with default parameters (v1.2.0). Features 21 to 24 were also obtained using PILP.

21. **π-stacking interactions (lig-RFC).** The presence of π-stacking interactions between the ligand and the WT/MT residue including information about the distance and group of interactions.
22. **π-cation interactions (lig-RFC).** The presence of π-cation interactions between the ligand and the WT/MT residue including information about the distance and atoms group involved in the interactions.
23. **Water bridges (lig-RFC).** The presence of water bridges between the WT/MT amino acid and crystallized waters molecules including the type of donor and acceptor atoms.
24. **Hydrophobic interactions (lig-RFC).** The presence of hydrophobic interactions between the ligand and the WT/MT amino acid including information about the distance of the interaction.

In summary, a total of 58 features were used for the aa-RFC and a total of 89 were used for the lig-RFC. The complete list of features and values for the training set are available as supplementary file.

Predictions and resistance score

We applied the aa-RFC to individually predict the phenotype of each of the EGFR and ERK2 mutations defined by the likelihood model. For each compound-protein-mutant, the aa-RFC assigns a confidence score for each the four possible phenotypes (SRES, RES, NEU and ISEN classes). The class-confidence scores addition is equal to 1. The highest class-confidence score corresponds to the predicted class. Next, we defined a global Resistance Score (RS) as the sum of the SRES and RES scores minus the ISEN and NEU weighted by the precision of each class in the aa-RFC training. The normalized RS measure aims at assessing the resistance-likeness of a mutation in a target for the studied drug. The RS score is defined as:

$$RS = \sum_{R=RES,SRES} S_R * P_R - \sum_{S=NEU,ISEN} S_S * P_S$$

where *R* are the two classes of resistance (*i.e.,* SRES and RES) and *S* are the two classes of non-resistance (*i.e.,* NEU and ISEN), S_x is the aa-RFC confidence score

for the class *x* and P_x is the global aa-RFC accuracy for the class *x* after a 10-fold cross validation. Finally, a normalized RS (<NRS>) was calculated by scaling all RS values within an experiment between 1.0 (that is, the highest RS) and 0.0 (that is, the lowest RS).

Creating a dataset of insensitive molecules

To identify compounds that may result insensitive to a particular mutation, and thus be alternative to a given treatment, we first manually extracted all compounds annotated in the Food and Drug administration (FDA, http://www.fda.gov) and the National Cancer institute (NCA; http://www.cancer.gov) web sites that are interacting with the studied protein target**.** Second, we collected all co-crystallized molecules with the protein target. Next, molecules with no experimentally measured binding affinity in BindingDB [53] were discarded. Due to the limited number of small molecules co-crystallized with ERK2, we extended the search to small-molecule ERK2 inhibitors with IC_{50} smaller or equal to 100 nM from the ChEMBL database [55]. Finally, we manually included compounds of interest into the final dataset, which resulted in a total of 19 and 75 possible non-resistant molecules to EGFR and ERK2 respectively.

Predicting molecules likely to be insensitive to a binding site mutation in the protein target

Once the dataset was built, we used it to identify molecules whose affinity may not decrease by a mutation in the protein target. Depending of the source of the molecule, the methodology to assess the sensitivity was different: (i) for the two first subsets (*i.e.,* those co-crystallized with the target) we defined the potential of the mutation to confer resistance using the crystal structure of the drug bound to the target; (ii) for those molecules extracted from ChEMBL and those manually included, we selected the top ranked pose by Autodock Vina [56] by performing virtual docking between the compounds and the target binding pocket. In both cases, each compound-target-mutation prediction was further scored by the normalized RS.

Predicting changes in affinity using AutoDock Vina

Finally, to assess the base-line accuracy when no additional information is given, a new classifier was trained using only the calculated binding affinity change by AutoDock Vina. For each wild type and mutated complex in the aa-RFC training set, we first calculated the predicted affinity of the top ranked pose by AutoDock Vina. Next, the two affinities were passed to a RFC classifier that eventually predicted the phenotypic class of the instance. The classifier parameters and the subsequent validation were performed using the same parameters than in the aa-RFC training. For each instance in the training set the fold change in the predicted affinity was calculated as the ratio between the wild type and the mutated predicted affinities.

RESULTS

Prediction of the drug binding affinity change upon single mutation

We tested the performance of the aa-RFC classifier using the Platinum database [44]. The average AUC of the classifier (0.77) together with a Kappa statistic of 0.40 [57] indicated an overall high accuracy of the classifier, especially for a four-class classifier (Fig 2a). The SRES class was the best predicted with a 0.81 AUC (0.63 precision and 0.62 recall). The second best predicted class was the ISEN class with a 0.79 AUC (0.59 precision and 0.55 recall). Despite the fact that these two classes performed similarly, the lower recall of the IS class indicated that this class had a higher number of false negatives (FN; *i.e.,* instances of the ISEN class miss-assigned to other class). This suggested that the classifier might have some difficulties in correctly finding the ISEN true positives (TP; *i.e.,* instances of the ISEN class correctly predicted). More specifically, of the total 180 ISEN instances, 43 were miss-classified as RES, 28 as SRES and 8 as NEU. Overall, the aa-RFC classified tended to over-assign instances to the RES class, which reflected to its performance metrics (0.74 AUC, 0.50 precision and 0.62 recall). Despite of this, it is remarkable that the aa-RFC resulted in a 0.50 precision for the RES class, which is twice the random value in a four classes classifier. Finally, the NEU class was the worst performing class (0.70 AUC, 0.48 precision, and 0.24 of recall). The low recall value (only one out of four NEU instances were assigned to the class) could be explained by the under-representation of the NEU instances in the training set (only 80 instances, versus 180 instances of the other classes).

To our knowledge this is the first classifier that predicts the resistance-associated phenotype of a mutation for a compound binding to a protein. However, there are multiple methods that predict the binding affinity of a drug-protein complex. These methods can be also applied to predict how a mutation can change the binding affinity of a particular binding compound. One of the most extensively used virtual screening methods is AutoDock Vina (ADV) [56]. Overall, the performance of the ADT classifier was worse, with an average AUC of 0.64 (0.77 of the A-RFC) and a Kappa statistic of 0.19 (Fig 2a). More specifically, the four phenotypic classes had considerably lower AUC values for the ADV predictions. The SRES class resulted in the greatest AUC drop compared to aa-RFC (0.81 to 0.65), followed by the NEU class (0.69 to 0.57), the ISEN class (0.79 to 0.68) and by the RES class (0.71 to 0.63). The individual values in change of affinity for each of the training cases

showed that only 13 (1.7%) instances had fold changes greater than 1.2, which suggests that virtual docking methods may have difficulties detecting large changes in affinity upon single mutation.

To assess the contribution of each of the 58 input variables to the aa-RFC classified, we sorted them by their mean decrease Gini [58], which describes how much each variable contributes to the homogeneity of the nodes and leaves in the resulting random forest. The most informative features were those associated with the change in the molecular surface area and solvent accessibility of the mutated amino acid (ranking positions 1st, 3rd, 4th, 8th-10th, Fig 2b). Change in the protein stability measured by I-Mutant 2.0 [49] was ranked in the second position. Multiple measures of the distance from the amino acid to the ligand were ranked from the 5th to the 7^{th,} positions, while other features such as the affinity of the wild type complex (20th) or the type of secondary structure of the amino acid (21st and 22nd) occupied the following positions. Features based in biochemical properties of the mutated amino acid were clearly overrepresented within the top 25 set (18 out of 25). Only the distance to the ligand and the wild type experimentally measured affinity were included within the top 25 features. Overall, these results showed that the classifier weighted more features based on biochemical properties of the amino acid while gave less relevance to those extracted from specific interaction with the ligand.

EGFR predicted mutational landscape in LUAD and LSCC cancer types

We studied the mutational probability landscape of EGFR in two different non-NSCLC cancer types: LUAD and LSCC (Fig 3a and 3b). The analysis of the mutational landscape indicated that each cancer type had their own underlying mutational mechanisms. Only 20 mutations (that is, ~5% of all binding-site mutations) were ranked in the same position in both cancer types and none of them had the same predicted likelihood. The main discrepancy may be associated to the contribution of signatures 4 and 5 (Supplementary Material). On the one hand, signature 4 is mainly characterized by C>A transversions caused by tobacco smoking [59]. LUAD has a slightly higher contribution from signature 4, resulting in 1.6 times greater average likelihood of C>A mutations in LUAD (0.0226 ± 0.0091 average estimated probability of mutation) than LSCC (0.0143 ± 0.0053). On the other hand, signature 5 has an unknown aetiology and it is associated with T>C substitutions at ApTpN context. Since signature 5 contribution to LSCC is higher than

to LUAD, it resulted in a 2.7 higher average likelihood of the ApTpN mutations in LSCC (0.0057 ± 0.0025) compared to LUAD (0.0037 ± 0.0021).

Analysis of the type of nucleotide change of the top likely mutations revealed an enrichment of C>A mutations in both cancer classes. The highest odds ratio of C>A mutations corresponded to position 21st, with an odds ratio value of 16.2 and position 29th with an odds ratio value of 10.0 in LUAD and LSCC, respectively. Additionally, seven (P741H, P794H, S720Y, P741T, L798I, L799M and L777M) and four (S720Y, P741T, P741H and P794H) mutations within the top 10 were C>A mutations in LUAD and LSCC, respectively (inner sets in Fig 3a and 3b). An exception to this trend was the top likely mutation, that is E762K, caused by T[G>A]A (T[C>T]A in pyrimidine base pair) mutation. This mutation was associated to signature 2, which had a very high frequency of T[C>T]A (41%), which has been attributed to activity of the AID/APOBEC family [60]. In fact, EGFR-E762K mutation has been observed in other cancer types associated with signature 2 [61]. The remaining of the top-10 mutations were associated to either C>T transitions (1 mutation in LUAD and 2 mutations in LSCC) or other nucleotide mutations (3 mutations in LSCC and 1 mutation in LUAD).

Next, EGFR mutations frequently observed in LUAD and LSCC patients were further analyzed. The T790M mutation, known to confer resistance to first-line targeted therapies in LUAD and LSCC, was ranked in positions 49th and 50th with a predicted likelihood of 0.015 and 0.011 in LUAD and LSCC, respectively. T790M is caused by a A[C>T]G nucleotide change, strongly associated with signature 1, which in turn correlate with age of diagnosis [39]. G719S is another EGFR mutation frequently observed in LUAD and LSCC patients. This mutation, ranked 79th in LUAD with a predicted likelihood of 0.010 and 118th with 0.007 likelihood in LSCC, is the result of a G[G>A]G nucleotide mutation, which has the highest probabilities in signatures 1, 6 and 16 (although the latest is not associated to LUAD). Therefore, we hypothesize that the emergence of this mutation can be associated to ageing (signature 1) and defective DNA mismatch repair (signature 6). Lack of association with signature 4 suggests that it is not particularly linked to tobacco smoking. Another interesting mutation is the recurrently reported R776H mutation, which activates EGFR in the absence of the activating EGF ligand R776H [62, 63]. This mutation was ranked 64th and 65th, with a predicted likelihood of 0.012 and 0.010 in LUAD and LSCC, respectively. R776H is caused by a C[G>A]C nucleotide mutation, strongly associated with signature 11. However, since this signature is not present in LUAD nor LSCC, the predicted probability value is the result of the sum of mild probabilities

of C[G>A]C in signatures 1, 2, 4 and 5. Consequently, this mutation is not particularly associated with any specific mechanism of mutation. Other clinically reported mutations such as G719A or G857V appeared beyond the top 100 mutations and were not particularly associated with any signature significantly contributing to either LUAD or LSCC.

