Mapping the 3D Space of Drug Resistance Variants

Inauguraldissertation
zur
Erlangung der Würde eines Doktors der Philosophie
vorgelegt der
Philosophisch-Naturwissenschaftlichen Fakultät
der Universität Basel

von

Erblin Asllanaj

2023

Genehmigt von der Philosophisch-Naturwissenschaftlichen Fakultät
auf Antrag von

Erstbetreuer:
Prof. Dr. Torsten Schwede

Zweitbetreuer:
Prof. Dr. Timm Maier

Externer Experte:
Prof Dr. Vincent Zoete

Basel, 21.02.2023

Prof. Dr. Marcel Mayor
Dekan

All happy families are alike; each unhappy family is unhappy in its own way.
- Leo Tolstoy

Erblin Asllanaj
Mapping the 3D space of drug resistance

PhD thesis, University of Basel, Basel, Switzerland (2023)
With references, with summary in English

CONTENTS

CHAPTER 1

Introduction

1.1 Variants

1.1.1 Variants: The Engine of Evolution

The concept of biological species arising from natural selection of beneficial traits began to spread among biologists at the beginning of the 18th century. This idea contributed to the creation of the field of evolutionary biology with the release of the book *On the Origin of Species* by Charles Darwin[1], where he stated: *"Nothing at first can appear more difficult to believe than that the more complex organs and instincts should have been perfected, not by means superior to, though analogous with, human reason, but by the accumulation of innumerable slight variations, each good for the individual possessor."*

Nowadays, it is known that the information on the growth, functioning, reconstruction and reproduction of all organisms and viruses is stored in the sequences of nucleic acids that constitute a genome. Mutations result from errors during replication, mitosis, meiosis and external damages to the DNA, which are repaired in such a way that the genetic sequence is altered. These mutations are the source of all genetic variation and provide the building blocks that drive the process of genetic change over multiple generations, leading to the emergence of new traits or characteristics of a species.

The importance of mutations cannot be understated. They play a vital role in driving evolution, cause the majority of cancers and enable our immune system to keep up with the evolution of pathogens. Understanding how mutations affect the traits of an organism is one of the major questions in the life sciences.

Mutations can be distinguished by their impact on the genetic code. Large-scale mutations, or chromosome abnormalities, are drastic alterations of the chromosomal structure which include deletions, duplications, inversions, insertions and translocations of chromosomal segments. These mutations are considered to be impactful on the affected cell as a whole due to the altering of large parts of the genetic code. Small-scale mutations, on the other hand, encompass local changes in the genetic sequence. It can include the insertion or deletion of several nucleotides in the DNA or the altering of nucleotides[2].

The location of the mutation in the DNA is essential to understand its

impact. An important distinction can be made between mutations which lie outside or inside of the protein-coding region. Mutations in non-coding, regulatory sequences such as promoters, silencers or enhancers, can alter the level of gene expression, but most likely will not affect the protein sequence itself. Mutations in the coding region could alter the protein product in such a way that its function is impacted. Due to the involvement of proteins in practically all biological processes of an organism and their participation in a complex network of molecular biological interactions, the slightest changes of the protein might alter the phenotype of an organism significantly[3].

Mutations in the coding region can be separated by their effect on the respective amino acid sequence. Frameshift mutations are caused by insertions or deletions of nucleotides in such a way that the reading frame of the protein shifts from its original frame, leading to drastic alterations of the protein product. The event in which a number of nucleotides evenly divisible by three is inserted or removed is called an "in-frame mutation" and will insert or remove multiple amino acids from the protein sequence. A point substitution changes a single nucleotide, which can either be synonymous if the amino acid encoded by the altered nucleotide triplet is the same as the original amino acid, or non-synonymous if the amino acid is different. A nonsense mutation refers to the replacement of an amino acid encoding nucleotide triplet coding with a triplet which gives rise to a stop codon. The presence of the new stop codon results in the production of a shortened protein that is likely non-functional. Finally, missense mutations refer to mutations that substitute an amino acid in the protein product with another amino acid[4].

The central questions in the research of mutations are (1) how are the function of genes and proteins altered by mutations, and (2) what those changes entail for the phenotype of the organism as a whole. Answering these questions requires the inspection of the event through the lens of multiple biological disciplines. This endeavour includes the investigation of the structural and functional changes of proteins upon mutations, the analysis of the impact of the changed proteins on the metabolism or system of protein interactions in the organism and the study of how these changes manifest in the phenotype of the organism in the end. This can bring us closer to strengthening our understanding of how and why species adapt and change.

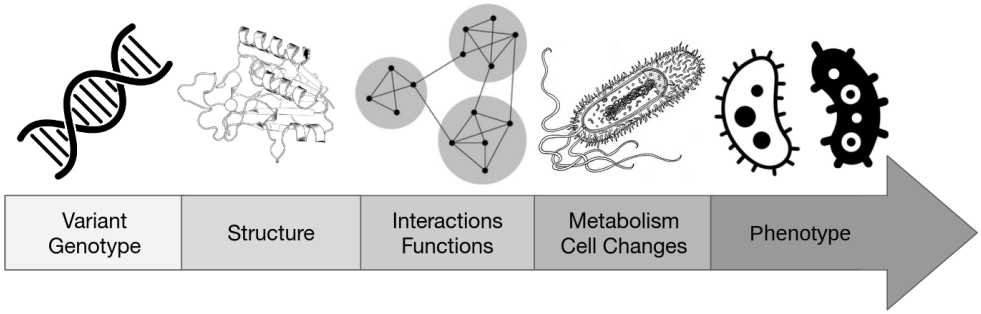

Figure 1.1: Schematic of the central scientific question of variant research: understanding the steps from genotype to phenotype.

Besides the academic interest in the investigation of mutations in the context of evolutionary biology, the topic also has practical relevance for medical research. The two main applications are research into disease-associated variants and the development of drug resistance.

The functioning of an organism relies on an intricate system of working interactions between thousands of proteins. Perturbations in this system, such as those caused by changes in the coding region of the DNA, can manifest themselves in a highly diverse set of diseases. Mutations in a germ cell can be passed to all cells of an organism. If such a mutation leads to harmful or damaging effects, it could result in an inherited disease. An example is albinism, which is caused by a mutation in the OCA1 or OCA2 gene and leads to the absence of melanin in the individual. This results in an increased risk of many types of cancers and impaired vision[5].

Somatic mutations that disrupt the balance between the proliferation of a cell and its apoptosis lead to uncontrolled cell division. This is considered to be the causative agent of most types of cancer. Due to the high mutation rate of cancer cells, the interpretation of variants plays a role not only in understanding how cancer emerges but also in how cancer cells adapt to evade the immune system and resist treatments using anticancer agents[6–8].

Changes in the DNA do not have to be harmful to an organism. Occasionally the genotypic effect is positive and allows a population to propagate and withstand environmental stresses better than organisms with the standard genotype (wild-type). If the organisms in question are

pathogens which are adapting in order to survive the medical drugs used to combat them, then we are facing a medical challenge[9].

Drug resistance is a major public health threat in the 21st century. The combination of the innate capacity of microbes to develop resistance at a rate that outpaces the development of new drugs leads to the spread of resistant pathogens for which treatment is more expensive, constitutes a higher burden for the patient and is more lethal than the susceptible pathogens. Understanding the underlying mutations in pathogens that cause the resistance phenotype is crucial to guide the development strategy of new drugs, facilitating the diagnosis of resistant pathogens and helping our understanding of resistance mechanisms.

1.1.2 Variants in the Context of Protein Structure Biology

The fundamental principle of protein structure biology postulates that the amino acid sequence of a protein determines its structure and biophysical properties[10]. The last part is not strictly true for protein chains which require the aid of chaperone proteins to assume their functional conformation[11], but a direct relationship between amino acid sequence and protein structure can be presumed. Protein structures are in general relatively robust to small alterations of the underlying amino acid sequence and are not expected to fundamentally change their structure and function upon the mutation of a few residues[12,13].

However, numerous examples do exist where even a single amino acid substitution leads to functional impairment, aggregation, conformational changes and unfolding of the whole protein[14–19]. The prediction of these impactful mutations is still a challenge due to computationally demanding methods with limited accuracy[20]. They could help us understand exactly how these single mutations lead to such a drastic structural change[21] and thus predict if and how specific conformational changes are induced by mutations.

The robustness of protein structures towards mutations and the difficulty in predicting large conformational changes caused by small mutations leads us as a first approximation for the purpose of structure-based variant interpretation to assume that the wild-type structure is not going to undergo drastic changes. Therefore, most structure-based analyses of mutations will rely on the wild-type conformation of the protein of in-

terest. Wild-type experimental structures are more readily available, and structure prediction of wild-type proteins has higher precision and quality. The impact of mutations in the context of the wild-type structure is then used to assess their potential to induce a phenotype change

For mutations that are expected to strongly disrupt the protein structure, structure-based variant interpretation cannot offer much information. Frameshifts or large deletions (or indels) will likely lead to a non-functional protein. The details on how the structure of the protein is impacted exactly are not considered to be necessary to arrive at the conclusion that the affected protein ceases to function. The structure of large-scale insertions into the protein sequence can be potentially predicted by the latest structure prediction methods, but the accuracy of the results requires further assessment[22].

The information protein structures can provide to aid the interpretation of small-scale mutations (which are usually understood to be single amino acid substitutions) is the comparison of the chemical environment of the mutated residue in the wild-type setting to the mutated setting. The wild-type environment of the mutation site can offer insights into important chemical interactions that the wild-type amino acid is a part of. The spatial pattern of hydrogen bonds, hydrophobic interactions, ionic bonds and disulfide bridges are the chemical building blocks which define the biophysical and functional properties of proteins. Hydrophobic interactions and interactions between opposite charge pairs between residues of two different protein chains are essential to make protein-protein interactions possible[23]. In some cases, a single mutation is enough to abolish the entire protein interaction[24] or to generate an entirely new one[25]. Certain amino acids are also essential in the formation of a ligand binding site, where a 3D pattern of chemical moieties allows a ligand to interact with the protein. Any change to that environment can weaken or even completely abolish the ligand interaction[26–28]. The same principle is also true for catalytic sites in enzymes, where a specific chemical reaction needs to be enabled by the same chemical 3D pattern next to the proper interaction with the target ligand[29–31]. A mutation that weakens or removes a participant in this system of chemical interactions can lead to a perturbation of the energy surface which abolishes the catalytic activity of the enzyme[32,33].

Inspecting the protein structure also reveals the available space in the immediate environment of a mutation site. Amino acids in a structured protein are usually packed together efficiently. Mutations of residues with a small volume to amino acids with a large volume will lead to steric clashes which disrupt the environment of the mutation site, independently of any potential effects on chemical interactions. A mutation towards a smaller amino acid can also cause issues for the stability of the protein by introducing voids in the core of the protein. The structure of the protein is required to accommodate for these disruptions, which can lead to the destabilisation of the affected protein region[34,35].

The 3D information of protein structures allows us to interpret the systems of chemical interactions and the available space of specific residue sites. Mutations either disrupt these interactions or introduce instabilities into the protein structure due to the loss of favourable interactions.

1.2 Antibiotic Drug Resistance And Protein Structure Biology

1.2.1 Overview of Antibiotic Drug Resistance

With the discovery of Penicillin by Sir Alexander Fleming in 1928[36] and its purification and testing as an antibacterial drug by Ernst Chain and Howard Florey in 1940[37], the antibiotic revolution began its march across the world. The compound proved its use right away in World War II, where it was mass-produced and used to treat infections in wounded soldiers. After 1945, penicillin was mass-produced for the public and antibiotics became a standard drug in the medical arsenal.

The treatment of infectious diseases was revolutionised worldwide, especially in Western countries. The introduction of antibiotics in the United States caused the shift of the leading cause of death to change from transmissible diseases to non-transmissible ones (for example, cardiovascular diseases, cancer, and stroke) and the average life expectancy to rise to 78.8 years[38]. They also enabled other medical fields in an unprecedented way. Complex surgeries such as organ transplants, joint replacements, or open-heart surgeries would be too risky without the ability to treat infections. Antibiotics have successfully prevented or

treated infections that can occur in patients who are receiving intensive treatments such as chemotherapy and who have chronic diseases such as diabetes. The time period after 1945 up to 1970 is known as the golden age of antibiotic discovery. Around 80% of the clinically relevant antibiotic classes were discovered in this time period[39].

Antibiotics also revolutionised global food production and enabled the spread of standardised animal monocultures across the world. In the example of poultry production in Brazil, a twenty-fold production increase between 1968 and 1990 was achieved and amounted to 12 million tons of poultry in 2009. In Europe and the United States, antibiotics are used as a standard agricultural treatment to not only treat disease but also to prevent their spread[40].

However, the efficacy of antibiotics did not last forever and antibiotic drug resistance started to become a major issue. Even before antibiotic drugs properly began their worldwide advance, occurrences of antibiotic resistance were recorded. Penicillin-resistant bacterial strains were already documented in 1942 in London when hospitalised patients infected with *Staphylococcus aureus* were found to resist treatment with penicillin[41]. Similarly, Streptomycin-resistant strains of *Mycobacterium tuberculosis* were detected only 5 years after the discovery of the antibiotic in 1943[42]. In 1945, Sir Alexander Fleming warned that the *"public will demand [the drug and] . . . then will begin an era . . . of abuses."*[43].

Resistance against multiple antibiotics in a single bacterial strain was first described a decade later in the late 1950s to the early 1960s[44–46] for *Escherichia coli*, *Shigella* and *Salmonella*. These strains were quite challenging to treat and often lead to deadly outcomes in poorer countries that do not have access to the latest antibiotics to circumvent the resistance. But at that time, antibiotic resistance was still perceived only as a medical curiosity and not a serious problem for the industrialised world. The attitude changed quickly in the 1970s with an outbreak of *Neisseria gonorrhoeae* with resistance to ampicillin[47] and *Haemophilus influenzae* with resistance to ampicillin, chloramphenicol and tetracycline[48–50].

The situation became more precarious when the discovery of new antibiotics stopped in the 90s. This so-called "Discovery Void" (Figure 1.2) is primarily attributed to the refocusing of the pharmaceutical industry to

more profitable therapeutic ventures[51]. Since the discovery of oxazolidinones in 1987 no new antibiotic family has been discovered. Some new antibiotic compounds were still developed and deployed by modifying existing and established antibiotic compounds like through the commercialisation of linezolid in 2003 and daptomycin in 2001, even though these compounds were already known as potential antibiotic compounds in the 1950s and 1980s respectively[52].

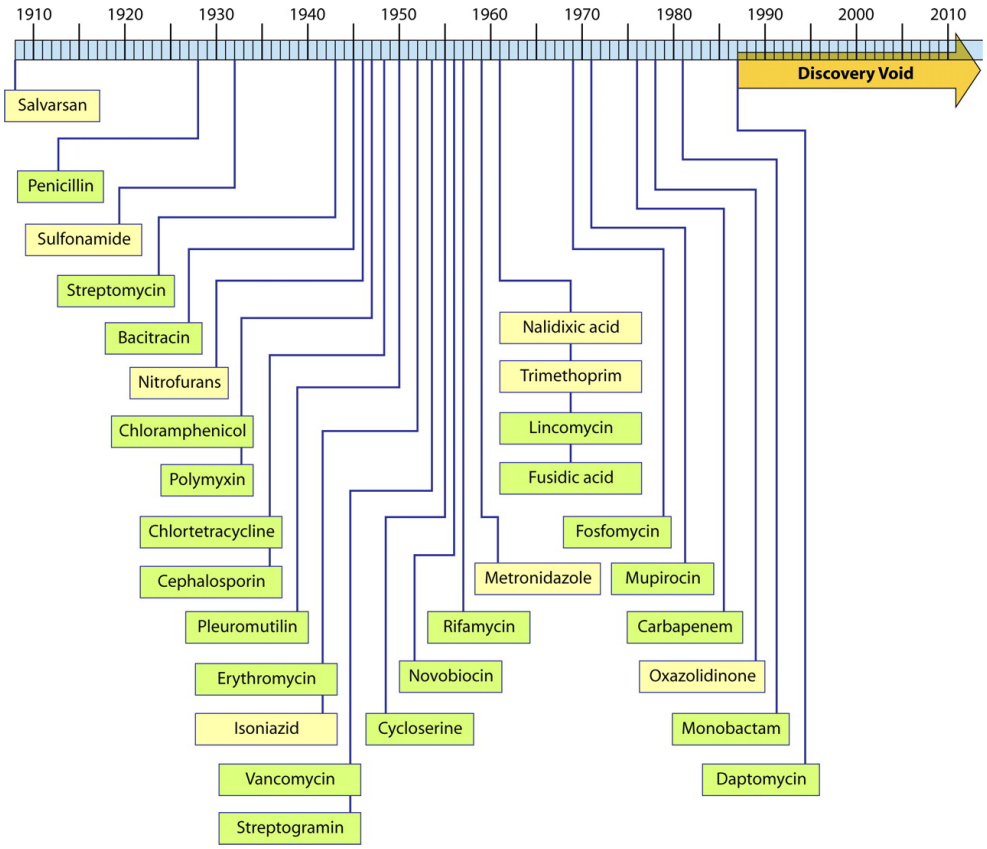

Figure 1.2: Time-line of the discovery of different antibiotic classes in clinical use. Figure 1 from[53]

Today, antibiotic resistance is considered one of the most challenging public health problems we are facing. Infectious diseases were the second most frequent cause of death right behind cardiovascular diseases before the start of the SARS-CoV-2 pandemic in 2020[54]. High rates of antibiotic resistance are encountered in bacterial infections such as sepsis, urinary tract infections and sexually transmitted diseases, which

is an indication that our antibiotic arsenal is running out. The rate of resistance to ciprofloxacin, which is used to treat urinary tract infections, varied from 8.4% to 92.9% in *Escherichia coli* and from 4.1% to 79.4% in *Klebsiella pneumonia* in countries that are reporting to the Global Antimicrobial Resistance and Use Surveillance System (GLASS). In 2019, 25 countries, territories and areas provided data to GLASS on bloodstream infections due to methicillin-resistant *Staphylococcus aureus* and 49 countries provided data on bloodstream infections due to *Echerichia coli*. The median rate observed for methicillin-resistant *S. aureus* was 12.11% and that for *E. coli* resistant to third-generation cephalosporins was 36.0%[55]. Antibiotic-resistant *Mycobacterium tuberculosis* (MTB) threatens the progress to containing the global tuberculosis epidemic. The World Health Organization (WHO) estimates half a million new rifampicin-resistant MTB cases globally in 2018, of which the majority are also multi-drug resistant MTB (MDR-TB). Less than 60% of those treated for MDR/RR-TB are successfully cured[54].

1.2.2 Improved Diagnostics is Essential for Combating Antibiotic Resistance

The development of new antibiotics is not the only way to thwart the resistance crisis. Quick and reliable diagnosis of the resistance status of infecting strains can provide crucial information to increase the efficiency of medical treatment. A medical professional with the knowledge of which antibiotics will work for the patient can drastically decrease the treatment time and increase treatment success.

Antibiotic resistance is classically determined by measuring the minimum inhibitory concentration (MIC). The MIC is the lowest concentration of an antibiotic necessary to kill a bacterial population. The resistance strength is proportional to the MIC increase. The MIC is usually measured in a broth dilution assay of a bacterial sample, where the concentration of an antibiotic is steadily increased across multiple fluid bacterial cultures with an incubation time of 17-20h. The lowest concentration that decreases the turbidity of the cultures to zero is the MIC[56]. While the measurement of the MIC is considered to be the gold standard of antibiotic resistance diagnosis, it is not always a practical option. The experimental setup requires a microbiology laboratory, which is not read-

ily available in less developed regions of the world. The time required to conduct the experiment can also be a bottleneck. While most bacterial species require 1 to 2 days until the culture is fully grown, some species like MTB require close to a month[57]. The bacterial sample growing successfully is also not a guarantee, which potentially means that experiments have to be repeated

With the next-generation sequencing revolution in the late 2000s, the large-scale analysis of bacterial genomes is possible and cost-effective. Specific genes and mutations were discovered to be characteristic of the resistance phenotype of certain antibiotic drugs. This can be used to infer the resistance phenotype by detecting these characteristic genetic elements instead of evaluating the ability of the bacterial cell to survive the presence of antibiotics in a laboratory setting[58].

This rapid and precise diagnosis of antibiotic-resistant strains provides the means to create a global surveillance system of the epidemiology of antibiotic-resistant strains and allows the deployment of local, national and global mitigation strategies.

The build-up of the infrastructure for the global surveillance of antibiotic resistance started in 2015 when the WHO created GLASS[59], which monitors the spread of antibiotic-resistant strains in 107 countries. GLASS provides a standardised methodology for genome analysis. Global surveillance is a game-changing tool for the monitoring of new outbreaks of resistance and for the evaluation of local, national and global containment and mitigation strategies.

Next to large-scale sequencing efforts, innovations in the diagnosis of resistance in a clinical setting were achieved through the development of the GeneXpert platform[60] for the diagnosis of rifampicin resistance in TB. Rifampicin resistance is known to be conferred by mutations in a specific region in the rpoB gene (Rifampicin Resistance Determining Region[61]). The GeneXpert platform can detect mutations in this region in the time span of 2 hours based on a sputum sample of the patient[62].

The technology however relies on knowing which variants are causative for resistance. While for certain antibiotic drugs that were under scientific investigation for decades, the genetic determinants are well characterised, the genetic resistance determinants for most antibiotics are not known.

A systematic and accessible workflow to increase the understanding of the mechanisms of resistance of a large number of resistance variants is necessary for clinicians worldwide. This framework could be used to decide the best course of action for the treatment of drug resistance by providing as much available information as possible.

1.2.3 Biological Basis of Antibiotic Resistance

Antibiotic resistance is defined as the ability of a bacterial organism to survive in an environment with a high concentration of antibiotic drugs[63]. Antibiotic drugs bind a specific type of biomolecule in the bacterial cell that is essential for its functioning. This interaction disables the target's function, and the consequence of this ceased or reduced function leads to the death of the bacterial cell (bactericidal antibiotics) or the prevention of further growth (bacteriostatic antibiotics).

Any phenotypic change that prevents this process from concluding in cell death or ceased cell growth is a resistance mechanism. There are multiple mechanisms for a bacterial cell to overcome its disruption by the antibiotic. They ultimately rely on the reduction of the effective concentration of the antibiotic compound in the cell to tolerable levels. It is important to stress that the complete abrogation of the drug effect is not necessary for antibiotic resistance to emerge. It is sufficient if the bacteria can tolerate the antibiotic at the given concentration, even if the bacterial cell is technically still susceptible to the drug. Antibiotic resistance is not a binary trait but can be seen as a spectrum of resilience dependent on antibiotic concentration[64]. This is reflected in the experimental assay for the determination of drug susceptibility, in which the MIC is measured[65]. It is important to keep in mind that the MIC is not necessarily a predictor of the clinical effectiveness of the antibiotic *in vivo*[66].

A mechanism of resistance is defined as the biochemical phenotype that counters the adverse effect of an antibiotic drug. These phenotypes can either be acquired through horizontal gene transfer (HGT) of resistance genes, through mutations in genes related to the mechanisms of action of antibiotics, i.e. encoding for the drug molecular target or other proteins and enzymes involved in drug transport and metabolism[67,68] (Figure 1.3).

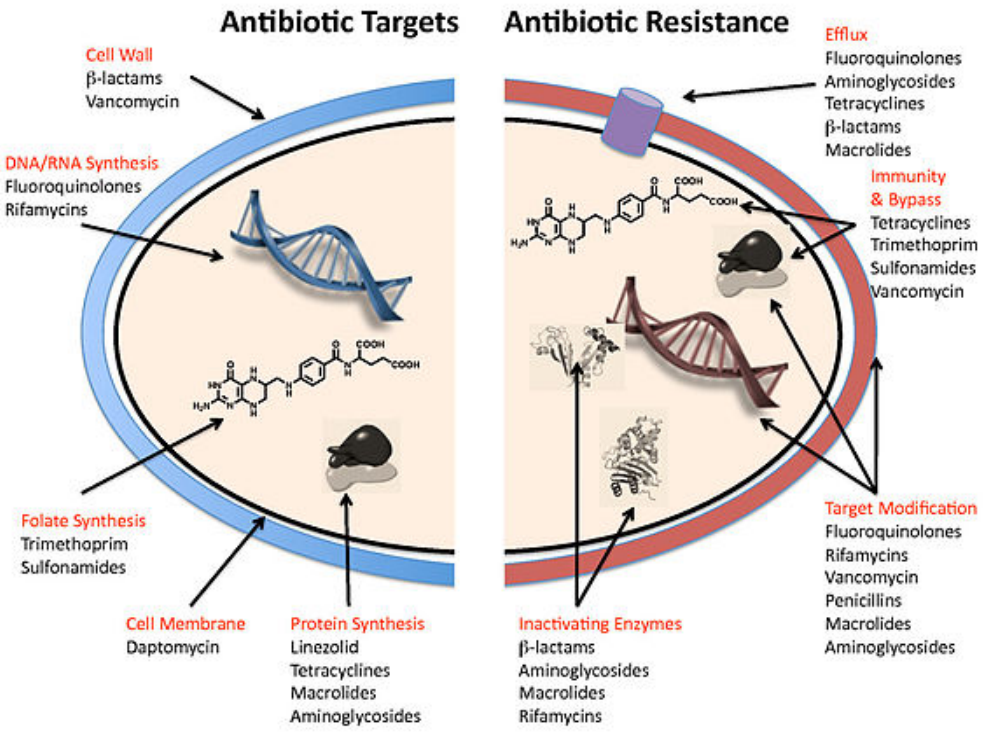

Figure 1.3: Overview over antibiotic drug targets and resistance mechanisms. Figure 1 from [69]

Drug inactivation occurs when catalysing enzymes specifically targeting antibiotic molecules are expressed in the bacterial cell. They are obtained through HGT of plasmids[70,71] and provide a direct route for the bacterial cell to decrease the concentration of the drug by breaking the drug molecule down into non-active compounds. Examples include the class of β-lactamases that break down the name-giving chemical moiety in β-lactames, the catalysis of tetracyclines by the monooxygenase tetX and different classes of aminoglycoside modifying enzymes which are encountered in drug-resistant gram-positive and gram-negative bacteria[72–74].

Drug-target alteration describes a biochemical change of the molecular target with which the antibiotic drug needs to interact to fulfil its func-

tion. The function of the drug target protein is significantly hindered by this interaction. Changes of the binding site that decrease the strength of the interaction between drug and target increase the tolerable antibiotic concentration in the cell[75]. This effect can also be achieved by a different but related resistance mechanism in which enzymes responsible for the post-translational modification of the drug binding site are affected. For example, resistance mutations in methyltransferases prevent methylations on the ribosome antibiotic binding sites which are necessary for the functioning of the drug[76].

Another drug resistance mechanism is the abrogation of prodrug activators. Some antibiotics are administered as pharmacologically inactive forms, which are activated by specific bacterial enzymes. It is observed in some cases that, while the drug target of the activated drug is essential for the survival of the bacterial cell, this is not always the case for the prodrug activator enzymes. In these cases, resistance can be achieved by deleterious mutations in these enzymes that interrupt the chain of reactions required for the prodrugs to transform into their activated form[77].

Other mechanisms provide ways to stop the antibiotic without hindering the binding of the drug to its designated target. Some mechanisms of resistance compensate for the antibiotic-disabled biological pathway by either activating an alternative metabolic pathway or by obtaining variations of the essential drug target through HGT. An example of this is the resistance of sulfonamides mediated through the HGT of alternative Dihydropteroate synthase plasmids[78]. Another example is the expression of tetracycline ribosomal protection proteins that are GTPases that catalyse the GTP-dependent release of tetracycline from the binding site, actively removing the antibiotic from the ribosome[79].

Decreased antibiotic concentration can also be achieved by altering the gene expression of certain proteins. Increased expression of active and passive transporter proteins and changes in the composition of the cell membrane can cause a general tolerance against antibiotics[80]. Mutations in promoter regions, leading to the overexpression of the drug target have been associated with resistance to drugs like Trimethoprim, Fluoroquinolones, β-lactames and sulfanilamides[81].

1.2.4 Tuberculosis as a Model Organism to Study Antibiotic Resistance Variants in their Structural Context

There are few illnesses which have accompanied humanity throughout history as consistently as tuberculosis. From the first report of the symptoms in Ancient Egypt to the description of the disease in the 19th century as the "white death" in art and literature, tuberculosis remained a constant companion of humanity. Tuberculosis is an infection of the lungs with the bacteria *Mycobacterium tuberculosis* (MTB). Although MTB is known to reproduce relatively slowly (roughly once per day), it is extremely resilient to the immune response of its host. Macrophages cannot digest the bacteria due to its thick mycolic acid protective layer. If the macrophage attempts to destroy the bacteria, it will reproduce in the macrophage and slowly eliminate it with secreted toxins. The infection attacks the lung tissue directly, causing small wart-like lesions (in Latin tubercles). Additionally, it is easily spread by air droplets from the host's mouth, which are produced even by speaking alone. Its spreadability, resilience towards the human immune system and devastating effect on the lung tissue of the host make tuberculosis a deadly disease[82].

Figure 1.4: Left: Picture of Mycobacterium tuberculosis. Scanning electron micrograph. Mag 15549X. Center for Disease Control US. Right: Postmortem examination of lungs of a 40 year old tuberculosis patient[83]

Today, tuberculosis is considered the second-deadliest infectious disease after COVID-19. In 2021, around 2.9 million people were confirmed with bacteriologically confirmed pulmonary TB. 71% of these patients were tested for rifampicin resistance. 150'000 multidrug-resistance MTB and 25'000 cases of extensively drug-resistant MTB were detected. With 10.6 million MTB cases occurring worldwide in the year 2021, it is assumed that many more cases of drug-resistant MTB remain unreported and, therefore, the data collected does not necessarily give the full picture of the situation[54].

Having a reliable and widespread way to diagnose drug resistance in patients allows health professionals to use treatments tailored to individual patients. Utilising the current antibiotic arsenal to its fullest extent can help to contain MDR-TB. The way drug resistance in MTB is diagnosed currently is by using a combination of bacteriological confirmation of MTB and testing for drug resistance using rapid molecular tests, culture methods or sequencing technologies. Given the slow growth rate, difficult growing conditions of MTB and the high expenses of a laboratory setting, culture methods alone are not considered to be a viable global testing strategy[84]. While characteristic resistance variants for commonly used antitubercular drugs like Rifampicin and Fluoroquinolones are well known and can be used to diagnose resistance with rapid molecular testing[85], many variants associated with resistance for other MTB antibiotics are only poorly understood or even not documented. Even the well-known resistance mechanisms of drugs like Rifampicin have many resistance-associated variants outside the binding site of the drug for whose variant impact is not characterised thoroughly.

1.3 Naturally Occurring Resistance of Antibody Therapeutics

1.3.1 Overview over Antibody Therapeutics

In the last four decades, the use of monoclonal antibodies changed drastically from a scientific tool to a powerful human therapeutic. The market for the pharmaceutical application of therapeutic antibodies is growing immensely: 50 antibody therapeutics were approved by the U.S. Food and Drug Administration (FDA) until 2015. The same number rose to

122 approved drugs in 2022[86]. The market share of antibody therapeutics in the world is estimated to be worth 186 billion USD in 2021 and based on ongoing clinical trials and preclinical studies the market is projected to grow to 445 billion USD by 2028[87].

Antibodies (Abs) or immunoglobulins are multimeric glycoproteins secreted by B-cells and formed by two identical light (L) and heavy (H) chains made by structurally similar domains, two for the light and four or more for the heavy chain (Figure 1.5). The antigen binding site is located on the upper tip of the molecule and is formed by the pairing of the VH and the VL variable domains, each contributing three hypervariable loops or complementary determining regions (CDR). The amino acid sequence, structure and length variability of the CDRs are the main determinant of the ability of antibodies to specifically recognise virtually any targets, called antigens.[88]

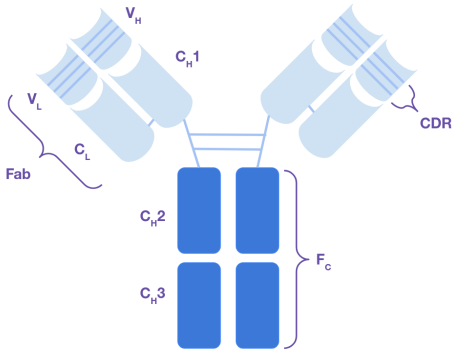

Figure 1.5: Schematic of the structure of an antibody. Immunoglobulin G (IgGs) are large proteins consisting of pairs of heavy and light chains connected through disulfide bridges. The heavy chain contains a variable domain VH, a hinge region and three constant domains (CH1, CH2 and CH3). The light chain contains a variable domain VL and only one constant domain CL. The antibody can be divided into a fragment antigen binding (Fab) region and a fragment crystallisation (Fc) region. The three hypervariable protein loops in each variable domain are also called the complementarity-determining region (CDR) and are responsible for the recognition of the respective antigen.

The immune response to an antigen is polyclonal, meaning that a diverse set of antibodies are produced that interact with different regions of the antigen (known as epitopes). Antibody therapeutics are monoclonal, referring to a homogenous population of antibodies to ensure a high specificity to a single epitope, low cross-reactivity and reproducibility of the drug product. The production of monoclonal antibodies (mAbs) was

described first in 1975 by Kohler and Milstein[89], where the authors succeeded in creating a fusion of myeloma cell lines with B cells to create hybridomas that could produce antibodies to a specific antigen. In the late 1980s, murine mAbs were tested in clinical trials. The therapeutics showed significant drawbacks, though. They caused strong allergic reactions together with the induction of anti-drug antibodies in the patient, which rendered the antibody therapeutic non-functional[90]. The first murine mAbs also had a short half-life in the patient and were relatively poor inducers of cytotoxicity when used for an oncological indication[91].