Prediction of likely resistant EGFR mutations in gefitinib binding-site

We applied the aa-RFC to predict the resistance score of the amino acid mutations for the binding of gefitinib (Fig. 3c). There was not observed correlation between the two predicted scores (Pearson correlation coefficient = -0.05). The red area gathered a total amount of 39 *likely-and-resistant* mutations (*i.e.,* mutations that are very likely to arise and predicted to confer resistance). Examples of these mutations included M793L, G719S, H835Y, G796V, D855N, G796V or C775Y, among others. This representation allowed for the identification of those mutations with high likelihood and high resistance potential. The analysis the number of mutations and mean normalized resistance score (<<NRS>>) values associated to each phenotypic class revealed similar predictive trends than the observed in the original training set. A total amount of 171 mutations (46%) were predicted to belong to the RES class (<<NRS>> 0.52 ± 0.13). The SRES class was the second in number of predicted mutations. It had 124 mutations (35%) with an <<NRS>> of 0.57 ± 0.13. The ISEN class had 72 instances (19%), with an average resistance score of 0.28 ± 0.09. None of the mutations were predicted to belong to the NEU class.

Mapping of likelihood and resistance-likeness into the 3D structure of EGFR

Mapping of the amino acid accumulated resistance score and the resistance-likeness into the 3D structure of the EGFR kinase domain revealed the structural localization of the major players in gefitinib resistance (Fig 3d). Residues with warmer colours represented amino acids whose mutation is more prone to decrease the gefitinib binding affinity (*i.e.,* higher resistance score), while the thickness of the ribbons represented the accumulated likelihood of that particular amino acid. The D855, localized in the DFG motif, was the amino acid with highest accumulated resistance score. More specifically, the D855A mutation was ranked as the top gefitinib-resistant mutation (1.0 <<NRS>>). D855 has been previously reported to play a major role in

gefitinib binding [64], and consequently, its mutation will likely decrease binding affinity to gefitinib. Interestingly, another D855 mutant (D855N) was ranked also within the *likely-and-resistant* mutations in LUAD (Fig 3c). Other gefitinib-binding key residues such as L792 or M793 (both in the hinge region), were also among the top predicted mutations conferring resistance (*e.g.,* M793, which has a stable hydrogen bonding to the gefitinib molecule) and its mutation can lead to a significant drop in gefitinib binding affinity [65]. Some M793 mutants were also included in the LUAD *likely-and-resistant* group, such as the cases of M793L or M793I (Fig 3c). The L792P mutation in turn, will introduce the proline side chain into the hinge region of binding site. The distinctive cyclic structure of proline alongside its exceptional conformational rigidity can cause a steric clash between the proline side chain and gefitinib, with consequences for its binding. G719, localized in the phosphate-binding loop (P-loop), had several mutations among the top predicted mutants (G719V, 0.86 <NRS> SRES class; G719S, 0.83 <NRS> RES class) as well as mutations with lower predicted resistance potential (G719R 0.68 <NRS>, G719C 0.65, G719D 0.64 and G719A 0.63 all of them RES class) (Fig 3c, inner panel). Specifically, to the G719S mutation, it has been previously shown that EGFR-G719S mutant, in fact, increases gefitinib binding affinity [66]. Therefore, it appears that the classification of the G719S as RES class corresponds to a false positive prediction. The factors leading to this miss-prediction could include a wrong structural modeling of the mutation, which may be unable to completely capture the important rearrangement of the P-loop, and the fact that experimentally measured cases of glycine mutations are enriched in loss of affinity (in our training set: 3 ISEN, 2 NEU, 9 SRES and 18 RES). Other mutants such as G719A/C/D/R have been also associated to increased sensitivity to TKis [67], although results are contradictory and further confirmation is needed [68]. No G719V data associated response to gefitinib treatment was found in the literature. Unlike mutations in G719, T790M was correctly predicted to increase the binding affinity of gefitinib (0.35 <NRS>, ISEN class). This result agrees with the discovery of the mechanism of resistance of the EGFR-T790M mutation. T790M causes an increment of both ATP and gefitinib binding affinity. Interestingly, the increment in affinity is not uniform for both ATP and gefitinib, which is ultimately reflected in a lower $Kd/Km_{[ATP]}$ ratio, an estimator of inhibitory potency [69]. Similarly, the R776H mutation was also predicted to belong to the ISEN class (0.24 <NRS>). Experimental evidence found in the literature suggests that this mutation increases the sensitivity for TKis EGFR inhibitors [70, 71]. A summary of the predictions and the experimental data

associated with each mutation can be found in Table 1. Altogether, these results show that the aa-RFC can predict the mutation-induced phenotype, although individual interpretation of each case is required to further validate the predictions.

EGFR-binders insensitive to the resistance-like mutations

To test whether our approach is able to systematically predict insensitive compounds to the EGFR's *likely-and-resistant* mutations, we ran the lig-RFC predictor against all known EGFR reversible inhibitors with experimentally reported 3D structure (Fig 3e). The gefitinib lig-RFC predictions were consistent with the predictions from the aa-RFC. The only exception found was the gefitinib-M793L, which had considerably lower value than for the aa-RFC (aa-RFC <NRS> 0.85, lig-RFC <NRS> 0.41), yet being labelled as SRES. The <NRS> decrease can by explained by the fact that the lig-RFC weighted more the conservation of the hydrogen bonding by the mutant leucine. Erlotinib, another FDA approved EGFR TKi used in the treatment of NSCLC malignancies, resulted in a very similar mutational profile compared to gefitinib, which agrees with pervious published data [72].

T790M, M793L and R776H resulted in a low predicted resistant profile indicating that those mutations would confer increased sensitivity to many of the tested compounds. Conversely, other mutations, such as C775Y, resulted in a mixed profile conferring resistance to several of the tested compounds (e.g. CHEMBL2347963 or its structural analogue CHEMBL2347965) and increased sensitivity to others (e.g. CHEMBL2322330 and CHEMBL1229592). Finally, there were a total of six mutations with a highly drug-resistant profile (G796V, L792P, G719C/V, H835Y and D855A). These mutations were generally predicted as non-targetable, although a few exceptions were found. For instance, the CHEMBL1090356 compound had a <NRS> of 0.12, 0.20 and 0.14 for G796V, G719C/V mutations, respectively. In fact, this compound had the lowest resistance profile among all the screened set. Structural details revealed that CHEMBL1090356 has an imidazothiazole scaffold, with an amide group that lays deeply in the hydrophobic pocket and a morpholine tail that extends to a solved exposed region of the pocket [73] (Fig 3f). This mode of binding is significantly different to other reversible ATP-competitive inhibitors of EGFR and explains its predicted distinctive profile. We propose that this compound might be an alternative EGFR inhibitor to patients resistant to gefitinib therapy.

ERK2 predicted mutational landscape in melanoma and colorectal cancer

The predicted ERK2 mutational landscape revealed significant differences across the likelihood of mutations between melanoma and colorectal cancers. Indeed, the probabilities of mutations of amino acids involved in the binding site of VTX11e [74], a compound with anti-proliferative activity, was different in melanoma [75, 76] and colorectal adenocarcinoma [74] (Fig 4a and 4b). Such discrepancy was the result of completely different signatures contributing to the mutational landscape. While melanomas mutations are mainly coming from C>T transitions associated to signature 7, colorectal cancer mutations are the result of multiple mechanisms associated to signatures 1,5, 6 and 10. Melanomas predicted likelihood fitted into in a long tailed distribution, with enrichment in C>T mutations (Fig 4a). More specifically, there were nine possible amino acid mutations originated from C/T[C>T]N changes; and all of them were ranked within the top-10 likely mutations (S153F, P58L, P58S, S29L, L112F, S41F, P152L, L150F, L107F) (Fig 4a. inner panel). The remainding C>T mutations were also enriched among the top-50 most likely set (C>T odd ratio = 15.4). Conversely, colorectal cancer resulted in a more heterogeneous predicted mutational landscape (Figure 4b). The two most likely mutations (L112I and S41Y) were coming from T[C>A]T mutations associated with signature 10, which has been proposed to be caused by altered activity of the error-prone polymerase POLE [77]. Furthermore, mutations resulting from C>T transitions were also enriched among the top-50 likely mutations (C>T odds ratio = 58.4). In fact, 7 out of the top-10 most likely mutations were the result C>T transitions (M38I, M108L, G85R, G169S, S29L, G34S and G37S) (Fig 4b, inner panel). Unlike melanoma, colorectal cancer C>T mutations were associated to multiple signatures, including signatures 1, 6 and 10.

Prediction of likely resistant mutations in ERK2-VTX11e binding-site

The *resistance-likeness* of all ERK2 amino acid mutations in the binding site of VTX11e was calculated using the aa-RFC classifier. The predictive pattern was consistent with the predictions in the training set and the EGFR case. There were 171 (40%) mutations classified as RES (0.48 <NRS> ± 0.15), 159 (38%) classified as SRES (0.43 <NRS> ± 0.14), 93 (21.9%) as ISEN (0.25 <NRS> ± 0.10) and 1 (0.1%) as NEU (0.17 <NRS>). Consistent with the observed for EGFR, the predicted likelihood and the <NRS> scores did no correlate (Fig 4c, Pearson Correlation Score

of 0.03 in melanoma and 0.01 in colorectal). Variations in the mutational landscape between the two cancer types were also demonstrated in the differences in the set of top *likely-and-resistant* mutations. Melanomas resulted in 79 mutations, including P58S/L/T, L150F, L107F, P152S/L, L157F, L112P, I84N, F168Y or G37S among others, as likely for the cancer type and predicted to confer resistance to VTX11e. There were 86 mutations, including G34S, G37S, H147Y, P152S, E33K, L155P, P58L/S/T or K114R among others, as likely to appear in colorectal cancer and predicted to confer resistance to VTX11e. Only 29 of the mutations were shared between the two *likely-and-resistant* groups.