To overcome the drawbacks of murine mAbs, the antibody protein needed to be more similar to its human counterpart. The first attempts to solve this problem were through the creation of chimeric mouse-human antibodies. The antigen-specific variable domain of a mouse Ab was grafted onto a human Ab scaffold through genetic engineering techniques, resulting in mAbs that were around 65% human[92]. The half-life was increased and the allergic reactions were reduced with this class of mAbs, but the propensity to induce a specific immune response against the therapeutic was still significant[93]. The mAB properties were improved by an increased degree of humanization. Only the murine hypervariable region was grafted on a human Ab framework. The resulting molecule was 95% human[94]. The humanised mAbs resolved most of the issues in the clinical application of murine Abs, but the process does have limitations and can be a work-intensive process.

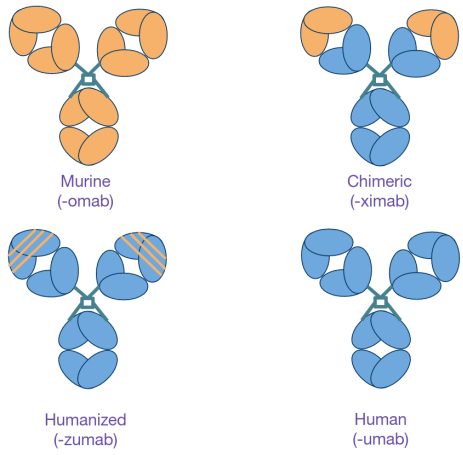

Figure 1.6: Overview over the four groups of monoclonal antibody therapeutics and their suffixes. The groups distinguish each other by the degree of humanisation.

Antibody drug discovery was revolutionised with the invention of the antibody phage display technology. Specific bacteriophages that display a protein of interest and contain its genetic information are screened against a library of human Abs. This allows the efficient identification of fully human candidate Abs for target proteins of interest[95]. Another way to generate fully human mAbs was by using transgenic mouse strains that express human variable domains[96–98]. The mAbs generated in transgenic mice or discovered through phage display show similar performance in the clinical setting. The compounds detected in a phage display screen however require more frequent additional lead optimization. This disadvantage is offset through direct lead isolation and control over the specificity and affinity of the mAb[99].

1.3.2 Naturally Occurring Polymorphisms Affect Epitope Recognition

In 1907, Paul Ehrlich observed during his foundational research on immunology that antibodies in the blood could attack invading pathogens specifically without causing any harm to the body[100]. Based on this observation he developed the scientific concept of a "magic bullet" [101], a molecule that only interacts with a specific target similar to the behaviour he observed for Abs.

The specificity of Abs arises from their highly variable complementary determining regions. The immune system has a wide arsenal of methods to increase the diversity of its Abs: somatic recombination[102], imprecise joining, random addition of nucleotides at the junction and somatic hypermutation[103,104] after the exposure to an antigen can be used to counter the diversity of pathogenic epitopes. Additionally, structural plasticity[105,106] adds to the process of generating even more antibody diversity. The antigen and antibody combining sites show a certain level of conformational flexibility to achieve complementarity at the antibody-antigen interface. This phenomenon was first described as "flexible keys and adjustable locks" by Edmundson et al.[107] when describing the binding of opioid peptides to an Mcg light chain dimer.

This arms race between pathogens and antibodies demonstrates that the immune system is required to regenerate new and flexible antibodies to combat the diversification of the surface epitopes of pathogens. An

antibody therapeutic, however, is a specific protein with a fixed amino acid sequence and will not undergo any change. Its efficacy is highly dependent on its respective antigen epitope to not undergo any change.

In 2014, a publication described the poor response of a subset of patients when treated with mAb Eculizumab[108]. This therapeutic agent is used to treat paroxysmal nocturnal hemoglobinuria (PNH), a rare, acquired and life-threatening haematological disease in which part of the innate immune system attacks the red blood cells of the patient. Eculizumab is a whole humanised antibody that targets the complement protein C5, a protein that is part of the cylindrical membrane attack complex that punctures the membrane of pathogenic cells, destroying them in the process. The binding of Eculizumab prevents the cleavage of the C5 protein into its two active products which activate an inflammatory response and build part of the attack complex[109].

The study assessed the sequences of the gene that encodes C5 in patients from Japan with PNH who either had a good response or a poor response to Eculizumab. Among the 345 patients with PNH treated with Eculizumab, 11 patients had a poor response. All 11 patients had a single heterozygous mutation, c.2654G→A, which translates into the single amino acid substitution p.Arg885His. It was found that the prevalence of this mutation in patients with PNH (3.2%) is reflecting the prevalence in the healthy Japanese population (3.5%).

This mutation did not arise due to any kind of evolutionary pressure but is present due to natural genetic diversity in a human population. Yet, the naturally occurring polymorphism can lower the efficacy of an Ab therapeutic and has real clinical consequences for the affected patients.

Structural biological investigations of the mutation in C5 could show that the p.Arg885His mutation is located in the centre of the antibody-antigen interface and that the arginine is interacting with an arginine binding pocket on the antibody interface. The histidine mutation is too short to fill this pocket, likely leading to a significant change in the conformation of the CDR3 loop of the opposing heavy chain, which leads to a loss of binding affinity[110] (Figure 1.7). Cases similar to the mutation in C5 are documented[111] but a systematic analysis of this phenomenon has to our knowledge not been conducted.

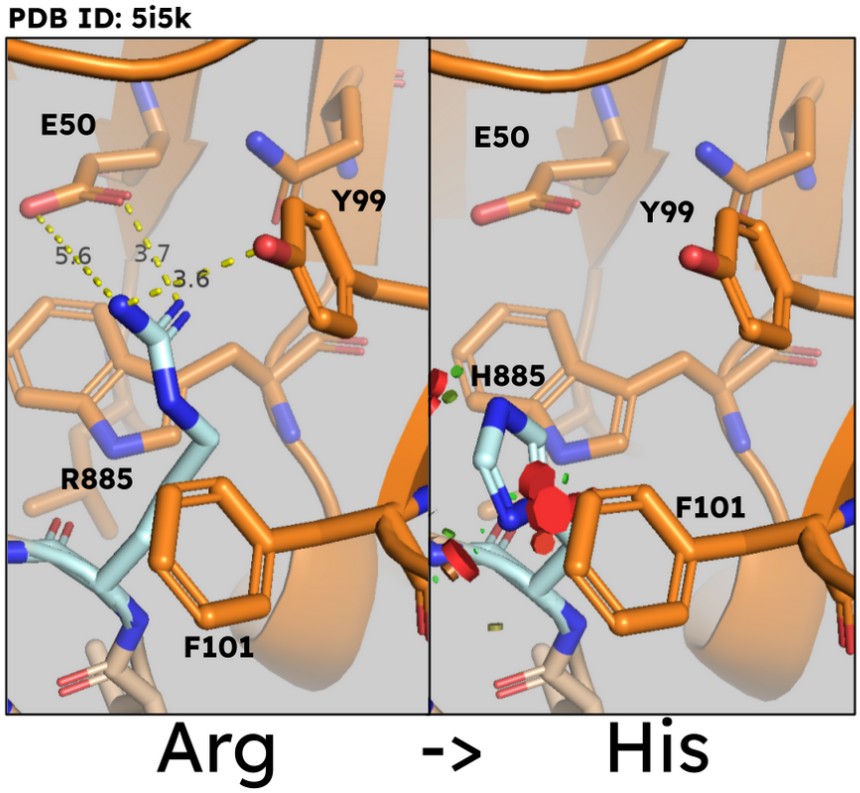

Figure 1.7: Variant p.Arg885His on complementary protein C5 in complex with eculizumab. The antigen is wheat coloured with the mutation site coloured in cyan. The antibody protein is coloured orange. The arginine residue (left panel) reaches deeply into the antibody protein having an interaction with an antibody glutamic acid and tyrosine. This interaction cannot be maintained if the residue is mutated to a histidine (right panel). Additionally, the most favourable rotamer still clashes with antibody phenylalanine. The visualisations were created with PyMol 1.30[112]

1.4 Thesis Objective

The main objective of this thesis is to advance the investigation of the impact of protein variants with a specific focus on their role in drug resistance mechanisms and epitope recognition.

The large set of variant and structure data necessitate software tools which facilitate their analysis. To this aim, I developed a new tool that allows the investigation of variants in their structural context. The variant annotation software, named Var3D, automatically aggregates and annotates variant and structural data. The results of these efforts are described in Chapter 2.

The further development of structure-based analysis tailored to resistance-associated variants in MTB can provide a framework for clinical researchers worldwide to form compelling hypotheses on their impact. Using Var3D and the resources of the protein structure modelling server SWISS-MODEL, I lead the development of TBvar3D, a web server for the analysis of protein variants of MTB in the context of protein structure information and antibiotic resistance variants. The tool allows the user to inspect their variants in the context of the "Catalogue of mutations in *Mycobacterium tuberculosis* complex and their association with drug resistance" [113] from the WHO, which provides a reference standard for the interpretation of mutations conferring resistance to all first-line and some second-line drugs. This server is the main topic of Chapter 3.

A systemic analysis of the impact of naturally occurring polymorphisms on epitope recognition has to our knowledge not been performed yet and could reveal variants which are critical for the clinical application of antibody therapeutics. Through the use of Var3D, I identified and annotated all known variants on therapeutic antibody-antigen interfaces. In collaboration with Pr. Dr. Lukas Jeker of the Department of Biomedicine at the University of Basel we aim to experimentally validate the impact of a subset of variants which were selected based on the calculated annotations. This subject is described in Chapter 4.

References

[1] Darwin, C. (2004). *On the origin of species, 1859*. Routledge.

[2] Alberts, B., Bray, D., Hopkin, K., Johnson, A. D., Lewis, J., Raff, M., Roberts, K., & Walter, P. (2015). *Essential cell biology*. Garland Science.

[3] Reece, J. B., Urry, L. A., Cain, M. L., Wasserman, S. A., Minorsky, P. V., Jackson, R. B. et al. (2014). *Campbell biology* volume 9. Pearson Boston.

[4] Lieberman, M., Marks, A. D., Smith, C. M., & Marks, D. B. (2007). *Marks' essential medical biochemistry*. Lippincott Williams & Wilkins.

[5] Summers, C. G. (2009). Albinism: classification, clinical characteristics, and recent findings. *Optometry and Vision Science*, *86*, 659–662.

[6] Hanahan, D., & Weinberg, R. A. (2000). The hallmarks of cancer. *Cell*, *100*, 57–70.

[7] Hanahan, D., & Weinberg, R. A. (2011). Hallmarks of cancer: the next generation. *Cell*, *144*, 646–674.

[8] Hanahan, D. (2022). Hallmarks of cancer: new dimensions. *Cancer discovery*, *12*, 31–46.

[9] Aslam, B., Wang, W., Arshad, M. I., Khurshid, M., Muzammil, S., Rasool, M. H., Nisar, M. A., Alvi, R. F., Aslam, M. A., Qamar, M. U. et al. (2018). Antibiotic resistance: a rundown of a global crisis. *Infection and drug resistance*, *11*, 1645.

[10] Anfinsen, C. B. (1973). Principles that govern the folding of protein chains. *Science*, *181*, 223–230.

[11] Ellis, R. J., & Van der Vies, S. M. (1991). Molecular chaperones. *Annual review of biochemistry*, *60*, 321–347.

[12] Chothia, C., & Lesk, A. M. (1986). The relation between the divergence of sequence and structure in proteins. *The EMBO journal*, *5*, 823–826.

[13] Kinch, L. N., & Grishin, N. V. (2002). Evolution of protein structures and functions. *Current opinion in structural biology*, *12*, 400–408.

[14] Narayan, A., & Naganathan, A. N. (2018). Switching protein conformational substates by protonation and mutation. *The Journal of Physical Chemistry B*, *122*, 11039–11047.

[15] Petrović, D., Risso, V. A., Kamerlin, S. C. L., & Sanchez-Ruiz, J. M. (2018). Conformational dynamics and enzyme evolution. *Journal of the Royal Society Interface*, *15*, 20180330.

[16] Nelson, G., Buzko, O., Spilman, P., Niazi, K., Rabizadeh, S., & Soon-Shiong, P. (2021). Molecular dynamic simulation reveals e484k mutation enhances spike rbd-ace2 affinity and the combination of e484k, k417n and n501y mutations (501y. v2 variant) induces conformational change greater than n501y mutant alone, potentially resulting in an escape mutant. *BioRxiv*, .

[17] Ie, S. I., Thedja, M. D., Roni, M., & Muljono, D. H. (2010). Prediction of conformational changes by single mutation in the hepatitis b virus surface antigen (hbsag) identified in hbsag-negative blood donors. *Virology journal*, *7*, 1–9.

[18] Hernández-Santoyo, A., del Pozo Yauner, L., Fuentes-Silva, D., Ortiz, E., Rudiño-Piñera, E., Sánchez-López, R., Horjales, E., Becerril, B., & Rodríguez-Romero, A. (2010). A single mutation at the sheet switch region results in conformational changes favoring λ6 light-chain fibrillogenesis. *Journal of molecular biology*, *396*, 280–292.

[19] Piel, F. B., Steinberg, M. H., & Rees, D. C. (2017). Sickle cell disease. *New England Journal of Medicine*, *376*, 1561–1573.

[20] Geng, C., Xue, L. C., Roel-Touris, J., & Bonvin, A. M. (2019). Finding the $\delta\delta$g spot: Are predictors of binding affinity changes upon mutations in protein–protein interactions ready for it? *Wiley Interdisciplinary Reviews: Computational Molecular Science*, *9*, e1410.

[21] Buel, G. R., & Walters, K. J. (2022). Can alphafold2 predict the impact of missense mutations on structure? *Nature Structural & Molecular Biology*, *29*, 1–2.

[22] Jilani, M., Turcan, A., Haspel, N., & Jagodzinski, F. (2022). Elucidating the structural impacts of protein indels. *Biomolecules*, *12*, 1435.

[23] Yan, C., Wu, F., Jernigan, R. L., Dobbs, D., & Honavar, V. (2008). Characterization of protein–protein interfaces. *The protein journal*, *27*, 59–70.

[24] Doud, M. B., Lee, J. M., & Bloom, J. D. (2018). How single mutations affect viral escape from broad and narrow antibodies to h1 influenza hemagglutinin. *Nature communications*, *9*, 1–12.

[25] Crepeau, R. H., Edelstein, S. J., Szalay, M., Benesch, R. E., Benesch, R., Kwong, S., & Edalji, R. (1981). Sickle cell hemoglobin fiber structure altered by alpha-chain mutation. *Proceedings of the National Academy of Sciences*, *78*, 1406–1410.

[26] Honegger, A., Dull, T., Felder, S., Van Obberghen, E., Bellot, F., Szapary, D., Schmidt, A., Ullrich, A., & Schlessinger, J. (1987). Point mutation at the atp binding site of egf receptor abolishes protein-tyrosine kinase activity and alters cellular routing. *Cell*, *51*, 199–209.

[27] Loftus, J. C., O'Toole, T. E., Plow, E. F., Glass, A., Frelinger III, A. L., & Ginsberg, M. H. (1990). A β3 integrin mutation abolishes ligand binding and alters divalent cation-dependent conformation. *Science*, *249*, 915–918.

[28] Luo, R., Jin, Z., Deng, Y., Strokes, N., & Piao, X. (2012). Disease-associated mutations prevent gpr56-collagen iii interaction. *PloS one*, *7*, e29818.

[29] Dodson, G., & Wlodawer, A. (1998). Catalytic triads and their relatives. *Trends in biochemical sciences*, *23*, 347–352.

[30] Bartlett, G. J., Porter, C. T., Borkakoti, N., & Thornton, J. M. (2002). Analysis of catalytic residues in enzyme active sites. *Journal of molecular biology*, *324*, 105–121.

[31] Agarwal, P. K. (2018). A biophysical perspective on enzyme catalysis. *Biochemistry*, *58*, 438–449.

[32] Redmond, T. M., Poliakov, E., Yu, S., Tsai, J.-Y., Lu, Z., & Gentleman, S. (2005). Mutation of key residues of rpe65 abolishes its enzymatic role as isomerohydrolase in the visual cycle. *Proceedings of the National Academy of Sciences*, *102*, 13658–13663.

[33] Toscano, M. D., Woycechowsky, K. J., & Hilvert, D. (2007). Minimalist active-site redesign: teaching old enzymes new tricks. *Angewandte Chemie International Edition*, *46*, 3212–3236.

[34] David, A., Razali, R., Wass, M. N., & Sternberg, M. J. (2012). Protein–protein interaction sites are hot spots for disease-associated nonsynonymous snps. *Human mutation*, *33*, 359–363.

[35] Cuff, A. L., & Martin, A. C. (2004). Analysis of void volumes in proteins and application to stability of the p53 tumour suppressor protein. *Journal of molecular biology*, *344*, 1199–1209.

[36] Fleming, A. (1929). On the antibacterial action of cultures of a penicillium, with special reference to their use in the isolation of b. influenzae. *British journal of experimental pathology*, *10*, 226.

[37] Chain, E., Florey, H. W., Gardner, A. D., Heatley, N. G., Jennings, M. A., Orr-Ewing, J., & Sanders, A. G. (1940). Penicillin as a chemotherapeutic agent. *The lancet*, *236*, 226–228.

[38] CDC National Centre for Health Statistics. Life expectancy. URL: `https://www.cdc.gov/nchs/fastats/life-expectancy.htm` Date Accessed: 2022-07-19.

[39] Durand, G. A., Raoult, D., & Dubourg, G. (2019). Antibiotic discovery: history, methods and perspectives. *International journal of antimicrobial agents*, *53*, 371–382.

[40] Kirchhelle, C. (2020). *Pyrrhic progress*. Critical Issues in Health and Medicine Series. New Brunswick, NJ: Rutgers University Press.

[41] Rammelkamp, C. H., & Maxon, T. (1942). Resistance of staphylococcus aureus to the action of penicillin. *Proceedings of the Society for Experimental Biology and Medicine*, *51*, 386–389.

[42] Crofton, J., & Mitchison, D. (1948). Streptomycin resistance in pulmonary tuberculosis. *British medical journal*, *2*, 1009.

[43] Spellberg, B., & Gilbert, D. N. (2014). The future of antibiotics and resistance: a tribute to a career of leadership by john bartlett. *Clinical infectious diseases*, *59*, S71–S75.

[44] Watanabe, T. (1963). Infective heredity of multiple drug resistance in bacteria. *Bacteriological reviews*, *27*, 87–115.

[45] Olarte, J. (1983). Antibiotic resistance in mexico. *APUA Newslett*, *1*.

[46] Levy, S. B. (2001). Antibiotic resistance: consequences of inaction. *Clinical Infectious Diseases*, *33*, S124–S129.

[47] Elwell, L. P., Roberts, M., Mayer, L. W., & Falkow, S. (1977). Plasmid-mediated beta-lactamase production in neisseria gonorrhoeae. *Antimicrobial agents and chemotherapy*, *11*, 528–533.

[48] De Graaff, J., Elwell, L. P., & Falkow, S. (1976). Molecular nature of two beta-lactamase-specifying plasmids isolated from haemophilus influenzae type b. *Journal of Bacteriology*, *126*, 439–446.

[49] Marshall, B., Roberts, M., Smith, A., & Levy, S. (1984). Homogeneity of transferable tetracycline-resistance determinants in haemophilus species. *Journal of Infectious Diseases*, *149*, 1028–1029.

[50] Van Klingeren, B., Van Embden, J., & Dessens-Kroon, M. (1977). Plasmid-mediated chloramphenicol resistance in haemophilus influenzae. *Antimicrobial Agents and Chemotherapy*, *11*, 383–387.

[51] Renwick, M., & Mossialos, E. (2018). What are the economic barriers of antibiotic r&d and how can we overcome them? *Expert opinion on drug discovery*, *13*, 889–892.

[52] Lewis, K. (2013). Platforms for antibiotic discovery. *Nature reviews Drug discovery*, *12*, 371–387.

[53] Silver, L. L. (2011). Challenges of antibacterial discovery. *Clinical microbiology reviews*, *24*, 71–109.

[54] World Health Organization (2022). Global tuberculosis report 2022. *World Health Organization*, .

[55] World Health Organization. Global antimicrobial resistance and use surveillance system (GLASS) report: 2021.

[56] Wheat, P. F. (2001). History and development of antimicrobial susceptibility testing methodology. *Journal of Antimicrobial Chemotherapy*, *48*, 1–4.

[57] Schön, T., Miotto, P., Köser, C. U., Viveiros, M., Böttger, E., & Cambau, E. (2017). Mycobacterium tuberculosis drug-resistance testing: challenges, recent developments and perspectives. *Clinical Microbiology and Infection*, *23*, 154–160.

[58] Weyer, K., Mirzayev, F., Migliori, G. B., Van Gemert, W., D'Ambrosio, L., Zignol, M., Floyd, K., Centis, R., Cirillo, D. M., Tortoli, E. et al. (2013). Rapid molecular tb diagnosis: evidence, policy making and global implementation of xpert mtb/rif. *European Respiratory Journal*, *42*, 252–271.

[59] World Health Organization (2015). *Global antimicrobial resistance surveillance system: manual for early implementation*. World Health Organization.

[60] Piatek, A. S., Van Cleeff, M., Alexander, H., Coggin, W. L., Rehr, M., Van Kampen, S., Shinnick, T. M., & Mukadi, Y. (2013). Genexpert for tb diagnosis: planned and purposeful implementation. *Global Health: Science and Practice*, *1*, 18–23.

[61] Zaw, M. T., Emran, N. A., & Lin, Z. (2018). Mutations inside rifampicin-resistance determining region of rpob gene associated with rifampicin-resistance in mycobacterium tuberculosis. *Journal of infection and public health*, *11*, 605–610.

[62] Helb, D., Jones, M., Story, E., Boehme, C., Wallace, E., Ho, K., Kop, J., Owens, M. R., Rodgers, R., Banada, P. et al. (2010). Rapid detection of mycobacterium tuberculosis and rifampin resistance by use of on-demand, near-patient technology. *Journal of clinical microbiology*, *48*, 229–237.

[63] Levin-Reisman, I., Brauner, A., Ronin, I., & Balaban, N. Q. (2019). Epistasis between antibiotic tolerance, persistence, and resistance mutations. *Proceedings of the National Academy of Sciences*, *116*, 14734–14739.

[64] Vega, N. M., & Gore, J. (2014). Collective antibiotic resistance: mechanisms and implications. *Current opinion in microbiology*, *21*, 28–34.

[65] Wiegand, I., Hilpert, K., & Hancock, R. E. (2008). Agar and broth dilution methods to determine the minimal inhibitory concentration (mic) of antimicrobial substances. *Nature protocols*, *3*, 163–175.

[66] Mattie, H. (2000). Antibiotic efficacy in vivo predicted by in vitro activity. *International journal of antimicrobial agents*, *14*, 91–98.

[67] Olivares, J., Bernardini, A., Garcia-Leon, G., Corona, F., B. Sanchez, M., & Martinez, J. L. (2013). The intrinsic resistome of bacterial pathogens. *Frontiers in microbiology*, *4*, 103.

[68] Reygaert, W. C. (2018). An overview of the antimicrobial resistance mechanisms of bacteria. *AIMS microbiology*, *4*, 482.

[69] Munita, J. M., & Arias, C. A. (2016). Mechanisms of antibiotic resistance. *Microbiology spectrum*, *4*, 4–2.

[70] Bennett, P. (2008). Plasmid encoded antibiotic resistance: acquisition and transfer of antibiotic resistance genes in bacteria. *British journal of pharmacology*, *153*, S347–S357.

[71] Taylor, D. E., Gibreel, A., Lawley, T. D., & Tracz, D. M. (2004). Antibiotic resistance plasmids. *Plasmid biology*, (pp. 473–491).

[72] Bush, K., & Bradford, P. A. (2019). Interplay between β-lactamases and new β-lactamase inhibitors. *Nature Reviews Microbiology*, *17*, 295–306.

[73] Markley, J. L., & Wencewicz, T. A. (2018). Tetracycline-inactivating enzymes. *Frontiers in microbiology*, *9*, 1058.

[74] Ramirez, M. S., & Tolmasky, M. E. (2010). Aminoglycoside modifying enzymes. *Drug resistance updates*, *13*, 151–171.

[75] Ndagi, U., Falaki, A. A., Abdullahi, M., Lawal, M. M., & Soliman, M. E. (2020). Antibiotic resistance: bioinformatics-based understanding as a functional strategy for drug design. *Rsc Advances*, *10*, 18451–18468.

[76] Osterman, I., Dontsova, O., & Sergiev, P. (2020). rrna methylation and antibiotic resistance. *Biochemistry (Moscow)*, *85*, 1335–1349.

[77] Nguyen, L. (2016). Antibiotic resistance mechanisms in m. tuberculosis: an update. *Archives of toxicology*, *90*, 1585–1604.

[78] Sköld, O. (2000). Sulfonamide resistance: mechanisms and trends. *Drug resistance updates*, *3*, 155–160.

[79] Connell, S. R., Tracz, D. M., Nierhaus, K. H., & Taylor, D. E. (2003). Ribosomal protection proteins and their mechanism of tetracycline resistance. *Antimicrobial agents and chemotherapy*, *47*, 3675–3681.

[80] Wright, G. D. (2010). Q&a: Antibiotic resistance: where does it come from and what can we do about it? *BMC biology*, *8*, 1–6.

[81] Palmer, A. C., & Kishony, R. (2014). Opposing effects of target overexpression reveal drug mechanisms. *Nature communications*, *5*, 1–8.

[82] Lawn, S. D., & Zumla, A. I. (2011). Tuberculosis. *The Lancet*, *378*, 57–72. URL: `https://doi.org/10.1016/s0140-6736(10)62173-3`.

[83] "Wikimedia Commons". "Cavitary tuberculosis". URL: `https://upload.wikimedia.org/wikipedia/commons/1/18/Cavitary_tuberculosis.jpg` by Yale Rosen.

[84] Stevens, W. S., Scott, L., Noble, L., Gous, N., & Dheda, K. (2017). Impact of the genexpert mtb/rif technology on tuberculosis control. *Microbiology spectrum*, *5*, 5–1.

[85] Brown, S., Leavy, J. E., & Jancey, J. (2021). Implementation of genexpert for tb testing in low-and middle-income countries: A systematic review. *Global Health: Science and Practice*, *9*, 698–710.

[86] Kaplon, H., Crescioli, S., Chenoweth, A., Visweswaraiah, J., & Reichert, J. M. (2023). Antibodies to watch in 2023. In *Mabs* (p. 2153410). Taylor & Francis volume 15.

[87] Lyu, X., Zhao, Q., Hui, J., Wang, T., Lin, M., Wang, K., Zhang, J., Shentu, J., Dalby, P. A., Zhang, H. et al. (2022). The global landscape of approved antibody therapies. *Antibody Therapeutics*, *5*, 233–257.

[88] Buss, N. A., Henderson, S. J., McFarlane, M., Shenton, J. M., & De Haan, L. (2012). Monoclonal antibody therapeutics: history and future. *Current opinion in pharmacology*, *12*, 615–622.

[89] Köhler, G., & Milstein, C. (1975). Continuous cultures of fused cells secreting antibody of predefined specificity. *nature*, *256*, 495–497.

[90] Ober, R. J., Radu, C. G., Ghetie, V., & Ward, E. S. (2001). Differences in promiscuity for antibody–fcrn interactions across species: implications for therapeutic antibodies. *International immunology*, *13*, 1551–1559.

[91] Stern, M., & Herrmann, R. (2005). Overview of monoclonal antibodies in cancer therapy: present and promise. *Critical reviews in oncology/hematology*, *54*, 11–29.

[92] Morrison, S. L., Johnson, M. J., Herzenberg, L. A., & Oi, V. T. (1984). Chimeric human antibody molecules: mouse antigen-binding domains with human constant region domains. *Proceedings of the National Academy of Sciences*, *81*, 6851–6855.

[93] Jones, P. T., Dear, P. H., Foote, J., Neuberger, M. S., & Winter, G. (1986). Replacing the complementarity-determining regions in a human antibody with those from a mouse. *Nature*, *321*, 522–525.

[94] Presta, L. G. (2006). Engineering of therapeutic antibodies to minimize immunogenicity and optimize function. *Advanced drug delivery reviews*, *58*, 640–656.

[95] McCafferty, J., Griffiths, A. D., Winter, G., & Chiswell, D. J. (1990). Phage antibodies: filamentous phage displaying antibody variable domains. *nature*, *348*, 552–554.

[96] Boulianne, G. L., Hozumi, N., & Shulman, M. J. (1984). Production of functional chimaeric mouse/human antibody. *Nature*, *312*, 643–646.

[97] Lonberg, N., Taylor, L. D., Harding, F. A., Trounstine, M., Higgins, K. M., Schramm, S. R., Kuo, C.-C., Mashayekh, R., Wymore, K., McCabe, J. G. et al. (1994). Antigen-specific human antibodies from mice comprising four distinct genetic modifications. *Nature*, *368*, 856–859.

[98] Green, L. L., Hardy, M., Maynard-Currie, C. a., Tsuda, H., Louie, D., Mendez, M., Abderrahim, H., Noguchi, M., Smith, D., Zeng, Y. et al. (1994). Antigen–specific human monoclonal antibodies from mice engineered with human ig heavy and light chain yacs. *Nature genetics*, *7*, 13–21.

[99] Lonberg, N. (2008). Fully human antibodies from transgenic mouse and phage display platforms. *Current opinion in immunology*, *20*, 450–459.

[100] Silverstein, A. M. (1996). History of immunology: Paul ehrlich: The founding of pediatric immunology. *Cellular immunology*, *174*, 1–6.

[101] Ehrlich, P. (1960). Experimental researches on specific therapy: On immunity with special reference to the relationship between distribution and action of antigens: First harben lecture. In *The Collected Papers of Paul Ehrlich* (pp. 106–117). Elsevier.

[102] Rudikoff, S., Pawlita, M., Pumphrey, J., & Heller, M. (1984). Somatic diversification of immunoglobulins. *Proceedings of the National Academy of Sciences*, *81*, 2162–2166.

[103] Janeway, C. A., Travers, P., Walport, M., & Shlomchik, M. (2017). Immunobiology: the immune system in health and disease. 2005. *New York: Garland Science*, *6*.

[104] Tonegawa, S. (1983). Somatic generation of antibody diversity. *Nature*, *302*, 575–581.

[105] Khan, T., & Salunke, D. M. (2014). Adjustable locks and flexible keys: plasticity of epitope–paratope interactions in germline antibodies. *The Journal of Immunology*, *192*, 5398–5405.

[106] Bhowmick, A., & Salunke, D. M. (2013). Limited conformational flexibility in the paratope may be responsible for degenerate specificity of hiv epitope recognition. *International immunology*, *25*, 77–90.

[107] Edmundson, A. B., Ely, K. R., Herron, J. N., & Cheson, B. D. (1987). The binding of opioid peptides to the mcg light chain dimer: flexible keys and adjustable locks. *Molecular Immunology*, *24*, 915–935.

[108] Nishimura, J.-i., Yamamoto, M., Hayashi, S., Ohyashiki, K., Ando, K., Brodsky, A. L., Noji, H., Kitamura, K., Eto, T., Takahashi, T. et al. (2014). Genetic variants in c5 and poor response to eculizumab. *New England Journal of Medicine*, *370*, 632–639.

[109] Dmytrijuk, A., Robie-Suh, K., Cohen, M. H., Rieves, D., Weiss, K., & Pazdur, R. (2008). Fda report: eculizumab (soliris®) for the treatment of patients with paroxysmal nocturnal hemoglobinuria. *The oncologist*, *13*, 993–1000.

[110] Schatz-Jakobsen, J. A., Zhang, Y., Johnson, K., Neill, A., Sheridan, D., & Andersen, G. R. (2016). Structural basis for eculizumab-mediated inhibition of the complement terminal pathway. *The Journal of Immunology*, *197*, 337–344.

[111] De Weers, M., Tai, Y.-T., Van Der Veer, M. S., Bakker, J. M., Vink, T., Jacobs, D. C., Oomen, L. A., Peipp, M., Valerius, T., Slootstra, J. W. et al. (2011). Daratumumab, a novel therapeutic human cd38 monoclonal antibody, induces killing of multiple myeloma and other hematological tumors. *The Journal of Immunology*, *186*, 1840–1848.

[112] Schrödinger. The pymol molecular graphics system, version 1.30.

[113] Walker, T. M., Miotto, P., Köser, C. U., Fowler, P. W., Knaggs, J., Iqbal, Z., Hunt, M., Chindelevitch, L., Farhat, M. R., Cirillo, D. M. et al. (2022). The 2021 who catalogue of mycobacterium tuberculosis complex mutations associated with drug resistance: a genotypic analysis. *The Lancet Microbe*, *3*, e265–e273.

CHAPTER 2

Var3D: Structure-Based Variant Analysis Framework

The work described in this chapter has been a collaborative effort between Erblin Asllanaj and Gabriel Studer.

Contributions: EA designed the overall software architecture, defined the variant format and selected the annotation tools. GS and EA implemented the software (names ordered by the amount of contribution).

Understanding the structural context of variants can provide valuable insights into their potential effects on protein function. However, the exponential growth of available genetic and structural data as well as the diversity of data formats and tools made it increasingly challenging to effectively analyse and interpret variant data. This creates the necessity for a software framework which is scalable, generically applicable to different variant and structure data sets and groups and applies various annotation methods efficiently to create a comprehensive overview of a mutation in its structural context.

To solve this problem, we created Var3D. It automates the integration of variant and structure data from various sources and provides interfaces for a diverse set of analysis tools. This allows to streamline annotation workflows which can be plugged together in a flexible manner, simplifying the analysis and interpretation of protein-coding variations from a structural perspective.

Using strategy patterns[1], we implemented reusable software for tasks common to all structure-based variant analysis pipelines. This approach reduces the amount of manual work required to implement a variant analysis pipeline and allows researchers to tailor the analysis to their specific needs and research goals.

In this chapter, we present Var3D, a general structure-based variant analysis framework which enables the customisation of data aggregation and data annotation processes for variant analysis in their 3D structure context. Var3D was used as the central framework for the implementation of TBvar3D, a web server for the automatic analysis of antibiotic resistance variants in MTB in Chapter 3 and the detection of critical polymorphisms in antibody-antigen interfaces in Chapter 4.