Mapping of likelihood and resistance-likeness into the 3D structure of ERK2

The significant discrepancies observed between the two cancer types were also observed in the 3D mapping of the mutations into the target structure of ERK2 (Fig 4d). Specifically, the significantly higher median likelihood observed in colorectal cancer (11.5 fold increase, colorectal median likelihood $2.6e^{-3}$; melanoma median likelihood $0.2e^{-3}$) was represented into the 3D space as thicker ribbons along the binding site of VTX11e. Similarly to the EGFR case, not a particular structural pattern was observed hosting the most likely mutations. Additionally, mapping of the amino acid accumulated resistance score into the ERK2 3D structure of the ERK2 kinase domain revealed the structural localization of those residues more prone to decrease VTX11e binding affinity (Fig 4d). Residues in the hinge region of the ATP binding-site showed the highest resistance scores. This region hosts the M108 residue, which is equivalent to the EGFR-M793, and is the major responsible of the hydrogen bonding between ERK2 and VTX11e. Examples of likely mutations of this amino acid included M108L (0.89 <NRS>, SRES class) and M108I (0.55 <NRS>, SRES class), being the later also included in the top *likely-and-resistant* group in colorectal cancer. ERK2-L107 was also predicted as one of the major contributors to resistance. Mutations of these amino acids included the L107P (1.0 <NRS>, SRES class) or L107F (0.41 <NRS>, RES class), being the later included in the *likely-and-resistant* set in melanoma. The importance of D167, localized at the DFG motif and structurally equivalent to the EGFR-D855, explains the high resistance score of the D167A mutation (<NRS> 0.97, SRES class). These residues were localized in the ATP-binding site of ERK2 and their potential to confer resistance might be explained by their ATP-binding site structural similarity with EGFR.

Proline 58 mutations were also classified as highly resistance-like. More specifically, P58L/S/T (0.60 <NRS>, 0.70 <NRS> and 0.67 <NRS>; SRES, SRES and RES class respectively) mutations were reported within the *likely-and-resistant* group in melanoma and colorectal cancer, suggesting these mutations are critical. These predictions agreed with the evidence of ERK2-P58L/S/T mutations found in VTX11e-resistant A375 melanoma cell line [78]. A complete summary of the VTX11e-resistant mutations previously described [78] alongside their predicted likelihood and resistance-likeliness is shown in Table2 and Fig 4c inner set. All the experimentally found resistant mutations were predicted as either SRES or RES by our model. Moreover, 5 out of 8 (62%) of the mutations were correctly predicted to belong to melanoma *likely-and-resistant* group. Altogether, these results probed the ability of the method to detect resistance-like mutations to ERK2-VTX11e interaction. Interestingly, the 3D mapping of the mutations from [78] revealed their clustering into an adjacent pocket to the ERK2 ATP binding site, which highlights the presence of ATP had an essential role in the emergence of mutations conferring resistance in ATP-competitive inhibitors. Other mutations ranked in the top 10 resistance-like mutations and not present in [78] included H147Y, I86M, L150P, G34V, F168I and E33D (Fig 4c inner set). Unfortunately, no experimental data was available at the time to confirm the resistance potential of these mutants.

ERK2-binders insensitive to the resistance-like mutations

Next, the lig-RFC classifier was applied to existing ERK2 reversible inhibitors to identify insensitive compounds to the resistance-like mutations previously identified. In this case, the limited number of co-crystallized ERK2 inhibitors, prompted us to extend the search to any known ERK2 inhibitor (see methods *Creating the dataset of candidate molecules*). Similarly to the EGFR example, some mutations had a highly resistance-like profile with a very limited number of compounds with low predicted sensitivity (Fig 4e). Such were the cases of L107F/P, I86M or P58S/T/L; which had few compounds with NSR below the average (0.50 <NRS> ±0.16). DEL22379 was one of the few compounds with low predicted sensitivity to highly resistant mutations. Interestingly, this compound resulted a highly insensitive profile among the all the screened mutations. DEL223790 unique sensitivity profile is explained by its completely different mode of action: it binds the ERK2 interface preventing its dimerization [79]. Other mutations resulted in low resistance-likeness profile,

including Y36N/H or C65Y. The results of the C65Y mutation were consistent with the predictions from the aa-RFC, which scored this mutation with a low <NRS> (Table 2). However, the Y36N/H predictions generally resulted in lower <NRS>. For instance, the control compound VTX11e, resulted in a lig-RFC <NRS> of 0.41 (Y36N) and 0.35 (Y36H) while the aa-RFC scored them with 0.58 and 0.66. Despite of the decrease in the <NRS>, the predicted class was maintained in both classifiers as SRES. The differences between the two classifiers might be caused by the fact that the lig-RFC does not contain all the amino acid based features used in the aa-RFC. Finally, the G37S mutation, which had previously been identified as resistant [78], was predicted to be in the *likely-and-resistant* group in both melanoma and colorectal cancer. G37 is localized in the ERK2 P-loop, and we hypothesize it may play an important role in the orientation of Y36 towards to the chlorobenzene group of VTX11e, which ultimately leads to the π-stacking interaction [80]. Remarkably, the lig-RFC provided several compounds with low resistance-likeness to G37S/V/C mutations. The compound with the lowest resistance profile for these mutations was E75 (Fig 4f, named as E75 due to their PDB accession code). The mutational profile of E75 had a <NRS> of 0.11, 0.08 and 0.06 for G37S/V/C, respectively. Unlike VTX11e, E75 is located distantly to the G37 residue, not interacting with the Y36 and mostly occupying the ERK2 hinge region (Fig 4f). Hence, the E75 binding mode might be compatible with G37 mutations, proposing an interesting candidate for overcoming resistance in tumors harboring ERK2-G37S/V/C mutations.

DISCUSSION

We have introduced a computational model that predicts the cancer-associated likelihood and the resistance-likeness of mutations in targets of small molecule targeted cancer therapies. Our approach first defines the mutational likelihood of amino acids involved in the binding of a small molecule drug. Our estimations rely on the tri-nucleotide mutational probabilities observed in the cancer-associated signatures previously described [35, 39]. We have shown that the EGFR mutational profile was not significantly different between LUAD and LSCC cancer types. Conversely, the ERK2 analysis revealed major discrepancies between melanoma and colorectal mutational landscape. Melanoma mutations are mainly originated from C>T transitions associated to ultraviolet light exposure. However, colorectal associated mutations are the result of more complex and heterogeneous processes. Interestingly, the discrepancies are also reflected to the global distribution of the probabilities. While melanoma seems to prioritize fewer ERK2-mutations with a very high likelihood, colorectal, presents higher median values with considerably lower peak values. The differences between the colorectal cancer and melanoma mutations are also reflected in the low overlapping between the likely-*and*-*resistant* groups of mutations. This result suggests that these two cancer types should have different pharmacological approaches to overcome resistance due to the spontaneous generation of cancer-associated mutations in the drug targets.

Interestingly, the nature of our approach enables the tracking of the association between each mutation and their underlying signatures, which ultimately can be translated into an individual mutation-mechanisms association. That is the case of the EGFR-T790M mutation, which we proposed to be mainly associated with ageing and not particularly linked to tobacco smoking. It is important to mention that our model only considers the probability of emergence of mutations in a cancer genomic context. Nevertheless, a significant number of mutations in a cancer cell can be also the result of germ-line variations or pre-malignant somatic mutations. For instance, it has been shown that the EGFR-T790M mutation can have both somatic and germ-line origin [81-83]. Another limitation of the likelihood model is that its predictions are based on the average probabilities from hundreds of samples for each cancer type. Therefore, the predicted likelihood shows global cancer trends but it is currently unable to capture specific trends in each individual cancer case. Future work might

thus focus on finding the mechanisms underlying each individual cancer case, which eventually would translate into the personalization of the likelihood predictions.

The structural mapping of the predicted likelihood did not reveal any association between the likelihood of an amino acid mutation and its structural localization. Perhaps, constraining the mutational likelihood with evolutionary restrains would lead towards an increase in less evolutionary conserved regions of the structure. Hence, the unfavorable phenotype linked to evolutionary restraints can partially explain the fact that some of the predicted mutations have not been observed in the clinic. This problem is chiefly evident in cancer, where tumor cell population has a fitness advantage over the healthy tissue. Another explanation is linked to the technical limitations of standard NGS sequencing of solid biopsies, which only allows for the detection of mutations present > 5% of tumor cells [7]. In fact, despite of tumors can harbour millions of mutations [22], only a small percentage of them are systematically reported. These low-frequency mutations may not have a critical effect during tumor progression, but the evolutionary pressure induced by a drug treatment regimen can transform them into drug resistance drivers. Thus, it is essential to detect not only the frequent cancer drivers but also the low-frequent mutations that can lead towards drug resistance. Recent studies using circulating tumor DNA (ctDNA) have shown very promising results for this purpose [84, 85]. However, there are many technological challenges to address prior to broader application of this technology. In the meantime, *in-silico* models can play a major role to comprehensively characterize the mutational burden of cancer samples.

We connected the mutational landscape of tumors with the drug-resistance phenotype due to spontaneously generated mutations in drug targets. To do so, the aa-RFC classifier predicts the effect of a single mutation to the drug binding affinity in a particular cancer target. The classifier was trained with the Platinum database [44], whose instances were split into four phenotypic classes depending of their drug binding affinity fold change. In our opinion, reducing the number of possible classes from four to two (*e.g.* into *loss-of-affinity* and *gain-of-affinity)* would increase the classifier performance, but it would also over simplify the spectrum of possible phenotypes. Evaluation of the performance of the aa-RFC showed that classes representing severe changes (*i.e.,* ISEN and SRES classes) outperformed those representing mild changes (*i.e.,* RES and NEU classes). More specifically, the lower performance of these classes is the result of over prediction towards the RES class

as well as under prediction of the NEU class. This limitation may be explained by the fact that many RES cases are very close to the NEU frontier (*i.e.,* cases with very small drop in affinity) and vice versa. In such cases, the classifier assigns the instances to the most populated class (*i.e.,* the RES class) since that is the one with higher probability. To address this limitation, we calculated the <NRS>, which provides a smoother way to assess the resistance-likeness by combining the confidence score of the four classes and correcting for the over-assignment of the most populated classes. To our knowledge this is the first method specifically developed to classify changes in drug binding affinity upon mutation. Comparison with a gold-standard methods for measuring drug-binding affinity revealed the difficulties of such methods in detecting large changes in affinity upon mutation. Rather, they are oriented to quantitatively estimate the drug binding affinity when the binding is known to occur.

Application of the aa-RFC to the EGFR and ERK2 cases showed its ability to identify the phenotype of previously reported mutations. Remarkably, the method correctly predicted the class of EGFR-T790M, conferring resistance by decreasing the $Kd/Km_{[ATP]}$ ratio; EGFR-R776H, ERK2-P58L/S/T or ERK2-G37S among others. However, it failed predicting the EGFR-G719S phenotype, which featured the problem that glycine mutations increasing the sensitivity of the drug are likely to be miss-classified. Additionally, our model proposed, in both cases, multiple new unseen mutations as candidates for conferring resistance to the studied treatments. Nevertheless, mutations negatively interfering with ATP might be non-functional. Mutations disrupting the ATP binding would lead to a non-functional protein kinase (*i.e.,* loss of function mutations), which is incompatible with their role in cancer progression. This hypothesis is also supported by previous findings indicating a cluster of ERK resistant mutations in an allosteric region next to the ATP binding site [78]. Moreover, our findings might explain why mutants of amino acids with an essential role in ATP binding, such as EGFR-L792 (ERK2-L107) or EGFR-M793 (ERK2-M108), have not yet been reported in the clinics.