2.1 Methods

2.1.1 Software Architecture

Var3D divides variant analysis into two tasks: data import and data annotation. The primary focus of the data import is to parse variant and structural data from generic sources (web-based, variant table, . . .) into a standardised data aggregation of a reference sequence, variants,

and protein structures. This serves as input for the data annotation task which manages and runs individual annotation processes. The output of each annotation process is collected into an annotation data set which links the outputs to the respective data points in the data aggregation.

The separation of the pipeline into standardised interfaces (yellow in Figure 2.1) for recurring tasks and their problem-specific implementation (green in Figure 2.1) allows for exchanging and adapting the import and annotation processes to construct the desired analysis pipeline.

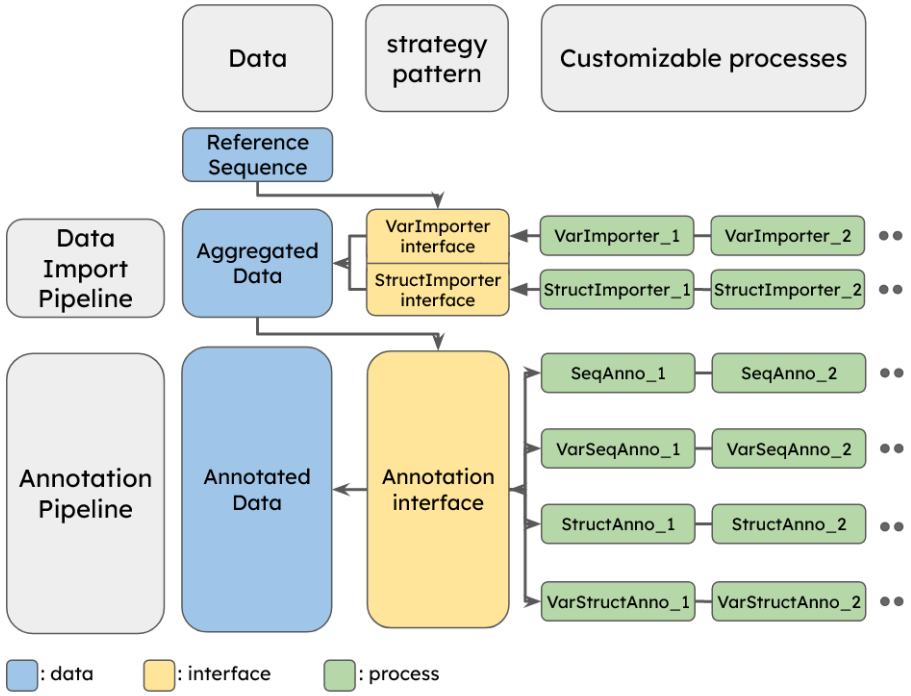

Figure 2.1: Schematic overview of the Var3D architecture. The central reference point is the protein reference sequence, to which any imported variant data and structure data refer to. After the data was aggregated through the VarImporter and StructImporter interface, annotation programs can be run on the aggregated data set. The annotation processes follow four categories: Annotations of the sequence (SeqAnno), annotations of variants only requiring sequence information (VarSeqAnno), Annotations of the structure (StructAnno) and annotations of variants requiring also structural information (VarStructAnno). The end result is an annotated data set in which the calculated annotations link to the respective data points in the aggregated data.

The task which needs to be implemented in the data import processes is the provision of variant and structure data in a compatible format. All other necessary tasks and checks are already provided in the *VarImporter*

and *StructImporter* interface. This includes checking the variant data set for redundancy, mapping variants, sequence and structures to each other and reporting data inconsistencies.

The annotation processes communicate through the annotation interface to i.) obtain the specific information they need to calculate or create their annotation successfully and ii.) return the results of the annotation process in a format which maps correctly to the respective data point in the data aggregation step.

2.1.2 Data Import

The central data piece and starting point of any Var3D pipeline is the protein reference sequence onto which variants and structures are mapped.

Each implementation of the *VarImporter* interface has the task to parse variant data into a human-readable (hr) variant format which consists of the following four parameters:

1. **Ref:** Reference sequence which will change upon mutation
2. **Pos:** Residue number of first amino acid changed in reference
3. **Alt:** Alternative sequence
4. **Variant type**

Variant Type	Description	Ref	Pos	Alt
SUBSTITUTION	Single amino acid substitution	A	176	G
INSERTION	Insertion of 1 or more amino acids	A	176	AHH
DELETION	Deletion of 1 or more amino acids	ART	176	A
INDEL	INSERTION and DELETION combined	ART	176	AHHG
STOP_GAINED	Special case of substitution, leading to stop codon	A	176	*
STOP_LOST	Special case of substitution, removing stop codon	*	237	A
START_LOST	Special case of substitution, removing start codon	M	1	L
FRAMESHIFT	Any event that leads to a frameshift	A	176	-
SYNONYMOUS	No effect on amino acid identity	A	176	A

Table 2.1: Overview of variant types which can be processed in Var3D together with examples of variants in the Var3D variant format.

The consistency of the indicated reference sequence is verified for every imported variant. Every variant which has a reference sequence, position and alternative sequence which is already present in the data aggregation will not be included again. It is also assessed if the variant type is consistent with the given variant information.

We provide two variant importers by default in the framework. One is the variant importer "HRVarImporter" (human readable variant importer) which parses strings in the format described in Table 1 (for example "A176G") into their respective Variant object. The second variant importer we provide is the "UniprotEntryVarImporter" which obtains the currently available variant data associated with a protein entry from the UniProt knowledge base[2], a database of protein sequence and functional information that is maintained by the UniProt consortium. Every variant associated with the provided entry (defined as all features which are annotated as "VARIANT") is obtained and added to the Var3D data aggregation.

Each implementation of the StructImporter interface needs to provide the means of loading structural data that refer to the underlying reference sequence. The representation of full macromolecular complexes beyond that reference sequence is explicitly supported in order to provide the full structural context for subsequent annotation processes. We use the Open-Source Computational Structural Biology Framework (OpenStructure)[3] to process structural data.

The Var3D framework contains two structure importers by default. The "FileSystemStructImporter" requires as the name indicates the path to a local protein structure file which is then added to the data aggregation. The "SMRStructImporter" utilises the application programming interface (API) of the SWISS-MODEL repository[4] to get all structure entries corresponding to a UniProt accession code. If structural information is available, the returned structures will include experimental structures from the Protein Data Bank (PDB)[5] or homology models from SWISS-MODEL[6].

The imported variants and structures are finally stored in a data aggregation which provides a mapping to the underlying reference sequence.

2.1.3 Data Annotation

The data annotation pipeline consists of several individual processes which are applied to the data aggregation created in the data import step. Depending on the chosen annotation tasks, the data annotation interface passes the necessary data to the annotator and performs consistency checks on the returned annotation results. In the example of

sequence-specific annotations, e.g. any suitable conservation score, the Var3D annotation interface sends a sequence to the annotator and expects a list with equal length as the annotation result.

The annotation process defines four interfaces which vary on the input required to perform the respective computations: sequence-level annotations (*SeqAnno*), variant annotations using only sequence features (*VarSeqAnno*), structure-level annotations (*StructAnno*), and variant annotations using both sequence and structure features (*VarStructAnno*).

SeqAnno: requires the reference sequence as input and is expected to return an annotation for every position in the sequence.

For this category, we provide two conservation scorers, the normalised Shannon entropy[7] and the ConSurf conservation score[8]. The Shannon entropy is a measure of information content for a protein residue in the context of a multiple sequence alignment on the UniRef90[9] protein sequence database. Lower information content correlates with higher conservation of the amino acid position. In contrast to the Shannon entropy, the ConSurf score explicitly considers the evolutionary relationships of the aligned homologous sequences. This score is based on an estimation of evolutionary rates and is expected to be more accurate than the conservation score based on Shannon entropy. Conservation scores can help to identify conserved regions in a protein in which mutations are expected to have a stronger impact[10].

Two additional sequence annotators were implemented for importing protein sequence annotations from different sources. One annotator imports functional annotations like binding sites, interaction sites and other variant annotations from the UniProt knowledge base. The other sequence annotator obtains functional and protein domain annotations from the InterPro database[11], which aggregates the results of functional annotations of proteins, classifications of proteins into families and predictions of domains and important sites from 13 different databases. Both annotators require that a UniProt accession code is provided. Functional sites which are located on or close to mutation sites can provide further context to hypothesise how the mutation could affect protein function. Variants located on the active site of an enzyme for example could indicate that the chemical capabilities of the enzyme are altered drastically.

VarSeqAnno: In addition to the reference sequence they require information on the amino acids which are exchanged by a mutation.

We provide the PROVEAN mutation impact scorer[12], which estimates the impact of mutations not limited to single amino acid substitutions. It measures the change in sequence similarity between the query sequence and a homologous sequence upon mutation of the query sequence. A set of homologs are searched for in the non-redundant protein sequence database[13] and an unbiased delta score is used to align the multiple delta scores into the PROVEAN score. If the introduced mutation reduces the similarity between the input sequence and many functional homologous proteins by a score less than -2.28, it is considered to be damaging. The PROVEAN score provides a quantitative measurement of the impact of a mutation. It can be considered to be a more precise measurement of the mutational impact than what is provided by conservation scores.

We also make use of the full capacity of the AAindex database[14] which contains a large set of numerical indices representing physicochemical and biochemical properties of amino acids and pairs of amino acids. The database consists of three sections: AAindex1 which contains 544 different amino acid indices corresponding to various chemical properties of amino acids which can be represented by a single numerical value. AAindex2 consists of 94 amino acid substitution matrices and AAindex3 has 47 different amino acid contact potential matrices. By using the respective AAindex accession code, the annotator can obtain the relevant values of the chosen index or matrix. Four AAindex properties are going to be used consistently in this thesis to describe the chemical distance between reference and alternative amino acid in a mutation: the hydrophobicity parameter pi[15], the molecular weight[16], the isoelectric point[17] and the STERIMOL length of the side chain[18]. These physicochemical features demonstrate how the size and the electrostatic properties in single amino acid substitutions are affected.

StructAnno: computed on the protein chains of all structures which map to the reference sequence.

We calculate the per-residue solvent accessibility using an OpenStructure implementation of an algorithm after Lee & Richards[19]. The annotator can return the absolute value of the accessible surface area or a relative value which is scaled by the theoretical maximum accessibility of the

respective amino acid. Buriedness is an important property to interpret variants. Mutations on buried sites for example are expected to have a higher impact on the protein stability of a single protein chain than on solvent-exposed surface sites.

The second structure annotation we provide is a predictor of transmembrane regions which is based on an implicit solvation model[20] also implemented in OpenStructure. The algorithm identifies transmembrane structures based on energetic and geometric criteria and estimates the orientation of the hypothetical membrane for predicted transmembrane proteins. The assumptions for the chemical environment of a mutation site change drastically if that site is located in a transmembrane region, which implies that the site is located in a hydrophobic lipid bilayer environment.

Residues located on protein-protein interfaces are detected and annotated by a specific structure annotator module. Every residue which is localised close to a different protein chain (the default detection distance is 5 Ångstrom) is annotated as an interface residue. The annotator distinguishes between homomeric interfaces if the protein chain in contact also maps to the reference sequence and heteromeric interfaces if the protein chain in proximity belongs to a different protein. Mutation sites on protein interfaces often have an impact on protein-protein interactions[21].

The last structure annotation we provide is the fully automated protein-ligand interaction profiler[22] (PLIP) applied to all the structures in the data aggregation. PLIP automatically detects all ligands present in the various structures and catalogues the chemical interactions between ligand and protein and identifies the respective protein residues involved in the interaction. Important chemical interactions with a ligand which cannot be maintained by the alternative amino acids are a strong indicator of an impactful mutation.

VarStructAnno: uses information on both the amino acid sequence changes introduced by the variant and 3D structure information.

For this category, we implemented the estimation of the free energy change upon mutation using the FoldX empirical force field[23]. The force field is based on an equation which sums up terms for hydrophobic interactions, polar and hydrophobic desolvations, hydrogen bond ener-

gies, electrostatic interactions, free energy change at protein interfaces of oligomeric proteins, entropy costs and clashes of amino acids. The difference between the energy in the wild-type structure and a structure. FoldX first estimates the total free energy in the wild-type structure. Then it introduces the mutation indicated by the currently processed variant into the structure and optimises the structure of the new protein variant. The free total energy is estimated again in the mutated structure and the difference between the wild-type energy estimation and the mutation energy estimation is returned. High free energy differences indicate that the protein is not coping well with the introduction of the respective mutation.

The results of all the annotation processes are collected and stored in an annotation data structure. This dictionary links the annotations to their respective data structures in the data aggregation and contains a log if errors in the single annotation processes were encountered.

VarImporter	
HRVarImporter	Parses variants from human readable (HR) strings, e.g. A123AB
UniprotEntryVarImporter	Fetches and parses variants from UniProt entry
StructImporter	
SMRStructImporter	Fetches structures from the SWISS-MODEL repository (SMR)
FilesystemStructImporter	Fetches user-defined structures from disk
SeqAnno	
EntropySeqAnno	Shannon entropy based on MSA
ConsurfSeqAnno	Annotation with ConsurfDB pipeline
UniProtSeqAnno	Functional annotations from UniProtKB
InterProSeqAnno	Functional and domain annotations from InterProKB
VarSeqAnno	
ProveanVarSeqAnno	Variant annotation based on Provean
AAIndexVarSeqAnno	Variant annotation based on the AAindex DB
StructAnno	
AccessibilityStructAnno	Annotations based on Solvent Accessibilities
TransmembraneStructAnno	Classifies if a structure has transmembrane-like properties. If yes, the optimal membrane positioning is added too.
InterfaceStructAnno	Annotates interface residues
PLIPStructAnno	Annotates protein-ligand interactions with PLIP
VarStructAnno	
FoldXVarStructAnno	Variant annotation based on the "BuildModel" function of FoldX

Table 2.2: Overview over Importer and Annotator functionalities implemented in the Var3D framework.

2.2 Results

2.2.1 Deployment of Var3D

The Var3D toolbox was implemented using Python and can be accessed at the following Git repository:
https://git.scicore.unibas.ch/schwede/var3d. To run Var3D, a Singularity container[24] must be set up to satisfy the software dependencies of Var3D and the collection of provided importers and annotators. While the container itself is not provided, the "def_builder.py" Python script can be used to create a recipe file for the Var3D container construction.

After the Var3D container is set up, the pipelines provided in the repository can be executed, or a new variant analysis pipeline can be developed. If new software dependencies are required, it needs to be ensured that the Var3D Singularity container is updated accordingly.

2.2.2 Var3D Pipelines

The first pipeline is a tutorial example to demonstrate and test the basis of the structure of a Var3D pipeline. There, we annotate the small protein crambin with a single example variant and apply the two mentioned conservation score annotators, an AAIndex annotation and the FoldX annotator to it. The structures for the crambin protein are obtained using the SMR structure importer. The full pipeline is implemented in a single short script and provides a good starting point for the implementation of a customised pipeline.

The second pipeline demonstrates an application of Var3D with biological relevance: the "sars_cov2" pipeline, which analyses variants that lie on the polyprotein of SARS-CoV-2. It uses the UniProt variant importer and the SMR structure importer to get variant and structure data from sources which will be updated periodically with the latest information on the pathogen. The pipeline then annotates the aggregated data with the two conservation scores and with the relative solvent accessibility.

The "tbvar3d" pipeline is the central piece of the TBvar3D web server and will be the object of discussion in the next chapter. The variant and structure data are obtained from a local source. All the annotators described in the subsection "Data annotation" were applied to this data

aggregation. Further details on the analysis pipeline can be found in Chapter 3.

The fourth pipeline is the "antibody" pipeline. Similar to the TBvar3D pipeline, variants and structures were obtained from a local data set. Besides the conservation scores, PROVEAN scores and certain AAindex features provided by the default Var3D framework, it was required to implement three new annotators: a detector of epitope residues and the separate calculations of the relative surface accessibility and free energy difference upon mutation for the structure of the full antibody-antigen complex and the apo form of the antigen. A more in-depth description of this pipeline can be found in Chapter 4.

var3d/pipelines/**crambin_annotation**.py	A simple tutorial pipeline which manually creates and imports one variant lying on the crambin protein in a single script. The annotations calculated are normalised entropy, ConSurf score, amino acid similarity and the FoldX free energy difference.
var3d/pipelines/**sars_cov2**/	Imports variants from the UniProt entries and structures from the SMR entries of the SARS-Cov2 polyprotein and calculates normalised entropy, ConSurf score and the relative solvent accessibility for all variants.
var3d/pipelines/**tbvar3d**/	Imports variants and structure from two manually defined databases and uses several annotations (Chapter 3)
var3d/pipelines/**antibody**/	Imports variants and structure from two manually defined databases and uses several annotations (Chapter 4)

Table 2.3: Pipelines provided in the Var3D Git repository. The path to the pipeline script or folder is provided in the left column, with the title of the pipeline highlighted. A short description is provided on the right. The last two pipelines are described in more detail in the indicated chapters of this thesis.

In Table 2.4 is an overview of the data processed by the presented pipelines together with the total computational time required to calculate the annotations indicated in Table 2.3. While a small number of variants and structures can be handled manually, the table demonstrates the ability of Var3D to process the combination of large variant and structure data sets.

Var3D Pipeline	Reference Sequences	Total Annos	Total Variants	Total Structures	Accumulated Run Time	Longest Running Task
crambin_annotation	1	4	1	26	0 h 11 min 47 sec	0 h 11 min 47 sec
sars_cov2	20	3	109	1722	5 h 47 min 24 sec	1h 35 min 48 sec
tbvar3d	69	10	10'719	64	110 h 12 min 54 sec	3 h 42 min 25 sec
antibody	62	11	10' 352	114	142 h 59 min 32 sec	5 h 15 min 15 sec

Table 2.4: Overview over the amount of data aggregated by the pipelines and the computation time. The computations were performed on a heterogeneous compute cluster (sciCORE, scientific computing centre at the University of Basel). The Var3D pipeline for each reference sequence was submitted as independent computation tasks.

2.3 Discussion

To efficiently process large amounts of variant, protein structure and annotation data, we required a robust framework and the definition of data standards. The usefulness of a stable workflow which enables the handling of a diverse range of requests for variant interpretation would not be only limited to the projects described in this dissertation.

With Var3D, we implemented a framework for the general analysis of a large scale of variant data together with their respective structure data. Processes required for every structure-based investigation of variants were automatized through the standardised software interfaces and we provide a set of importers and annotators which allow the construction of a custom variant analysis pipeline.

We demonstrate the capabilities of Var3D by the implementation of four different variant analysis pipelines, two of which will be prominently featured in this dissertation. The next chapter will show how the Var3D framework can also be used as the central server-side software component in the implementation of a variant analysis web server.

The development of new prediction tools based on artificial intelligence relies on large, consistent and homogeneously generated datasets. A data standard and framework for variants would open up new possibili-

ties for the application of these groundbreaking methods to predict the impact of mutations. The Var3D framework enables the generation of large, standardised data sets which can be utilised for machine learning applications.

Var3D is focused on the investigation of the effect of variants on the level of protein structures but as discussed in the Introduction, this would only constitute the first step (Figure 1.1). The next step would be to investigate how the change in the protein level would impact the biological system in the cell. To enable researchers to build on top of this framework, the code is made available as open-source.

2.4 Publication and Code Availability

The code base has been published on https://git.scicore.unibas.ch/schwede/var3d. Description of the framework will be published in the context of a manuscript describing the TBvar3D web server in chapter 3.

References

[1] Gamma, E., Johnson, R., Helm, R., Johnson, R. E., & Vlissides, J. (1995). *Design patterns: elements of reusable object-oriented software*. Pearson Deutschland GmbH.

[2] UniProt Consortium (2019). Uniprot: a worldwide hub of protein knowledge. *Nucleic acids research*, *47*, D506–D515.

[3] Biasini, M., Schmidt, T., Bienert, S., Mariani, V., Studer, G., Haas, J., Johner, N., Schenk, A. D., Philippsen, A., & Schwede, T. (2013). Openstructure: an integrated software framework for computational structural biology. *Acta Crystallographica Section D: Biological Crystallography*, *69*, 701–709.

[4] Bienert, S., Waterhouse, A., De Beer, T. A., Tauriello, G., Studer, G., Bordoli, L., & Schwede, T. (2017). The swiss-model repository—new features and functionality. *Nucleic acids research*, *45*, D313–D319.

[5] Burley, S. K., Berman, H. M., Kleywegt, G. J., Markley, J. L., Nakamura, H., & Velankar, S. (2017). Protein data bank (pdb): the single global macromolecular structure archive. *Protein Crystallography*, (pp. 627–641).

[6] Waterhouse, A., Bertoni, M., Bienert, S., Studer, G., Tauriello, G., Gumienny, R., Heer, F. T., de Beer, T. A. P., Rempfer, C., Bordoli, L. et al. (2018). Swiss-model: homology modelling of protein structures and complexes. *Nucleic acids research*, *46*, W296–W303.

[7] Shannon, C. E. (1948). A mathematical theory of communication. *The Bell system technical journal*, *27*, 379–423.

[8] Ben Chorin, A., Masrati, G., Kessel, A., Narunsky, A., Sprinzak, J., Lahav, S., Ashkenazy, H., & Ben-Tal, N. (2020). Consurf-db: An accessible repository for the evolutionary conservation patterns of the majority of pdb proteins. *Protein Science*, *29*, 258–267.

[9] Suzek, B. E., Wang, Y., Huang, H., McGarvey, P. B., Wu, C. H., & Consortium, U. (2015). Uniref clusters: a comprehensive and scalable alternative for improving sequence similarity searches. *Bioinformatics*, *31*, 926–932.

[10] Celniker, G., Nimrod, G., Ashkenazy, H., Glaser, F., Martz, E., Mayrose, I., Pupko, T., & Ben-Tal, N. (2013). Consurf: using evolutionary data to raise testable hypotheses about protein function. *Israel Journal of Chemistry*, *53*, 199–206.

[11] Blum, M., Chang, H.-Y., Chuguransky, S., Grego, T., Kandasaamy, S., Mitchell, A., Nuka, G., Paysan-Lafosse, T., Qureshi, M., Raj, S. et al. (2021). The interpro protein families and domains database: 20 years on. *Nucleic acids research*, *49*, D344–D354.

[12] Choi, Y., & Chan, A. P. (2015). Provean web server: a tool to predict the functional effect of amino acid substitutions and indels. *Bioinformatics*, *31*, 2745–2747.

[13] Pruitt, K. D., Tatusova, T., & Maglott, D. R. (2007). Ncbi reference sequences (refseq): a curated non-redundant sequence database of genomes, transcripts and proteins. *Nucleic acids research*, *35*, D61–D65.

[14] Kawashima, S., Pokarowski, P., Pokarowska, M., Kolinski, A., Katayama, T.,

& Kanehisa, M. (2007). Aaindex: amino acid index database, progress report 2008. *Nucleic acids research*, *36*, D202–D205.

[15] Fauchere, J.-L. (1983). Hydrophobic parameters pi of amino-acid side chains from the partitioning of n-acetyl-amino-acid amides. *Eur. J. Med. Chem.-Chim. Ther.*, *18*, 369–375.

[16] Fasman, G. (1976). Handbook of biochemistry and molecular biology. *3rd ed., Proteins - Volume 1*, .

[17] Zimmerman, J., Eliezer, N., & Simha, R. (1968). The characterization of amino acid sequences in proteins by statistical methods. *Journal of theoretical biology*, *21*, 170–201.

[18] FAUCHÈRE, J.-L., Charton, M., Kier, L. B., Verloop, A., & Pliska, V. (1988). Amino acid side chain parameters for correlation studies in biology and pharmacology. *International journal of peptide and protein research*, *32*, 269–278.

[19] Lee, B., & Richards, F. M. (1971). The interpretation of protein structures: estimation of static accessibility. *Journal of molecular biology*, *55*, 379–IN4.

[20] Lomize, A. L., Pogozheva, I. D., Lomize, M. A., & Mosberg, H. I. (2006). Positioning of proteins in membranes: a computational approach. *Protein Science*, *15*, 1318–1333.

[21] Yates, C. M., & Sternberg, M. J. (2013). The effects of non-synonymous single nucleotide polymorphisms (nssnps) on protein–protein interactions. *Journal of molecular biology*, *425*, 3949–3963.

[22] Adasme, M. F., Linnemann, K. L., Bolz, S. N., Kaiser, F., Salentin, S., Haupt, V. J., & Schroeder, M. (2021). Plip 2021: expanding the scope of the protein–ligand interaction profiler to dna and rna. *Nucleic acids research*, *49*, W530–W534.

[23] Schymkowitz, J., Borg, J., Stricher, F., Nys, R., Rousseau, F., & Serrano, L. (2005). The foldx web server: an online force field. *Nucleic acids research*, *33*, W382–W388.

[24] Kurtzer, G. M., Sochat, V., & Bauer, M. W. (2017). Singularity: Scientific containers for mobility of compute. *PloS one*, *12*, e0177459.

CHAPTER 3

TBvar3D: Mapping Antibiotic Resistance Variants in Mycobacterium Tuberculosis on 3D Protein Structures

The work described in this chapter has been a collaborative effort between Erblin Asllanaj, Andrew Waterhouse, Gabriel Studer and Rosalba Lepore.

Contributions: RL designed and supervised the project. EA and RL designed the pipeline and performed the analysis of the catalogue. EA and GS implemented the backend of the TBvar3D web server. AW and EA implemented the front end of the web server. All names are ordered by the amount of contributions.

The antibiotic resistance crisis is one of the greatest worldwide challenges to medical research and society. The continuous spread of multi-drug resistant bacterial strains in combination with a shortage of new antibiotic drugs puts healthcare systems worldwide in a precarious situation[1]. In 2018, the World Health Organisation (WHO) reported about 500'000 new cases of multidrug-resistant tuberculosis (MDR-TB), which are resistant against the two most powerful antitubercular drugs, Rifampicin and fluoroquinolones, with an estimation of further 1 million unreported cases[2].

Besides the need for new antibiotic drugs, quick and reliable diagnosis of the drug resistance status of pathogens is essential to exercise better and immediate effective control over the situation. Variants with a strong resistance phenotype are used as genetic markers of resistance for antibiotic resistance diagnosis. But these well-characterised variants are vastly outnumbered by variants for which the phenotypic consequence is not known.

TB researchers across the world would highly benefit from an accessible web service which facilitates a structure-based analysis of new variants in the context of established resistance variant data and relevant structure models of variant target proteins. This enables the exploration of variants with an unknown consequence concerning drug resistance and could enable the selection of promising candidates for further experimental characterisation.

In this chapter, we describe the variant analysis web server TBvar3D. The web server is based on the Var3D software presented in the previous chapter and provides the user with an up-to-date web-based environment that streamlines data integration, analysis and hypothesis generation on the role of given variants in MTB drug resistance. TBvar3D has no login requirement and is freely available at https://swissmodel.expasy.org/var3d/.

3.1 Methods

3.1.1 Analysis Pipeline

The backend of the TBvar3D web server is based on two Var3D pipelines: one for precomputing annotations of all variants of the WHO catalogue

of drug resistance mutations in the MTB complex[3] and a second for processing user-submitted variants in real-time for the TBvar3D web server (Figure 3.1).

The precomputation pipeline uses a customised variant importer which directly parses the unprocessed WHO variant catalogue file. The structures are loaded through the file system structure importer from the curated TBvar3D structure database (see subchapter 3.1.3). The annotation pipeline contains all annotators which were described in Chapter 2 under the "Data annotation" section. The output is stored for future use in the TBvar3D web server, where the catalogue variants can be explored on their own.

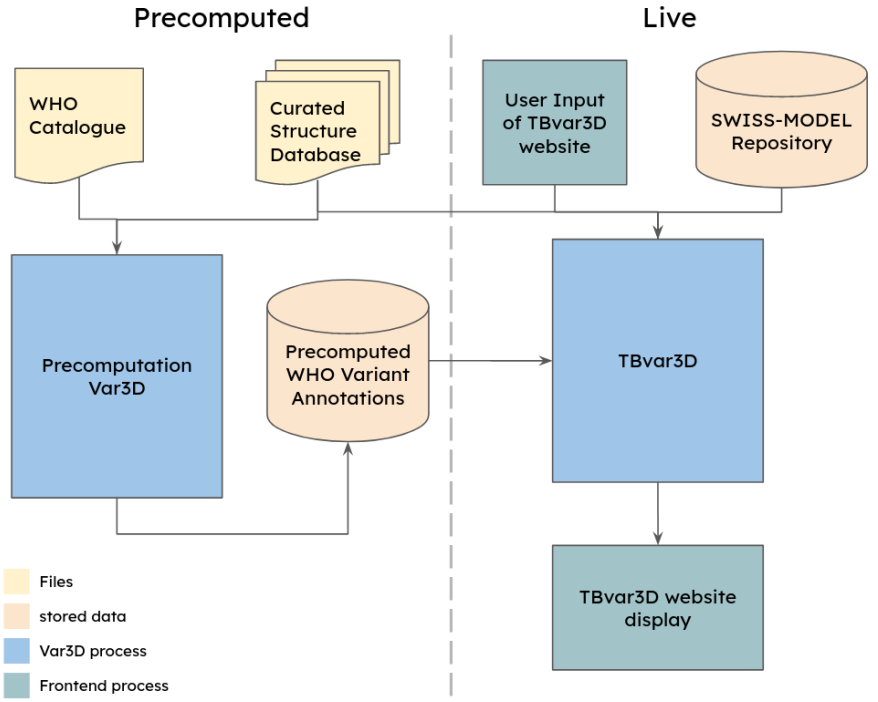

Figure 3.1: Schema of the overall TBvar3D software architecture. The left pipeline precomputes variant annotations in the WHO catalogue using the protein structure models from the curated structure database. The results are stored and then reused in the TBvar3D web server. The user of the server can submit variants on any protein in the MTB proteome. If the protein is part of the resistance target set represented in the curated structure database, then the structure and precomputed WHO variant annotations will be obtained from there, else it is provided by the SWISS-MODEL repository.

The web server pipeline mirrors the precomputation pipeline but processes a user input which consists of a protein together with a list of variants located on it. The structure data is either obtained from the TBvar3D structure database if the indicated protein is represented in it, or obtained from the SWISS-MODEL repository otherwise. Variants from the WHO catalogue are included if they map to the user-provided input protein.

3.1.2 Curation of the WHO MTB Mutation Catalogue

The WHO catalogue of drug resistance mutations in the MTB complex is a worldwide effort to categorise and standardise the current knowledge of antibiotic resistance variants in MTB[3]. The catalogue provides a reference standard for the interpretation of mutations conferring antibiotic resistance. The current release contains 18 '446 genetic mutations and their resistance phenotype towards 13 antibiotics commonly used in MTB treatment.

We implemented a Var3D variant importer which processes the variant list from the raw data set of the variant genome indices made available by the WHO. The variants in the catalogue use as reference the protein sequences from Mycobrowser[4], a TB-specific gene and protein sequence database with computationally generated and manually reviewed information dedicated to complete genomes of *Mycobacterium tuberculosis*, *Mycobacterium leprae*, *Mycobacterium marinum* and *Mycobacterium smegmatis*. All the Mycobrowser sequences in the catalogue were mapped to proteins in the UniProtKB proteome ID UP000001584 (MTB strain ATCC 25618 / H37Rv, access date 01.02.2022) which TBvar3D is using as the reference proteome from this point on. The mapping to UniProt is required for the use of the Var3D SMR structure importer and the Var3D Uniprot feature annotator. The alignments were performed using the semi-global alignment method (Needleman/Wunsch without gap penalty) from OpenStructure[5].

In three cases, the UniProt protein sequence starts at a later position than the respective Mycobrowser sequence. 23 variants were mapped to a part of the protein which was not covered by the respective UniProt sequence and could not be processed.

A mismatch of the starting amino acid is observed for 27 proteins but

was not considered as relevant for the analysis. Modifications of the starting codon prevent the expression of the full protein. UniProt protein sequences consistently start with a methionine while MycoBrowser protein sequences can start with other amino acids: Valine (25 times in this protein set) or Leucine (2 times) which is supported by the fact that alternative starting codons for MTB are expected[6].

3.1.3 Curation of the TBvar3d Target Structure Database

The structure of proteins not represented in the WHO catalogue is extracted by the Var3D SMR structure importer. The SMR[7] has a weekly update cycle and provides a ranked selection of PDB experimental structures and SWISS-MODEL homology models for every protein in the MTB proteome.

The structures of the proteins in the WHO catalogue were manually curated. The two goals of the selection are (i) the presence of the antibiotic ligand in the proper binding site if the protein is a known drug target and (ii) the proper oligomeric state of the protein target.

Whether an antibiotic ligand was expected to be present in a protein was decided by conducting an extensive literature search on the resistance mechanisms of all targets (Supplementary, Table 3.1). The oligomeric state predictions were provided by the SWISS-MODEL repository (SMR)[8].

Experimentally resolved structures of the targets were obtained directly from the Protein Data Bank wherever possible. Other targets were obtained from the AlphaFold DB[9] which contains structure predictions of single protein chains. The modelling of homo-oligomeric targets without experimental structures was performed by using AlphaFold-Multimer (v.2.1.1)[10].

Three protein targets are part of a complex oligomeric structure which could not be modelled with AlphaFold due to size limitations. These proteins were the ATP-synthase subunits atpE and subunit atpB which are part of the large ATP synthase machinery (an 18-mer consisting of 8 different proteins, Figure 3.2)) and a homo-6-mer of the large (848 residues) ATP-dependent Clp protease. We used SWISS-MODEL[11] to create a homology model of these complexes using as templates structure

of the ATP synthase in complex with Bedaquiline in *Mycobacterium smegmatis* (PDB ID 7jga) and of the Clp protease in *Escherichia Coli* (PDB ID 5og1). The average model confidence for the ATP synthase model was on the higher end with a QMEANDisCo score[12] of 0.82 while the confidence for the Clp protease was scored at 0.55.