The last step of the model application consisted on the search for non-resistant molecules to the mutations detected by the aa-RFC. To do so, we used a lighter and more ligand centric version of the aa-RFC called lig-RFC. The performance of both classifiers is also illustrated by the consistency of the EGFR-gefitinib and ERK2-VTX11e predictions. However, small discrepancies in the <NRS> score were

observed for the ERK2Y36H/N and EGFR-M793L mutations. In both cases the differences respond to the fact that the lig-RFC weights more the ligand-based features (*e.g.*, hydrogen bonding conservation) that changes in the amino acid biochemical properties. It is essential to mention that our models (the aa-RFC or lig-RFC) do not consider information relative to covalent bonding between the compound and the protein target. That explains why EGFR irreversible inhibitors such as afatinib, olmutinib, rociletinib or WZ-3146 were not included in the study.

In both, ERK2 and EGFR cases, we generally observed three groups of mutations. The first group is composed by those mutations with very limited or no compounds with low resistance score. That is, these mutations are generally predicted as non-targetable. However, some exceptions were found, including the dimerization inhibitor del22379 predicted to be insensitive to the majority of the ERK2-mutations. Similarly, CHEMBL1090356 was predicted as insensitive to several of the highly resistance-like EGFR-mutations. These two cases are the result of very distinctive mode of actions, which ultimately was reflected in their resistance profile. Other group of mutations was composed by those predicted to increase the affinity of most of the screened compounds. Interestingly, despite of EGFR-T790M being known to confer resistance to most of EGFR reversible inhibitors, it was classified into this group of mutations. This is because this mutation confers resistance by decreasing the $Kd/Km_{[ATP]}$ ratio. Therefore, non-resistant reversible inhibitors need to maintain such ratio to conserve their inhibitory potency. The last group of mutations are those with heterogeneous profile. This group is probably the most interesting from a resistance perspective, since they have molecules with highly diverse predicted phenotypes. The diversity alongside the structural nature of our predictions facilitated the study of their mechanism of resistance. For instance, thorough the ERK2-G37S example, we showed the ability of the method in forecasting non-resistant alternatives cancer targeted therapies. Future applications of the model would benefit from the inclusion of both new candidate molecules and information about resistant mutants. Moreover, further application in other systems would ultimately lead towards a comprehensive characterization of the resistant mutational landscape of targeted cancer therapies. To achieve this goal, it is also important that some of our predictions get validated. We encourage the scientific community to experimentally validate our predictions to get closer to one of the final goals in cancer treatment: the personalized design of non-resistant cancer therapies.

CONCLUSIONS

This manuscript presents a computational framework that aims to connect the mutational landscape of tumors with the drug-resistance phenotype generated by cancer-associated mutations in drug targets. We first introduced a computational model that predicts the probability of generation of mutations in anticancer drug targets. Application of the likelihood model demonstrated that the mutational profile of drug targets is cancer specific. We then introduced a RFC that uses structural information to predict the drug binding affinity change upon mutation. To our knowledge, this is the first classifier used to systematically detect the resistance-associated phenotype caused by mutations in drug targets. Application of the model to the EGFR-gefitinib and ERK2-VTX11e targeted cancer therapies identified some of the known resistant mutants alongside other new unseen mutations predicted to confer resistance to these therapies. Interestingly, the structural localization of most of the known resistance mutations suggests that ATP plays an essential role in the emergence resistant mutants. Consequently, some of the predicted unseen resistance mutations can be indeed non functional. Finally, the search for alternative non-resistant treatments provided a sensitivity map that connects the pharmacological space with the mutational landscape of EGFR and ERK2. Thorough several concrete examples we exemplified how this model can be used to rational design personalized non-resistant cancer therapies. Further application in other systems will ultimately lead towards a comprehensive characterization of the resistant mutational landscape of targeted cancer therapies.

REFERENCES

1. Vadlapatla RK, Vadlapudi AD, Pal D, Mitra AK: **Mechanisms of drug resistance in cancer chemotherapy: coordinated role and regulation of efflux transporters and metabolizing enzymes**. *Current pharmaceutical design* 2013, **19**(40):7126-7140.
2. Sawyers C: **Targeted cancer therapy**. *Nature* 2004, **432**(7015):294-297.
3. Gonzalez-Angulo AM, Hortobagyi GN, Ellis LM: **Targeted therapies: peaking beneath the surface of recent bevacizumab trials**. *Nature reviews Clinical oncology* 2011, **8**(6):319-320.
4. Morita S, Okamoto I, Kobayashi K, Yamazaki K, Asahina H, Inoue A, Hagiwara K, Sunaga N, Yanagitani N, Hida T *et al*: **Combined survival analysis of prospective clinical trials of gefitinib for non-small cell lung cancer with EGFR mutations**. *Clin Cancer Res* 2009, **15**(13):4493-4498.
5. Chapman PB, Hauschild A, Robert C, Haanen JB, Ascierto P, Larkin J, Dummer R, Garbe C, Testori A, Maio M *et al*: **Improved survival with vemurafenib in melanoma with BRAF V600E mutation**. *The New England journal of medicine* 2011, **364**(26):2507-2516.
6. Al-Lazikani B, Banerji U, Workman P: **Combinatorial drug therapy for cancer in the post-genomic era**. *Nat Biotechnol* 2012, **30**(7):679-692.
7. Schmitt MW, Loeb LA, Salk JJ: **The influence of subclonal resistance mutations on targeted cancer therapy**. *Nat Rev Clin Oncol* 2016, **13**(6):335-347.
8. Holohan C, Van Schaeybroeck S, Longley DB, Johnston PG: **Cancer drug resistance: an evolving paradigm**. *Nature reviews Cancer* 2013, **13**(10):714-726.
9. Debatin KM, Krammer PH: **Death receptors in chemotherapy and cancer**. *Oncogene* 2004, **23**(16):2950-2966.
10. Lowe SW, Cepero E, Evan G: **Intrinsic tumour suppression**. *Nature* 2004, **432**(7015):307-315.
11. Triller N, Korosec P, Kern I, Kosnik M, Debeljak A: **Multidrug resistance in small cell lung cancer: expression of P-glycoprotein, multidrug resistance protein 1 and lung resistance protein in chemo-naive patients and in relapsed disease**. *Lung cancer* 2006, **54**(2):235-240.
12. Gottesman MM, Fojo T, Bates SE: **Multidrug resistance in cancer: role of ATP-dependent transporters**. *Nature reviews Cancer* 2002, **2**(1):48-58.
13. Sharma SV, Lee DY, Li B, Quinlan MP, Takahashi F, Maheswaran S, McDermott U, Azizian N, Zou L, Fischbach MA *et al*: **A chromatin-mediated reversible drug-tolerant state in cancer cell subpopulations**. *Cell* 2010, **141**(1):69-80.
14. Bell DW, Gore I, Okimoto RA, Godin-Heymann N, Sordella R, Mulloy R, Sharma SV, Brannigan BW, Mohapatra G, Settleman J *et al*: **Inherited susceptibility to lung cancer may be associated with the T790M drug resistance mutation in EGFR**. *Nature genetics* 2005, **37**(12):1315-1316.
15. Shih JY, Gow CH, Yang PC: **EGFR mutation conferring primary resistance to gefitinib in non-small-cell lung cancer**. *The New England journal of medicine* 2005, **353**(2):207-208.

16. Walter AO, Sjin RT, Haringsma HJ, Ohashi K, Sun J, Lee K, Dubrovskiy A, Labenski M, Zhu Z, Wang Z *et al*: **Discovery of a mutant-selective covalent inhibitor of EGFR that overcomes T790M-mediated resistance in NSCLC**. *Cancer Discov* 2013, **3**(12):1404-1415.
17. Cross DA, Ashton SE, Ghiorghiu S, Eberlein C, Nebhan CA, Spitzler PJ, Orme JP, Finlay MR, Ward RA, Mellor MJ *et al*: **AZD9291, an irreversible EGFR TKI, overcomes T790M-mediated resistance to EGFR inhibitors in lung cancer**. *Cancer Discov* 2014, **4**(9):1046-1061.
18. Liao BC, Lin CC, Yang JC: **Second and third-generation epidermal growth factor receptor tyrosine kinase inhibitors in advanced nonsmall cell lung cancer**. *Curr Opin Oncol* 2015, **27**(2):94-101.
19. Barouch-Bentov R, Sauer K: **Mechanisms of drug resistance in kinases**. *Expert Opin Investig Drugs* 2011, **20**(2):153-208.
20. Duong-Ly KC, Devarajan K, Liang S, Horiuchi KY, Wang Y, Ma H, Peterson JR: **Kinase Inhibitor Profiling Reveals Unexpected Opportunities to Inhibit Disease-Associated Mutant Kinases**. *Cell reports* 2016, **14**(4):772-781.
21. Fisher R, Pusztai L, Swanton C: **Cancer heterogeneity: implications for targeted therapeutics**. *Br J Cancer* 2013, **108**(3):479-485.
22. Ling S, Hu Z, Yang Z, Yang F, Li Y, Lin P, Chen K, Dong L, Cao L, Tao Y *et al*: **Extremely high genetic diversity in a single tumor points to prevalence of non-Darwinian cell evolution**. *Proc Natl Acad Sci U S A* 2015, **112**(47):E6496-6505.
23. Schmitt MW, Kennedy SR, Salk JJ, Fox EJ, Hiatt JB, Loeb LA: **Detection of ultra-rare mutations by next-generation sequencing**. *Proc Natl Acad Sci U S A* 2012, **109**(36):14508-14513.
24. Lipinski KA, Barber LJ, Davies MN, Ashenden M, Sottoriva A, Gerlinger M: **Cancer Evolution and the Limits of Predictability in Precision Cancer Medicine**. *Trends Cancer* 2016, **2**(1):49-63.
25. Morrissy AS, Garzia L, Shih DJ, Zuyderduyn S, Huang X, Skowron P, Remke M, Cavalli FM, Ramaswamy V, Lindsay PE *et al*: **Divergent clonal selection dominates medulloblastoma at recurrence**. *Nature* 2016, **529**(7586):351-357.
26. Altrock PM, Liu LL, Michor F: **The mathematics of cancer: integrating quantitative models**. *Nature reviews Cancer* 2015, **15**(12):730-745.
27. Anderson AR, Weaver AM, Cummings PT, Quaranta V: **Tumor morphology and phenotypic evolution driven by selective pressure from the microenvironment**. *Cell* 2006, **127**(5):905-915.
28. Williams MJ, Werner B, Barnes CP, Graham TA, Sottoriva A: **Identification of neutral tumor evolution across cancer types**. *Nat Genet* 2016, **48**(3):238-244.
29. Attolini CS, Cheng YK, Beroukhim R, Getz G, Abdel-Wahab O, Levine RL, Mellinghoff IK, Michor F: **A mathematical framework to determine the temporal sequence of somatic genetic events in cancer**. *Proc Natl Acad Sci U S A* 2010, **107**(41):17604-17609.
30. Komarova NL, Wodarz D: **Drug resistance in cancer: principles of emergence and prevention**. *Proc Natl Acad Sci U S A* 2005, **102**(27):9714-9719.