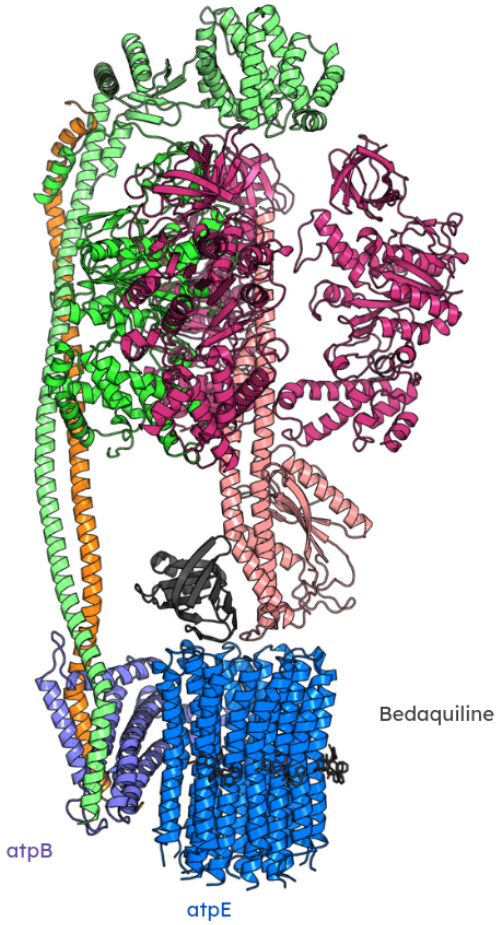

Figure 3.2: Homology model of the ATP-synthase complex based on the template 7jga. The template contained the full complex of the protein machinery with Bedaquiline, a novel antibiotic. Presumed resistance variants are located either on the atpE subunits or the atpB subunits which are both predicted to be immersed in the membrane of the bacteria.

The drug targets gyrase subunit B and subunit A have incomplete experimental structures of the gyrase complex with DNA and the two fluoroquinolones Moxifloxacin and Gatifloxacin[13]. Roughly 60% of the N-terminal gyrA sequence and 40% of the C-terminal gyrB sequence are

covered in the available experimental structures. Using AF2 single chain models, the structure was completed by superposing the chains on the experimental structure with the UCSF Chimeras MatchMaker algorithm[14] (Figure 3.3). We detected no clashes between the two DNA chains, the two ligand molecules and the gyrase subunits using the clash detector of UCSF Chimera.

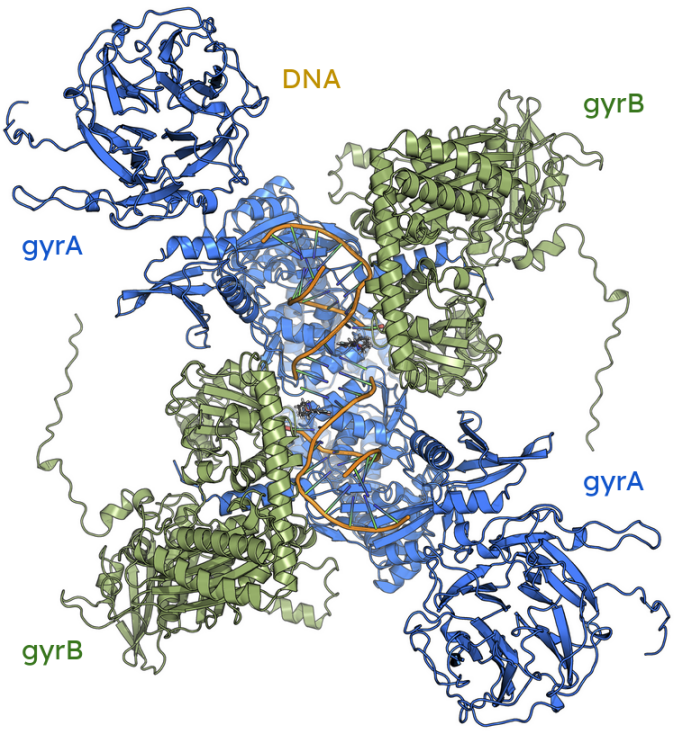

Figure 3.3: Assembled model of the gyrase complex using an experimental scaffold and AF2 predictions of the single subunits. The model shows for the first time a predicted beta-propeller domain in the gyrA subunit with an unknown function and an additional domain in the gyrB subunit.

For targets with no experimental structure of the drug-ligand complex but with a homologous experimental structure which contained the drug of interest, the ligand could be modelled by superposition of the ligand template with the target structure. The ligand was either transferred directly by the modelling tool (example in Figure 3.2) or transferred using the UCSF Chimeras MatchMaker to superpose the ligand template with the structure of the target.

For the drug target of Linezolid, the 50S ribosomal protein rplC, an experimental structure exists of the complex with a Linezolid analogue (Linezolid-114)[15]. We superposed the Linezolid molecule (obtained from PubChem[16]) with its analogue by using the field-based ligand alignment of the ligand-based design software Flare™[17] from Cresset®. The algorithm uses a combination of atomic and electrostatic features of the ligand and the binding pocket to calculate a more accurate superposition between ligand molecules.

We used the molecular docking program AutoDock Vina 1.2.0[18] to predict the binding pose of three targets (Two fragments of Pyrazinamide (123.11 Da) and Ethionamide (166.244 Da), and Delamanid (534.48 Da)) in their respective protein targets (pncA (Experimental), ethA (AF2), and ddn (AF2)). The docking of Pyrazinamide could be confined to a 10x10x10 Ångstrom search space due to an active site annotation. The best poses were selected through a combination of binding affinity estimations, manual inspection[19], and interaction analysis. However, it should be noted that the results may not be highly accurate due to the fragment ligands and AF2 models as targets[20,21].

Structures and models of the curated TBvar3d target structure database were deposited to ModelArchive [9]. ModelArchive is a database for protein structure models where users can deposit their model structure data sets with detailed information on the methods (parameters, the version number of tools, etc.) used for their structure predictions. The ModelArchive entry of the structure database (Accession code "10.5452/matbvar3d") contains detailed descriptions of the data sources, pipelines and tools used.

3.2 Results

3.2.1 WHO Mutation Catalogue Data Set

While parsing the 18'446 genetic mutations from the catalogue, the TBvar3D variant importer discarded variants which are located outside protein-coding regions (1'347), synonymous variants (4'922), variants with ambiguous or missing genetic locus (991), duplicated variant entries (836) and ambiguous descriptions of the amino acid mutations (the nature of the mutation is not described) or sequence mapping mismatches

(57). This leaves 10'293 protein-coding variants across 66 unique protein targets, broken down as follows (Figure 3.4): 928 (9%) are classified as resistance variants and 110 (1.06%) are categorised as susceptible or neutral variants. The remaining variants (9'255) lack a definite grading and are thus considered uncertain.

The majority of the variant data set contains single amino acid substitutions. 9' 139 of the variants (88.7%) in the data set belong to this variant type with the overwhelming majority of these substitutions being labelled as uncertain (8' 736). Susceptible variants in this data set are, with one exception in the indel group, represented by this variant type with 109 variants. The next group of represented variant types are frameshifts which constitute 723 (7%) of the variant data set. In this variant group, half are considered as resistant while the other half was labelled "uncertain". The other variant types follow the same pattern. Mutations leading to a loss of the start codon are all labelled uncertain.

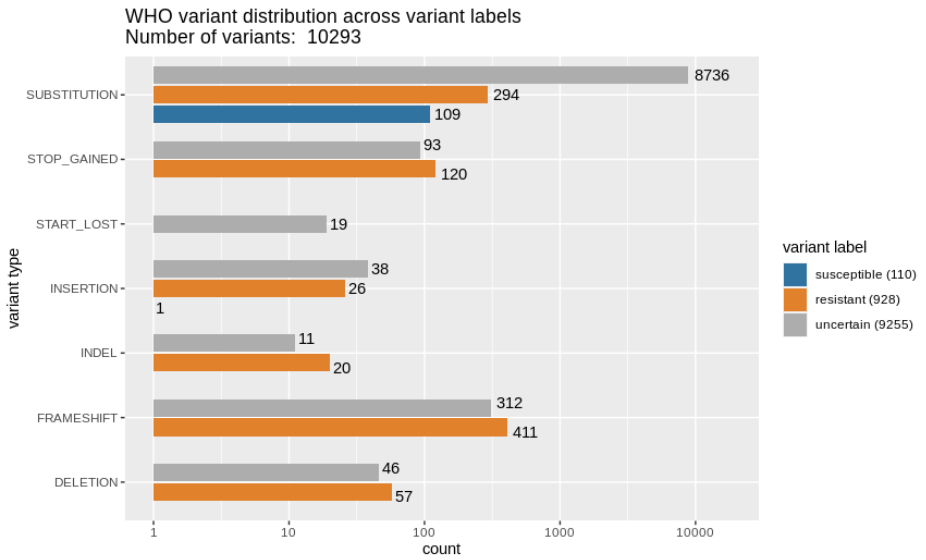

Figure 3.4: Annotations of the WHO mutation catalogue variant data set. The variants are split according to their variant type (see Chapter 2, Table 2.1) and the variant label they received from the WHO.

The distribution of resistance variants within the catalogue across protein targets can be observed in Figure 3.5. Prodrug activators exhibit the highest number of resistance variants (625), with notable examples including the enzymes katG, which activates the first-line drug Isoniazid and pncA, which activates the first-line drug Pyrazinamide. Drug targets exhibit the second highest number of resistant variants (152), with the RNA polymerase subunit rpoB, a target of the first-line antibiotic Rifampicin, predominating in this category (122). The two methyltransferases rmsG and tlyA comprise the third largest group of resistant variants (147) due to their mechanism of resistance, which involves the disruption of methylation function, similar to prodrug activators. Methylation of ribosomal RNA is required for the proper function of aminoglycoside antibiotics. The distribution of variants classified as resistance is heavily skewed towards proteins whose mechanism involves the disruption of function, as well as prominent drug targets.

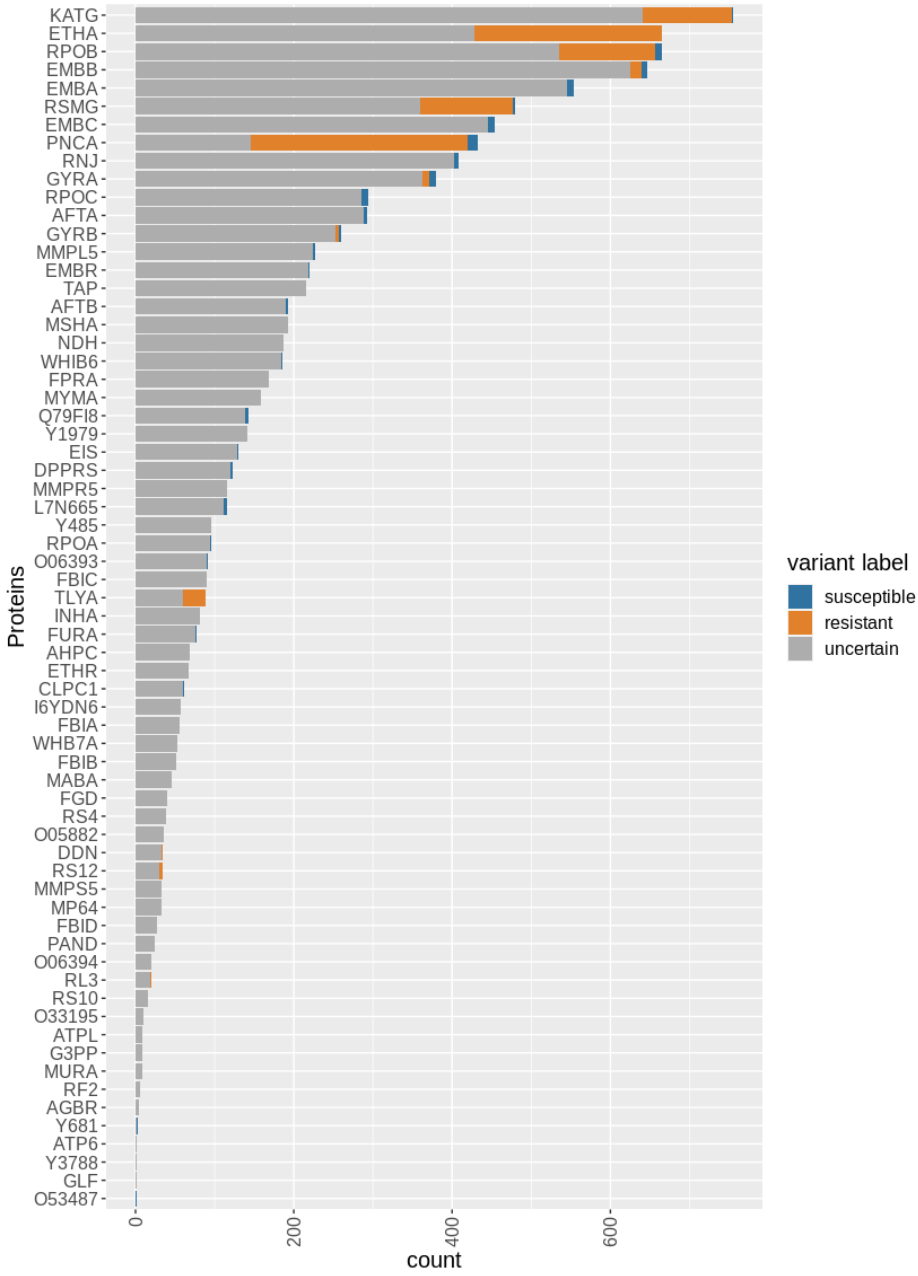

Figure 3.5: Distribution of variants in the mutation catalogue across targets in the data set.

The distribution of the ConSurf conservation scores (Figure 3.6) suggests that resistance variants tend to be located at more conserved sites than both susceptible variants and uncertain variants. The comparisons of distributions show a significant difference (Wilcoxon rank sum to test for the equality of the means) between resistance variants and susceptible variants.

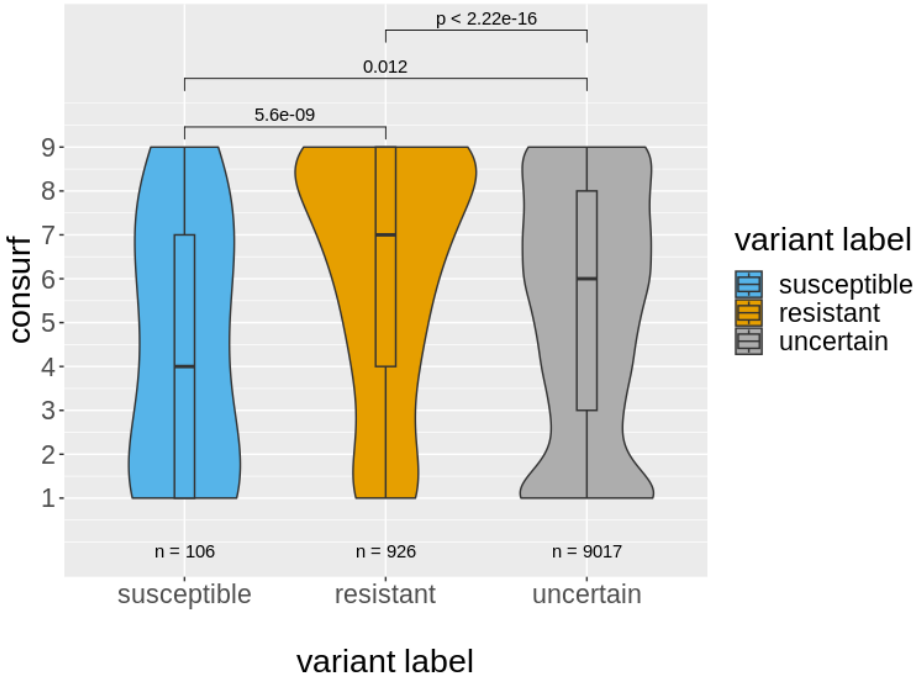

Figure 3.6: Conservation score distributions in the WHO variant data set. The violin plot above shows the distribution of the ConSurf score which ranges from 1 (not conserved) to 9 (very conserved). The displayed statistical testing results between the variant label categories used the Wilcoxon rank sum to test for the equality of the means.

The alignment-based PROVEAN score follows a similar trend (Figure 3.7). Variants which were labelled as resistant tend to have lower PROVEAN scores than susceptible variants. The variant type can have a drastic effect on the score: deletions, insertions and indels are scored with strongly negative values. The distribution of scores in the resistant variants tends to be lower than the PROVEAN deleteriousness threshold.