31. Iwasa Y, Nowak MA, Michor F: **Evolution of resistance during clonal expansion**. *Genetics* 2006, **172**(4):2557-2566.
32. Chmielecki J, Foo J, Oxnard GR, Hutchinson K, Ohashi K, Somwar R, Wang L, Amato KR, Arcila M, Sos ML *et al*: **Optimization of dosing for EGFR-mutant non-small cell lung cancer with evolutionary cancer modeling**. *Sci Transl Med* 2011, **3**(90):90ra59.
33. Siravegna G, Mussolin B, Buscarino M, Corti G, Cassingena A, Crisafulli G, Ponzetti A, Cremolini C, Amatu A, Lauricella C *et al*: **Clonal evolution and resistance to EGFR blockade in the blood of colorectal cancer patients**. *Nature medicine* 2015.
34. Bozic I, Reiter JG, Allen B, Antal T, Chatterjee K, Shah P, Moon YS, Yaqubie A, Kelly N, Le DT *et al*: **Evolutionary dynamics of cancer in response to targeted combination therapy**. *Elife* 2013, **2**:e00747.
35. Alexandrov LB, Nik-Zainal S, Wedge DC, Aparicio SA, Behjati S, Biankin AV, Bignell GR, Bolli N, Borg A, Borresen-Dale AL *et al*: **Signatures of mutational processes in human cancer**. *Nature* 2013, **500**(7463):415-421.
36. Qin J, Xin H, Nickoloff BJ: **Specifically targeting ERK1 or ERK2 kills melanoma cells**. *J Transl Med* 2012, **10**:15.
37. Wong DJ, Robert L, Atefi MS, Lassen A, Avarappatt G, Cerniglia M, Avramis E, Tsoi J, Foulad D, Graeber TG *et al*: **Erratum to: Antitumor activity of the ERK inhibitor SCH722984 against BRAF mutant, NRAS mutant and wild-type melanoma**. *Mol Cancer* 2015, **14**:128.
38. Hatzivassiliou G, Liu B, O'Brien C, Spoerke JM, Hoeflich KP, Haverty PM, Soriano R, Forrest WF, Heldens S, Chen H *et al*: **ERK inhibition overcomes acquired resistance to MEK inhibitors**. *Mol Cancer Ther* 2012, **11**(5):1143-1154.
39. Alexandrov LB, Jones PH, Wedge DC, Sale JE, Campbell PJ, Nik-Zainal S, Stratton MR: **Clock-like mutational processes in human somatic cells**. *Nat Genet* 2015, **47**(12):1402-1407.
40. Webb B, Sali A: **Protein structure modeling with MODELLER**. *Methods Mol Biol* 2014, **1137**:1-15.
41. Webb B, Sali A: **Comparative Protein Structure Modeling Using MODELLER**. *Curr Protoc Bioinformatics* 2014, **47**:5 6 1-32.
42. Frank E, Hall M, Trigg L, Holmes G, Witten IH: **Data mining in bioinformatics using Weka**. *Bioinformatics* 2004, **20**(15):2479-2481.
43. Liaw A, Wiener M: **Classification and Regression by randomForest.** *R News* 2002, **2**(3):18–22.
44. Pires DE, Blundell TL, Ascher DB: **Platinum: a database of experimentally measured effects of mutations on structurally defined protein-ligand complexes**. *Nucleic Acids Res* 2015, **43**(Database issue):D387-391.
45. DeLano WL: **The PyMOL Molecular Graphics System on World Wide Web** http://www.pymol.org. In.; 2002.
46. Kabsch W, Sander C: **Dictionary of protein secondary structure: pattern recognition of hydrogen-bonded and geometrical features**. *Biopolymers* 1983, **22**(12):2577-2637.

47. Hamelryck T: **An amino acid has two sides: a new 2D measure provides a different view of solvent exposure**. *Proteins* 2005, **59**(1):38-48.
48. Cock PJ, Antao T, Chang JT, Chapman BA, Cox CJ, Dalke A, Friedberg I, Hamelryck T, Kauff F, Wilczynski B *et al*: **Biopython: freely available Python tools for computational molecular biology and bioinformatics**. *Bioinformatics* 2009, **25**(11):1422-1423.
49. Capriotti E, Fariselli P, Casadio R: **I-Mutant2.0: predicting stability changes upon mutation from the protein sequence or structure**. *Nucleic acids research* 2005, **33**(Web Server issue):W306-310.
50. Altschul SF, Gish W, Miller W, Myers EW, Lipman DJ: **Basic local alignment search tool**. *Journal of molecular biology* 1990, **215**(3):403-410.
51. Henikoff S, Henikoff JG: **Amino acid substitution matrices from protein blocks**. *Proceedings of the National Academy of Sciences of the United States of America* 1992, **89**(22):10915-10919.
52. Kojetin DJ, Thompson RJ, Cavanagh J: **Sub-classification of response regulators using the surface characteristics of their receiver domains**. *FEBS letters* 2003, **554**(3):231-236.
53. Liu T, Lin Y, Wen X, Jorissen RN, Gilson MK: **BindingDB: a web-accessible database of experimentally determined protein-ligand binding affinities**. *Nucleic Acids Res* 2007, **35**(Database issue):D198-201.
54. Salentin S, Schreiber S, Haupt VJ, Adasme MF, Schroeder M: **PLIP: fully automated protein-ligand interaction profiler**. *Nucleic Acids Res* 2015, **43**(W1):W443-447.
55. Gaulton A, Bellis LJ, Bento AP, Chambers J, Davies M, Hersey A, Light Y, McGlinchey S, Michalovich D, Al-Lazikani B *et al*: **ChEMBL: a large-scale bioactivity database for drug discovery**. *Nucleic acids research* 2012, **40**(Database issue):D1100-1107.
56. Trott O, Olson AJ: **AutoDock Vina: Improving the speed and accuracy of docking with a new scoring function, efficient optimization, and multithreading**. *Journal of computational chemistry* 2009.
57. Cohen J: **A Coefficient of Agreement for Nominal Scales**. *Educational and Psychological Measurement* 1960, **20**(1):37-46.
58. Menze BH, Kelm BM, Masuch R, Himmelreich U, Bachert P, Petrich W, Hamprecht FA: **A comparison of random forest and its Gini importance with standard chemometric methods for the feature selection and classification of spectral data**. *BMC Bioinformatics* 2009, **10**:213.
59. Alexandrov LB, Ju YS, Haase K, Van Loo P, Martincorena I, Nik-Zainal S, Totoki Y, Fujimoto A, Nakagawa H, Shibata T *et al*: **Mutational signatures associated with tobacco smoking in human cancer**. *bioRxiv* 2016.
60. Nik-Zainal S, Alexandrov LB, Wedge DC, Van Loo P, Greenman CD, Raine K, Jones D, Hinton J, Marshall J, Stebbings LA *et al*: **Mutational processes molding the genomes of 21 breast cancers**. *Cell* 2012, **149**(5):979-993.
61. Weber F, Fukino K, Sawada T, Williams N, Sweet K, Brena RM, Plass C, Caldes T, Mutter GL, Villalona-Calero MA *et al*: **Variability in organ-specific EGFR mutational spectra in tumour epithelium and stroma**

may be the biological basis for differential responses to tyrosine kinase inhibitors. *Br J Cancer* 2005, **92**(10):1922-1926.
62. Ruan Z, Kannan N: **Mechanistic Insights into R776H Mediated Activation of Epidermal Growth Factor Receptor Kinase**. *Biochemistry* 2015, **54**(27):4216-4225.
63. van Noesel J, van der Ven WH, van Os TA, Kunst PW, Weegenaar J, Reinten RJ, Kancha RK, Duyster J, van Noesel CJ: **Activating germline R776H mutation in the epidermal growth factor receptor associated with lung cancer with squamous differentiation**. *Journal of clinical oncology : official journal of the American Society of Clinical Oncology* 2013, **31**(10):e161-164.
64. La Motta C, Sartini S, Tuccinardi T, Nerini E, Da Settimo F, Martinelli A: **Computational studies of epidermal growth factor receptor: docking reliability, three-dimensional quantitative structure-activity relationship analysis, and virtual screening studies**. *J Med Chem* 2009, **52**(4):964-975.
65. Doss GP, Rajith B, Chakraborty C, NagaSundaram N, Ali SK, Zhu H: **Structural signature of the G719S-T790M double mutation in the EGFR kinase domain and its response to inhibitors**. *Scientific reports* 2014, **4**:5868.
66. Yoshikawa S, Kukimoto-Niino M, Parker L, Handa N, Terada T, Fujimoto T, Terazawa Y, Wakiyama M, Sato M, Sano S *et al*: **Structural basis for the altered drug sensitivities of non-small cell lung cancer-associated mutants of human epidermal growth factor receptor**. *Oncogene* 2013, **32**(1):27-38.
67. Taron M, Ichinose Y, Rosell R, Mok T, Massuti B, Zamora L, Mate JL, Manegold C, Ono M, Queralt C *et al*: **Activating mutations in the tyrosine kinase domain of the epidermal growth factor receptor are associated with improved survival in gefitinib-treated chemorefractory lung adenocarcinomas**. *Clin Cancer Res* 2005, **11**(16):5878-5885.
68. Wu JY, Yu CJ, Chang YC, Yang CH, Shih JY, Yang PC: **Effectiveness of tyrosine kinase inhibitors on "uncommon" epidermal growth factor receptor mutations of unknown clinical significance in non-small cell lung cancer**. *Clin Cancer Res* 2011, **17**(11):3812-3821.
69. Yun CH, Mengwasser KE, Toms AV, Woo MS, Greulich H, Wong KK, Meyerson M, Eck MJ: **The T790M mutation in EGFR kinase causes drug resistance by increasing the affinity for ATP**. *Proc Natl Acad Sci U S A* 2008, **105**(6):2070-2075.
70. Beau-Faller M, Prim N, Ruppert AM, Nanni-Metellus I, Lacave R, Lacroix L, Escande F, Lizard S, Pretet JL, Rouquette I *et al*: **Rare EGFR exon 18 and exon 20 mutations in non-small-cell lung cancer on 10 117 patients: a multicentre observational study by the French ERMETIC-IFCT network**. *Ann Oncol* 2014, **25**(1):126-131.
71. Kobayashi S, Canepa HM, Bailey AS, Nakayama S, Yamaguchi N, Goldstein MA, Huberman MS, Costa DB: **Compound EGFR mutations and response to EGFR tyrosine kinase inhibitors**. *J Thorac Oncol* 2013, **8**(1):45-51.

72. Nguyen KS, Kobayashi S, Costa DB: **Acquired resistance to epidermal growth factor receptor tyrosine kinase inhibitors in non-small-cell lung cancers dependent on the epidermal growth factor receptor pathway**. *Clin Lung Cancer* 2009, **10**(4):281-289.
73. Fidanze SD, Erickson SA, Wang GT, Mantei R, Clark RF, Sorensen BK, Bamaung NY, Kovar P, Johnson EF, Swinger KK *et al*: **Imidazo[2,1-b]thiazoles: multitargeted inhibitors of both the insulin-like growth factor receptor and members of the epidermal growth factor family of receptor tyrosine kinases**. *Bioorg Med Chem Lett* 2010, **20**(8):2452-2455.
74. Aronov AM, Tang Q, Martinez-Botella G, Bemis GW, Cao J, Chen G, Ewing NP, Ford PJ, Germann UA, Green J *et al*: **Structure-guided design of potent and selective pyrimidylpyrrole inhibitors of extracellular signal-regulated kinase (ERK) using conformational control**. *J Med Chem* 2009, **52**(20):6362-6368.
75. Gonzalez-Cao M, Rodon J, Karachaliou N, Sanchez J, Santarpia M, Viteri S, Pilotto S, Teixido C, Riso A, Rosell R: **Other targeted drugs in melanoma**. *Ann Transl Med* 2015, **3**(18):266.
76. Morris EJ, Jha S, Restaino CR, Dayananth P, Zhu H, Cooper A, Carr D, Deng Y, Jin W, Black S *et al*: **Discovery of a novel ERK inhibitor with activity in models of acquired resistance to BRAF and MEK inhibitors**. *Cancer Discov* 2013, **3**(7):742-750.
77. Shinbrot E, Henninger EE, Weinhold N, Covington KR, Goksenin AY, Schultz N, Chao H, Doddapaneni H, Muzny DM, Gibbs RA *et al*: **Exonuclease mutations in DNA polymerase epsilon reveal replication strand specific mutation patterns and human origins of replication**. *Genome Res* 2014, **24**(11):1740-1750.
78. Goetz EM, Ghandi M, Treacy DJ, Wagle N, Garraway LA: **ERK mutations confer resistance to mitogen-activated protein kinase pathway inhibitors**. *Cancer research* 2014, **74**(23):7079-7089.
79. Herrero A, Pinto A, Colon-Bolea P, Casar B, Jones M, Agudo-Ibanez L, Vidal R, Tenbaum SP, Nuciforo P, Valdizan EM *et al*: **Small Molecule Inhibition of ERK Dimerization Prevents Tumorigenesis by RAS-ERK Pathway Oncogenes**. *Cancer Cell* 2015, **28**(2):170-182.
80. Chaikuad A, Tacconi EM, Zimmer J, Liang Y, Gray NS, Tarsounas M, Knapp S: **A unique inhibitor binding site in ERK1/2 is associated with slow binding kinetics**. *Nat Chem Biol* 2014, **10**(10):853-860.
81. Gazdar A, Robinson L, Oliver D, Xing C, Travis WD, Soh J, Toyooka S, Watumull L, Xie Y, Kernstine K *et al*: **Hereditary lung cancer syndrome targets never smokers with germline EGFR gene T790M mutations**. *J Thorac Oncol* 2014, **9**(4):456-463.
82. Lou Y, Pecot CV, Tran HT, DeVito VJ, Tang XM, Heymach JV, Luthra R, Wistuba, II, Zuo Z, Tsao AS: **Germline Mutation of T790M and Dual/Multiple EGFR Mutations in Patients With Lung Adenocarcinoma**. *Clin Lung Cancer* 2016, **17**(2):e5-e11.
83. Yu HA, Arcila ME, Harlan Fleischut M, Stadler Z, Ladanyi M, Berger MF, Robson M, Riely GJ: **Germline EGFR T790M mutation found in multiple members of a familial cohort**. *J Thorac Oncol* 2014, **9**(4):554-558.

84. Bettegowda C, Sausen M, Leary RJ, Kinde I, Wang Y, Agrawal N, Bartlett BR, Wang H, Luber B, Alani RM *et al*: **Detection of circulating tumor DNA in early- and late-stage human malignancies**. *Sci Transl Med* 2014, **6**(224):224ra224.
85. Sacher AG, Paweletz C, Dahlberg SE, Alden RS, O'Connell A, Feeney N, Mach SL, Janne PA, Oxnard GR: **Prospective Validation of Rapid Plasma Genotyping for the Detection of EGFR and KRAS Mutations in Advanced Lung Cancer**. *JAMA Oncol* 2016.

TABLES

Table 1. Summary of the EGFR mutations discussed in the manuscript alongside their aa-RFC predicted phenotype and, when available, the experimentally reported effect to gefitinib treatment found in literature.

Mutation	NMR	Predicted class	*Gefitinib phenotype*
L792P	0.88	SRES	Proposed resistant (Unconfirmed)
M793L	0.85	SRES	
D855N	1.0	SRES	
G719S	0.83	RES	Increase Sensitivity to gefitinib and erlotinib
G719A	0.63	RES	Contradictory
G719R	0.68	RES	
G719C	0.65	RES	
G719D	0.64	RES	
G719V	0.86	SRES	
T790M	0.34	ISEN	Increase Sensitivity gefitinib and erlotinib
R776H	0.24	ISEN	Increase Sensitivity gefitinib and erlotinib
C775Y	0.66	SRES	Unknown
H835Y	0.84	SRES	
G796V	0.85	SRES	

Table 2. Predicted aa-RFC phenotype of the ERK2-VTX11e resistant mutants reported in [78]. The top *likely-and-resistant* columns indicate their presence among the mutations included in the red area from Figure 5b and 5c.

Mutation	**NMR**	**Predicted class**	***Top Likely-and-resistant Melanoma?***	***Top likely-and-resistant Colorectal?***
P58L	0.60	SRES	YES	YES
P58S	0.70	SRES	YES	YES
P58T	0.67	RES	YES	YES
G37S	0.74	RES	YES	YES
Y64N	0.23	RES	NO	NO
Y36H	0.66	SRES	YES	NO
Y36N	0.58	SRES	NO	NO
C65Y	0.38	RES	NO	NO

FIGURES

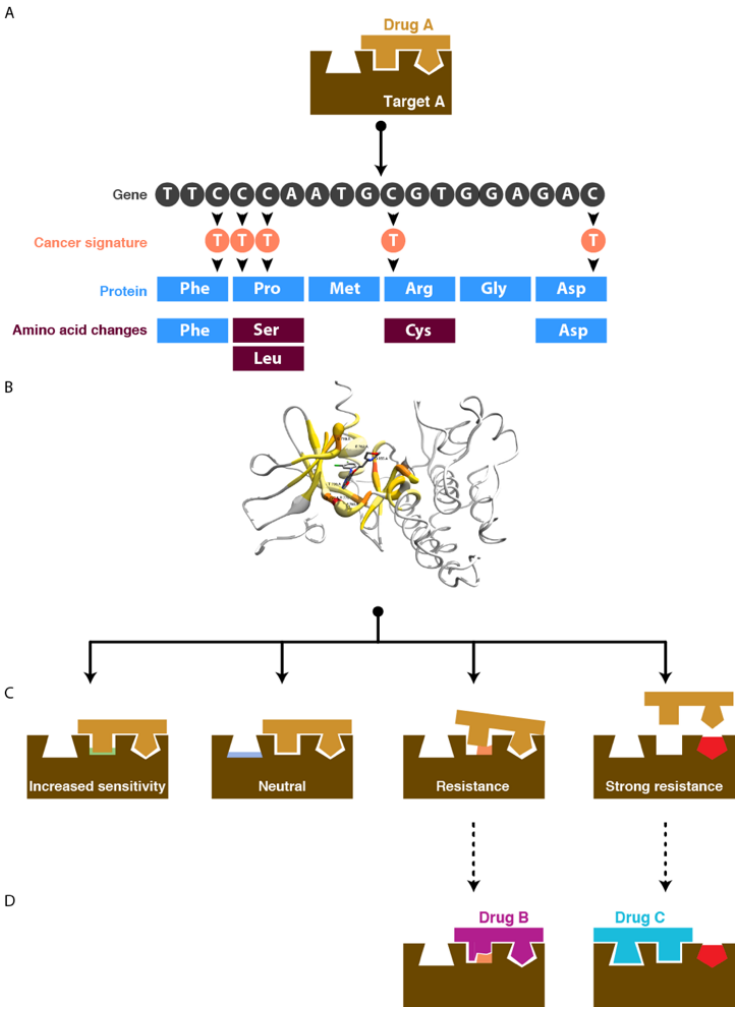


Fig. 1. Schematic representation of the developed framework. **a** For a particular targeted cancer therapy, the most likely mutations of the protein target are defined using the mutational signatures associated with that cancer class [35]. **b** 3D models of the mutations in the target structure are generated using the MODELLER package. **c** Structural and sequential information of the 3D-mutant models is used by a Random Forest Classifier (RFC) to predict the resistance potential of these mutations. **d** For the mutations classified as resistance-like, the model proposes alternative non-resistant compounds/drugs that may skip resistance.

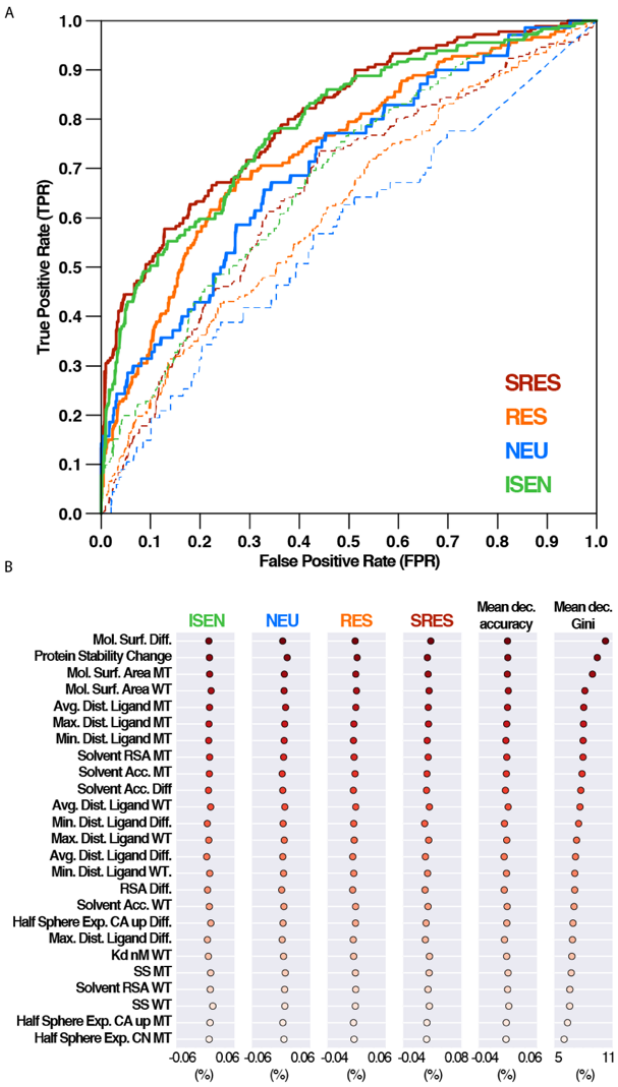


Fig. 2. RFC accuracy. **a** Receiver operating characteristic (ROC) curves of the four phenotypic classes (that is, strong resistance “SRES”, resistance “RES”, neutral “NEU” and increased sensitivity “ISEN”) after 10-fold cross validation. Solid lines correspond to the results of our RFC classifier; dashed lines correspond to the results of a non-trained approach based on the AutoDock Vina results. **b** Relative importance of the top 25 most informative variables used by the aa-RFC. Features are ranked by the mean decreased Gini score based on the Gini impurity index [58]. The rest of aa-RFC features are not shown for clarity.

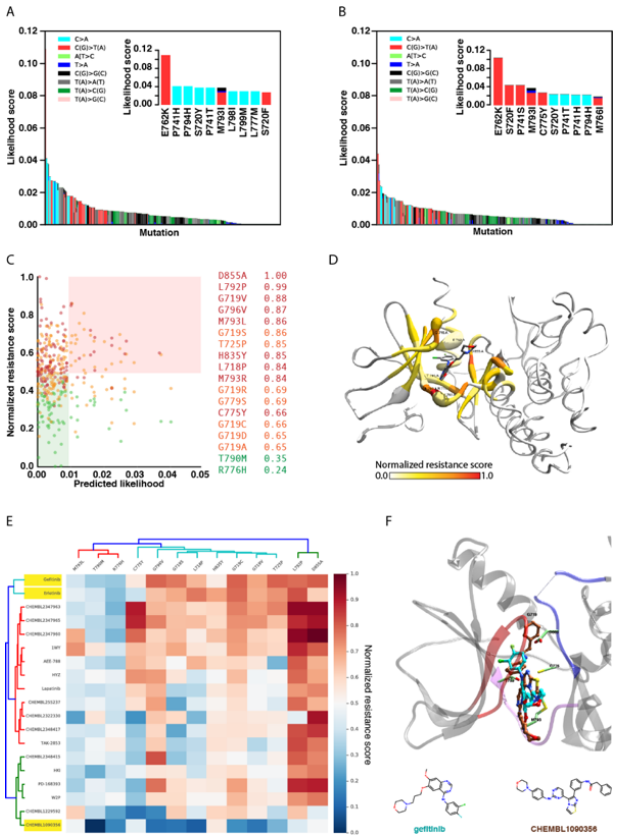


Fig. 3. a Predicted cancer associated-likelihood of mutations in the EGFR binding site for gefitinib for LUAD. Bar high indicates the likelihood of an amino acid mutation and its color the type of nucleotide change that leads to the amino acid mutation. When several nucleotide mutations lead to the same amino acid change, the probabilities were stacked. Inner sets show the top 10 likely mutations. **b** Predicted cancer associated-likelihood of mutations in the EGFR binding site for gefitinib for LSCC. Representation as in panel a. **c** Predicted likelihood and normalized resistance score (<NRS>) for EGFR mutations in the binding site of gefitinib for the LUAD cancer type. Each mutation is represented by a dot, which color indicates the predicted class by the aa-RFC (SRES in dark red, RES in orange, NEU in blue and ISEN in green). The red area encompasses mutations with predicted likelihood higher than the median value of all mutations and <NRS> higher than 0.5. The green area encompasses mutations predicted likelihood lower than the median value of all mutations and <NRS> lower than 0.5. The top ten resistance mutations, along with other mutations mentioned in the text are listed ordered by their <NRS>. **d** LUAD

mutation likelihood and normalized resistance score (<NRS>) in the 3D structure of EGFR-gefitinib complex (PDB: 4WKQ). The thickness of the ribbons indicates the accumulated mutational likelihood for that particular amino acid. The color represents the accumulated <NRS> score. Ligands are displayed as sticks. Mutations of amino acids beyond the binding site of the compounds were not considered. **e** Predicted sensitivity map for EGFR mutations in the binding site of gefitinib. Columns represents mutations, rows represent the screened compounds. The colour of the cells represents the predicted <NRS> by the lig-RFC. Name of the compounds are either the generic names for FDA approved drugs or drugs in clinical trials, the ChEMBL accession codes or the PDB accession code for those compounds lacking of an entry in ChEMBL. Compounds mentioned in the text are highlighted with a yellow background. **f** Structural mapping of the predicted resistance mutations in the wild type EGFR interaction with gefitinib (cyan) and CHEMBL1090356 (brown). PDB entries: 4WKQ and 3LZB for gefitinib and CHEMBL1090356 respectively. Side chains of the most important contributors to the binding are shown as sticks. The P-Loop is coloured in red, the hinge region in purple and A-Loop in blue.

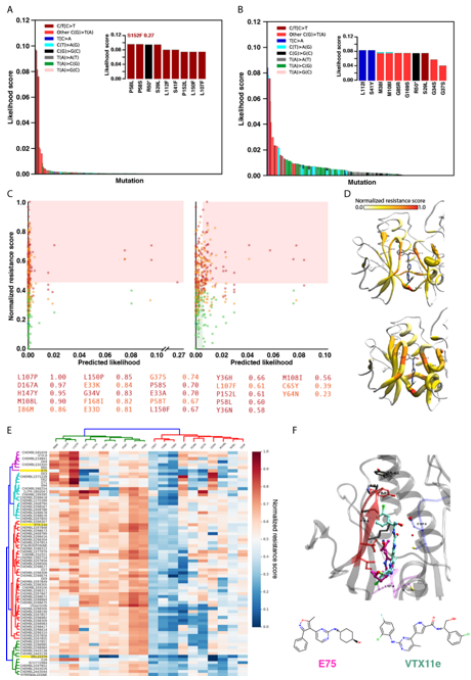


Fig. 4. a Predicted cancer associated-likelihood of mutations in the ERK2 binding site for VTX11e for melanoma. Represented as in Fig 3a. **b** Predicted cancer associated-likelihood of mutations in the ERK2 binding site for VTX11e for colorectal cancer. Representation as in panel a. **c** Predicted likelihood and normalized resistance score (<NRS>) for ERK2 mutations in the binding site of VTX11e for the melanoma (left) and colorectal (rigth) cancer types. Represented as in Fig 3c. **d** Melanoma (top) and colorectal (bottom) mutation likelihoods and normalized resistance scores (<NRS>) in the 3D structure of ERK2-VTX11e complex (PDB: 4QTE). Represented as in Fig 3d. **e** Predicted sensitivity map for ERK2mutations in the binding site of VTX11e. Represented as in Fig 3e. **f** Structural mapping of the predicted resistance mutations in the wild type ERK2interaction with VTX11e (cyan) and E75 (magenta). PDB entries: 4QTE and 4FUX for VTX11e and E75, respectively. Represented as in Fig 3f.

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. Particularly, Chapter 3.1 presented a comparative docking method (Subsection 1.1.9) 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 1.5 introduced a computational model that predicts cancer associated mutations with the highest chance 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 pre-

dictions 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 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 be analyzed in future studies. Specifically, 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 method 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 approaches. 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 (Subsection 1.5.2), 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 been 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 (Subsection 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 [244] to find the shortest pathways between compounds and protein targets. Other popular network search algorithms include random walk [245] or network propagation [246].

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. This application also exemplified how computational methods can play a significant role in the drug discovery process. Particularly, compounds were originally screened at the Tres Cantos Open Lab Foundation of GSK and were subsequently used as input of the ligand-target prediction methods developed by two academic research institutes. Finally, the open access profile of the conducted study gave raise to the delivery of the compounds and predictions. To the best of our knowledge, this is the first open access large-scale screening of antitubercular compounds, paving the way for future nonprofit R&D against TB. From a logistic perspective, this project also illustrated how Academia-Industry collaborations can improve the efficiency of R&D programs.
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 [247].

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 (Subsection 1.5) and bacterial infections (Subsection 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 in Subsection 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

This chapter presented a computational framework aiming to connect the mutational landscape of tumors with the drug-resistance phenotype generated by cancer-associated mutations in drug targets. Firstly, it introduced a computational model that predicts the probability of generation of spontaneous mutations in drug targets. The application of such model showed that the mutational profile of drug targets is cancer specific. Next, it introduced a Random Forest Classifier (RFC) that uses structural and sequential information of the drug target complex to predict the drug binding affinity change upon mutation. Application of the model to the EGFR-gefitinib and ERK2-VTX11e targeted cancer therapies identified some of the known resistant mutants alongside other new unseen mutations predicted to confer resistance to these therapies. Interestingly, the structural localization of the resistance mutations suggested that ATP plays an essential role in the spontaneous emergence of resistant mutants. To conserve the kinase activity, drug-resistant mutants need to conserve, or increase, the binding affinity of the ATP-analog substrate. Consequently, predicted drug-resistance mutations negatively interfering with ATP binding might be, in reality, non-functional.

Finally, for those mutations labeled as resistance-like, the model also predicted alternative non-resistant target inhibitors. Large-scale prediction of alternative small molecules sensitivity enabled the connection between the pharmacological space and the mutational landscape of EGFR and ERK2 in a cancer specific context.

4.3.2. Impact of the presented research

To our knowledge, this is the first machine learning model specifically applied to *de-novo* detect the resistance-associated phenotype caused by mutations in drug targets. The model does not require information about drug sensitivity across cell lines, patient-derived tumor xenografts or patient's genetic profile. Rather, the predictions are only based on sequential and structural features of the drug target interaction. Thus, it can be easily applied to any drug-target-mutation structural complex.

The study showed that the mutational landscape of drug targets varies across different cancer classes. Hence, the emergence of drug resistance to a given tar-

geted therapy is associated to the treating cancer class, at least, in those cases where the generation of spontaneous target mutations is the responsible mechanism of resistance. This result supports the conception of treating each cancer uniquely, pointing out the importance of studying the patient's genetic profile prior to apply any cancer treatment.

The framework can be applied to attain two main objectives: i) to anticipate which are the target mutations that may induce drug resistance to a given targeted cancer therapy and ii) to detect alternative non-resistant molecules for mutations conferring resistance to a targeted cancer therapy. On the one hand, the first objective helps to identify those mutations with adverse pharmacological profile given a particular targeted cancer therapy. The second objective, on the other hand, may guide the search for alternative molecules once the patient has already developed drug resistance as well as when the patient's genetic profile looks unfavorable for a particular targeted therapy.

In practical terms, we analyzed the pharmacological profile of more than 400 EGFR-mutations for gefitinib treatment and more than 400-ERK2 mutations for VTX11e therapy. In both cases, we also provided a compound-mutation sensitivity map of alternative small molecule inhibitors. All the analyzed data alongside the predictions will be freely available online once the manuscript is published.

Overall, this study illustrated how cancer therapies may benefit from the development of *in-silico* models assisting to the selection of the optimal treatment.

4.3.3. Limitations

This model focuses on identifying mutations decreasing the binding affinity of drugs acting on a protein target. This is one of the most frequently reported mechanism of drug resistance in targeted cancer therapy. Nevertheless, there are multiple alternative mechanisms of drug resistance not considered by the model [213]. Moreover, the model only works for single amino acid mutations. That is, it does not consider double, triple or multiple simultaneous mutants.

The likelihood model only considers the probability of generation of mutations associated with cancer. Therefore, it does not consider pre-malignant somatic mutations or germline variations. Such variations may also have an important impact in the emergence of drug resistance. Another limitation of the likelihood model is that its predictions are based on average probabilities from hundreds of patients samples. Therefore, the predicted likelihood shows global trends

but it is currently unable to capture specific characteristics of individual cancer cases. Finally, the likelihood predictions are static (*i.e.,* the model does not explicit consider any information about tumor size, clonal evolution or time). Including *dynamic* information about the tumor evolution would help to estimate the probability of emergence of mutations in a tumor more accurately.

The resistance model also suffers from other limitations. One of the major limitations is related to the structural nature of its predictions. It does require the 3D complex of the drug with the target to perform its predictions. In spite of most of the FDA-approved drugs have been co-crystallized with their main protein targets, there are several lacking of experimentally determined structure.
A distinct questionable aspect is the relatively small size of the training set. Moreover, the training set was originally unbalanced, which could induce a bias towards the most populated classes. This problem was soften by balancing the original training set.
Another aspect not explicitly considered by the model is the role of ATP in the emergence of resistance mutations in kinases. We proposed that ATP plays an important role by confining the number of mutations that both preserve the catalytic activity of the kinase and decrease the affinity of the kinase inhibitor. Therefore, those predicted resistance-like mutations adversely interfering with ATP may be in reality non-functional.

Finally, it is important to keep in mind that some external errors, which negatively affect the performance of the model (*e.g.,* modeling errors or virtual docking miscalculations), might have been introduced over the training and prediction processes.

4.3.4. Future perspectives

One of the short-term goals is to extend the application of the model to other targeted cancer therapies. The very likely increase in the number of experimentally determined 3D structures will expand the spectrum of targeted therapies amenable to be studied by the model. Furthermore, the growth would enlarge the number of candidate alternative compounds included in the search for non-resistant therapies. Finally, the model's training set, would also benefit from the inclusion of new instances, which would be eventually translated into a more accurate models. Therefore, we believe that the future can only increase the quality and the quantity of the predictions.

The likelihood model and the resistance predictor are also susceptible to be improved. As mentioned before, the likelihood predictions are average-based.

The average do not necessarily represent particular patient cases and therefore, the estimations may lose individual trends. However, including information about the patient's genetic profile, such as pre-existing somatic mutations, germline variations or the individual mutational signatures; would give a more personalized estimation of the mutational likelihood. The resistance predictor may also incorporate new features. For instance, the development a classifier allowing double or triple mutants. To do so, we would need a training set including such types of mutants, which at this moment is too narrow. Another interesting feature, which was discussed above, is the role of ATP (or ATP-analogues) in the generation of functional mutants in protein kinases. In the current model the role of each mutation for the binding of ATP is not explicitly considered. Including such feature would enable the prioritization of those mutations more likely to be functional.

Populations of tumor cells are very heterogeneous and may contain a large amount of somatic mutations. As a consequence, a target protein might harbor multiple distinct mutations in different subpopulations of tumor cells. Hence, in order to skip drug resistance, the anticancer drug needs to be simultaneously active against all the co-existing mutations. This an extremely difficult task, specially in highly-mutated tumors such as colorectal cancer or melanoma. I propose an alternative consisting of a rationally designed optimal combination of molecules able to overcome resistance to mutations present in a population of cells. Combining the predictions from the resistance model and information about tumor evolution, might lead to the development of non-resistant combinatorial targeted cancer therapies. Future work ought to focus on this idea too.

As discussed above in the other projects , experimental validation of the predictions would significantly enhance the value of the model. It would also enable the identification of both those cases were the model is more suitable and those cases where its performance decreases significantly.

Finally, the resistance model has been applied here to predict mutations likely to confer resistance to targeted cancer therapies. However, it might be applied to other systems and organisms. An example might be an application to detect mutations likely to confer resistance to antibiotics targeting bacterial proteins (such as those mentioned in Subsection 3.2).

5 Conclusions

- We developed nAnnolyze, a network-based comparative docking method that enables large scale and structurally detailed prediction of ligand target interactions.
- nAnnolyze was applied to predict the human targets for all FDA-approved small molecule drugs. The application identified some of the known molecular targets and also provided new potential drug-target interactions.
- The method alongside the predictions are available online in `http://nannolyze.cnag.cat`.
- nAnnolyze was also applied to predict the bacterial targets of two sets of antitubercular compounds. The predictions were combined with the predicted targets of two orthogonal approaches exploring different methodological spaces.
- Most of the predicted molecular targets were involved in amino acid metabolism pathways essential for the survival of *Mycobacterium tuberculosis*.
- We developed a model predicting the cancer-associated mutations likely to be responsible of resistance to a particular targeted cancer therapy.
- The model first estimates the likelihood of a mutation using the mutational signatures associated to the treating cancer class. Next, we used a RFC that leverages structural and sequential features of the ligand-target interaction to predict resistance-likeness of each target mutation.
- For those mutations classified as treatment-threatening, the model identified alternative non-resistant molecules predicted to overcome drug resistance.

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. 1).

[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 pp. 2, 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. 2).

[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. 2).

[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. 4).

[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. 4).

[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. 4, 10, 12, 149).

[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. 4, 149).

[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. 5).

[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. 5).

[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 p. 7).

[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. 7).

[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. 8).

[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. 8).

[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. 8).

[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. 9, 13).

[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. 10).

[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. 10).

[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. 10, 11).

[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. 10).

[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. 10).

[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. 10).

[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. 10).

[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. 11).

[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. 11).

[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. 11).

[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. 12).

[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. 12).

[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. 12).

[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. 12).

[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. 12).

[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. 12).

[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. 12).

[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. 12).

[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. 12).

[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. 12).

[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. 12).

[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. 12).

[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. 13).

[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. 13).

[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. 13).

[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. 13).

[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. 13).

[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. 13).

[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. 13).

[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. 13).

[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. 13).

[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. 13).

[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. 13).

[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. 13).

[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. 13).

[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. 13).

[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. 13).

[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. 13).

[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. 13).

[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. 13).

[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. 13).

[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. 13).

[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. 13).

[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. 13).

[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. 14).

[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. 14).

[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. 14).

[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. 14).

[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. 14).

[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. 14).

[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. 14).

[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. 14).

[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. 17).

[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. 17).

[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. 17).

[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. 17).

[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. 17).

[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. 17).

[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. 17).

[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. 17, 20).

[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. 17).

[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. 18).

[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. 18).

[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. 18, 44).

[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. 18).

[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. 18).

[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. 18, 44).

[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. 18).

[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. 18).

[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. 18).

[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. 20).

[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. 20).

[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. 20).

[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. 20).

[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. 20).

[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. 20).

[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. 20).

[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. 20).

[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. 20).

[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. 20).

[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. 20).

[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. 20, 46, 147).

[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. 20).

[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. 20).

[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. 20).

[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. 20).

[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. 20).

[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. 20).

[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. 20).

[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. 21).

[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. 21).

[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. 21).

[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. 21).

[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 pp. 21, 22).

[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. 21).

[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. 22).

[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. 23).

[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. 22).

[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. 25).

[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. 25).

[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. 25).

[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. 25).

[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. 25).

[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. 25, 31, 32).

[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. 25).

[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. 25).

[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. 25).

[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. 25, 32).

[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. 25).

[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. 26).

[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. 26).

[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. 26).

[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 p. 26).

[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. 26).

[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. 26).

[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. 26).

[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. 26).

[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. 26).

[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. 26).

[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. 26).

[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. 26).

[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. 27).

[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. 27).

[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. 27).

[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. 27, 28, 30).

[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. 27).

[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. 27).

[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 p. 28).

[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. 28).

[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. 28, 151).

[159] *Philanthropists unite to accelerate global fight against tuberculosis with combined $20 million gift to Broad Institute* (cit. on p. 28).

[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. 29).

[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. 29).

[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. 29).

[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. 29).

[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. 30).

[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. 30).

[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. 30).

[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. 30).

[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. 30).

[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. 30).

[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. 31).

[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. 31, 32).

[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. 31, 32).

[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. 31, 32).

[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. 31, 32).

[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. 31, 32).

[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. 31, 33).

[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. 31, 33).

[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. 31, 32).

[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. 31, 32).

[180] María Jose Rebollo-Lopez et al. «Release of 50 new, drug-like compounds and their computational target predictions for open source antitubercular drug discovery.» In: *PloS one* 10.12 (2015), e0142293. ISSN: 1932-6203. DOI: 10.1371/journal.pone.0142293 (cit. on p. 32).

[181] International Agency for Research on Cancer. *World Cancer Report 2014*. Tech. rep. 2014 (cit. on p. 33).

[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. 34, 35).

[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. 34).

[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. 34).

[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. 34).

[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. 34).

[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. 34).

[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. 34).

[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. 35).

[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 p. 35).

[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. 35).

[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. 35).

[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. 35).

[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. 35).

[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. 36, 43).

[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. 36).

[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. 36).

[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. 36, 41).

[199] Susanne Müller et al. «The ins and outs of selective kinase inhibitor development». In: *NATURE CHEMICAL BIOLOGY www.nature.com/naturechemicalbiology* 11.11 (2015), pp. 818–821. ISSN: 1552-4450. DOI: 10.1038/nchembio.1938 (cit. on pp. 36, 39).

[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. 36).

[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. 38).

[202] Kinase Inhibitor et al. «Exploration of Type II Binding Mode : A Privileged Approach for». In: (2016). DOI: 10.1021/cb500129t (cit. on p. 38).

[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. 38).

[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. 38).

[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. 39).

[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. 40).

[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. 40).

[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. 40).

[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. 40).

[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. 40).

[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. 41).

[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. 41).

[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. 42, 45, 154).

[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. 42).

[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. 42).

[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. 42).

[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. 42).

[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. 42).

[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. 42).

[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. 42, 43).

[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. 43).

[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. 43).

[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. 43).

[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. 43).

[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. 43).

[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. 43).

[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. 43).

[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. 43).

[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. 43).

[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. 43).

[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. 43).

[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. 43).

[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. 43).

[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. 44).

[235] Antoni Ribas and Jedd D. Wolchok. «Combining cancer immunotherapy and targeted therapy». In: *Current Opinion in Immunology* 25.2 (2013), pp. 291–296. ISSN: 09527915. DOI: 10.1016/j.coi.2013.02.011 (cit. on p. 44).

[236] 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. 44).

[237] 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. 44).

[238] 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. 44).

[239] 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. 44).

[240] 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. 44).

[241] 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. 44).

[242] 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. 44).

[243] 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. 44).

[244] 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. 150).

[245] 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. 150).

[246] 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. 150).

[247] 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. 151).