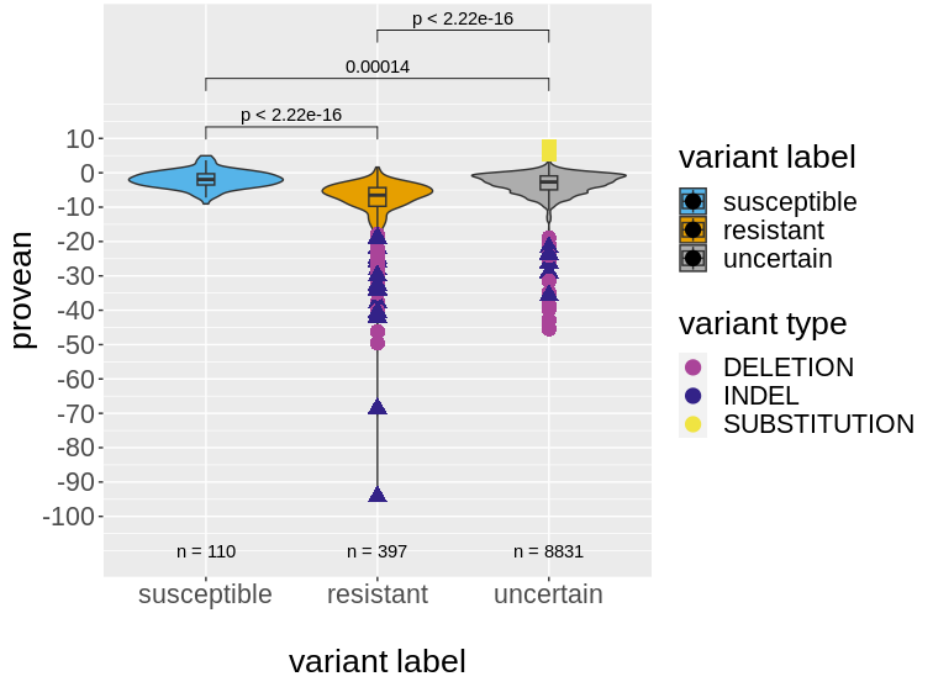

Figure 3.7: PROVEAN score distributions in the WHO variant data set. The violin plot above shows the distribution of the numeral score. Variants which have a value below -2.28 (red line) are considered deleterious The displayed statistical testing results between the variant label categories used the Wilcoxon rank sum to test for the equality of the means. The outlier points are coloured by variant type.

The distribution of relative solvent accessibility (RSA) values (Figure 3.7) also suggests that resistant variants tend to be more buried in the protein than susceptible variants. The mean RSA value for resistant variants is 15% lower than for susceptible variants. The largest contribution to this effect comes from the large number of resistance mutations mapping to the Rifampicin Resistance Determining Region (RRDR) which is part of the deeply buried Rifampicin binding site in rpoB.

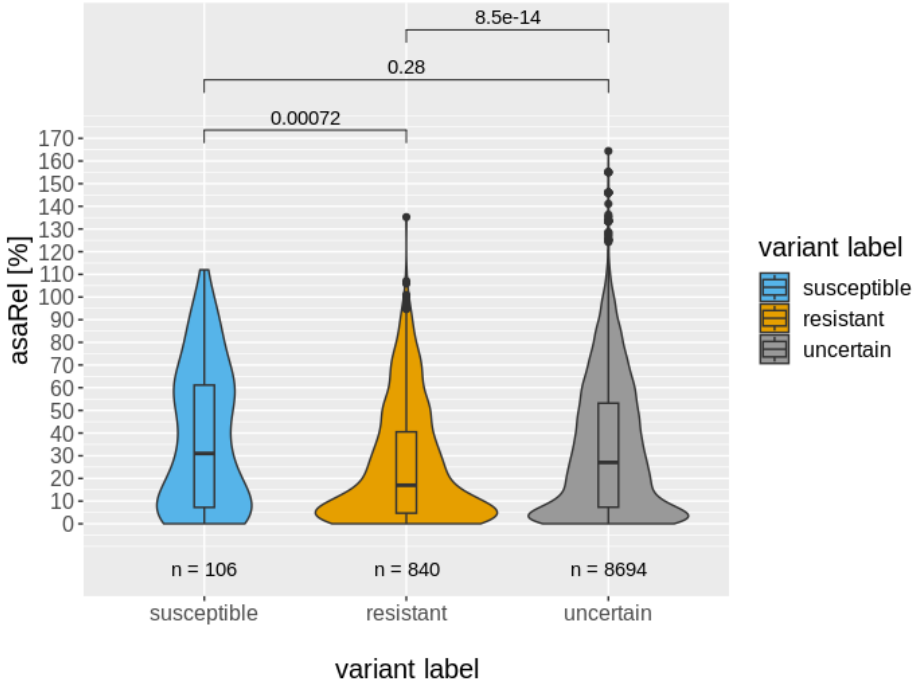

Figure 3.8: Relative solvent accessibility score distributions of the WHO variant data set. The value is usually a percentage between 0% and 100% but can be higher than 100% for highly exposed residues (improper reference values for normalisation can lead to RSA values above 1[22]). The displayed statistical testing results between the variant label categories used the Wilcoxon rank sum to test for the equality of the means.

3.2.2 TBvar3D Structure Data Set

The curated TBvar3D structure data set was created by extending experimental structures of TB proteins with a diverse set of structure prediction methods (Figure 3.9). Among the 66 protein targets in the WHO mutation catalogue, 28 have experimental structures present in the Protein Data Bank which could be used directly. Further, 35 proteins have been modelled using AlphaFold2 (AF2). 23 of these proteins were predicted to have a monomeric state which meant that the prediction from the AlphaFold DB could be used directly while 10 structures predicted to adopt a homo-oligomeric state were modelled using AlphaFold-Multimer. 3 targets required the modelling of large heteromeric complexes and could be modelled using SWISS-MODEL (see Methods). The partial experimental structures of the gyrase complex containing the two targets gyrA and gyrB were completed by superposing the respective AF2 models on the experimental structure (see Methods).

Among the 66 protein targets, 20 targets were determined to require the presence of a drug ligand in the structure. 10 of these targets have a resolved experimental structure of the drug target complex. The SWISS-MODEL pipeline transferred the ligand from the manually selected template for two protein targets. For four targets, we found homologous experimental structures which contained a drug of interest and through superposition they were transferred to the AF2 structure of the original target. One target had the experimentally resolved structure of a complex with an antibiotic analogue which could be replaced using field-based alignment. The remaining three targets had no template information on the ligand, so the ligand pose was predicted by molecular docking using AutoDock-Vina (see Methods).

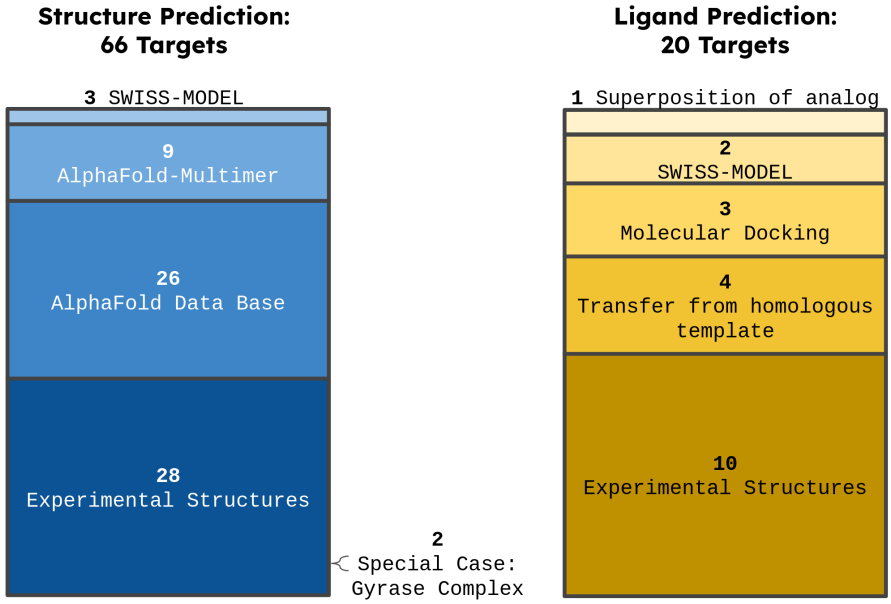

Figure 3.9: Overview over methods used to obtain the structures of the targets and the drug target complexes.

3.2.3 Usage Of TBvar3D

TBvar3D has been developed as a web server for mapping and analysing MTB variants in the context of protein structures and the comprehensive resistance variant catalogue provided by the WHO. The TBvar3D homepage (Figure 3.10) allows a user to initiate the variant analysis by entering a UniProt accession code for an MTB protein and a list of variants of interest in the protein. Upon entering a valid UniProt accession code, the corresponding protein sequence is displayed to assist the user in entering their variants of interest. Any errors in the variant input are interactively communicated to the user and prompted for correction. The user can also access previously submitted variant analysis projects through the projects panel located at the bottom of the page, which is stored for a period of two weeks. Additionally, the user has the option to access a target exploration page for the inspection of WHO catalogue variants through the exploration button located at the top of the webpage.

The target exploration page (Figure 3.11) consists of a table which lists the links to the results page of 66 protein targets and contains information on the antibiotics linked to the target, a categorization of the mechanism of resistance linked to the target and the number of resistant, susceptible and uncertain variants.

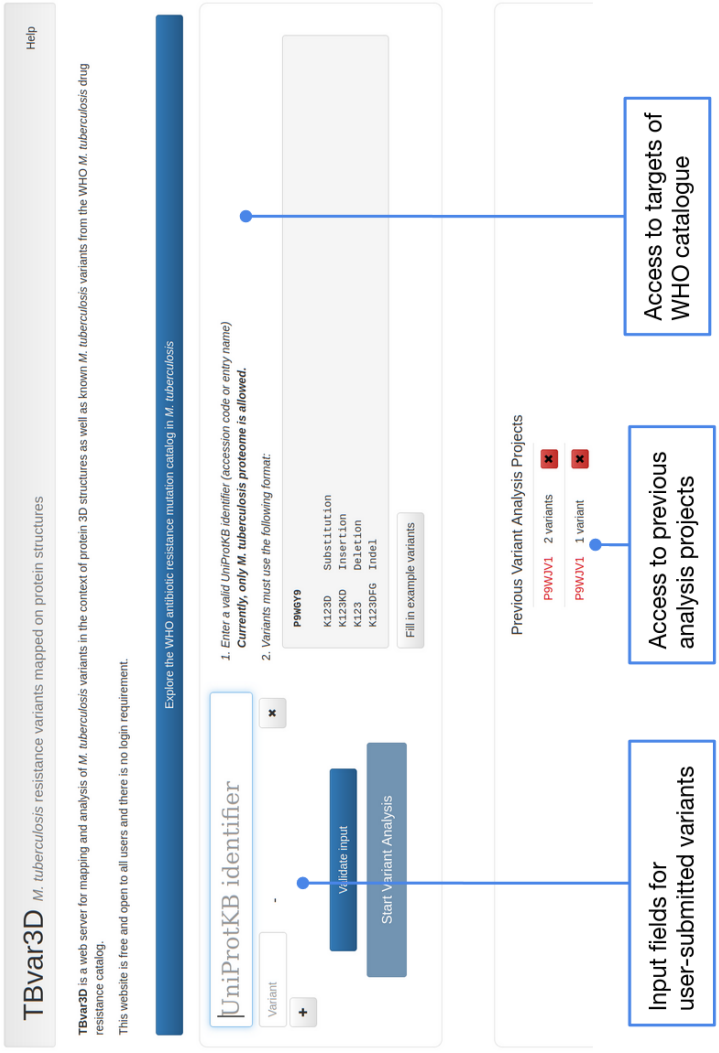

Figure 3.10: Overview over the TBvar3D home page. The user can either enter their own variants which will be analysed by the TBvar3D web server, get access to their previous projects or explore variants from the WHO variant catalogue.

Figure 3.11: Overview over the TBvar3D exploration page. The table gives an overview of all 66 antibiotic resistance targets present in the WHO catalogue with information on the drugs targeting it, associated mechanisms of resistance and the number of different types of variants located on the target.

The results page (Figure 3.12) shows the aggregation of variant information, sequence information and structure information.

The sequence panel (depicted in Figure 3.12A) presents a comprehensive overview of the variants, functional annotations, and sequence features of a given protein. The displayed variant data is separated into four categories: user-submitted variants, WHO resistance variants, neutral variants, and uncertain variants. This panel provides information to examine individual variants or groups of variants within the context of all sequence annotations, including easy access to relevant information such as proximity to interaction sites or regions of conservation. The bottom of the sequence panel displays a list of structures associated with the current protein target and also allows for the inspection of surface accessibility and transmembrane prediction features linked to the respective structure. If multiple structures are available, the structure display for the current protein target can be switched here.

The variant panel (Figure 3.12B) presents the results linked to a specific variant selected in the sequence panel. This includes the chemical distance for single amino acid substitutions and the PROVEAN score. For variants listed in the World Health Organization (WHO) catalogue, the panel also displays the antibiotics to which the variant confers resistance, as well as the original WHO classification and a link to the PubChem database for the respective antibiotic drug.

The structure panel (Figure 3.12C) uses the protein viewer implementation of SWISS-MODEL which is based on the NGL viewer[23] and the PV viewer[24]. In addition to the visualisation functionalities of a protein viewer, the user can map features of the sequence panel (like variants) onto the structure. Computationally predicted protein structure models display a global quality estimation score (pLDDT for AF models[25], QMEANDisCo for homology models[12]) in the structure title and the structure can be coloured by local quality scores.

Selecting the drug (Figure 3.12D) in the variant panel focuses the display on the drug-binding pocket in the structure panel (Figure 3.13). Mapping of variants and annotations can still be performed in the ligand-centric view. The size of the ligand environment around the drug can be extended up to 10 Ångstroms.

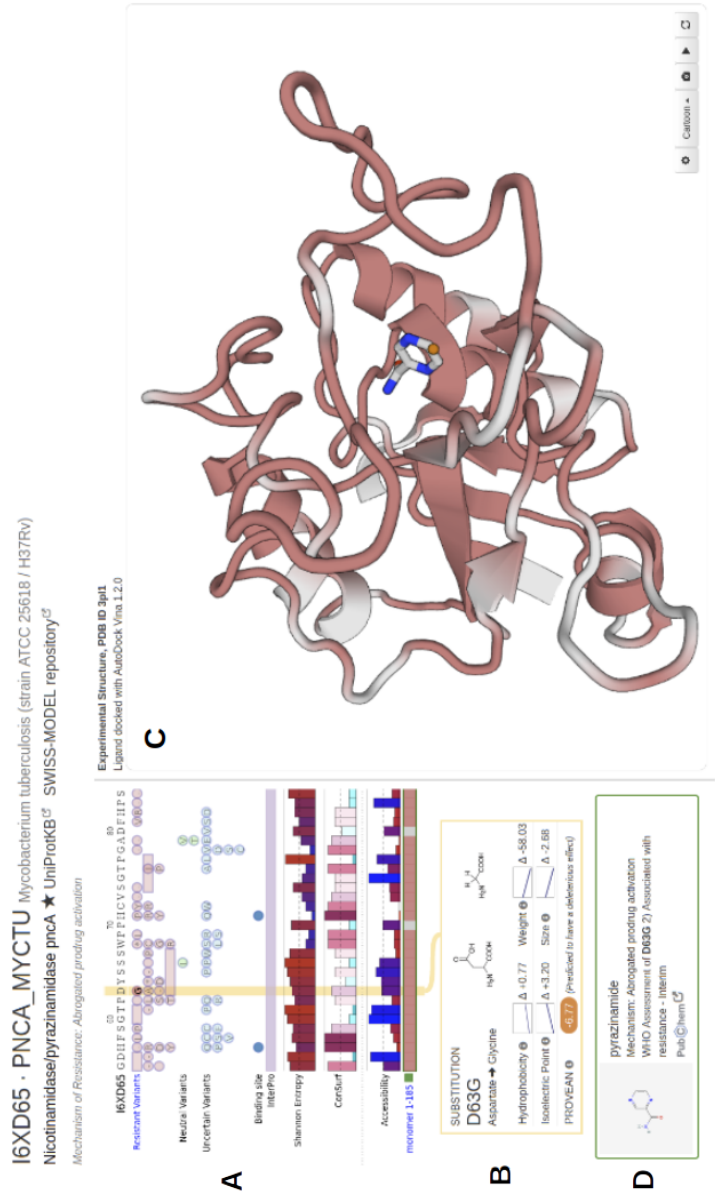

Figure 3.12: Overview of the TBvar3D results page. (A) sequence panel with the variant data, sequence features and structural features aligned to the reference sequence. (B) The variant panel features directly related to the variant are shown here. The box below the variant panel indicates the antibiotics which are present in the currently selected structure view and shows variant annotations from the WHO mutation catalogue. (C) Protein structure viewer, through the sequence panel the mapping of various features can be performed on the protein structure view. (D) The ligand panel shows information on the antibiotic, by clicking on it the structure view is centred on the ligand.

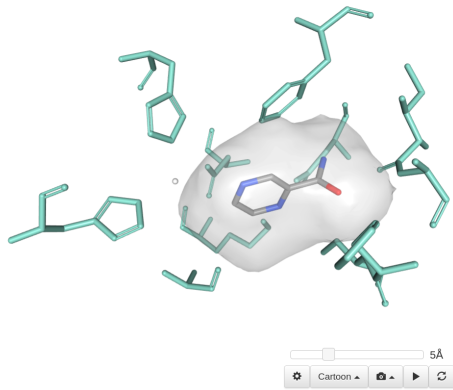

Figure 3.13: The structure view upon clicking on the ligand panel. Now the antibiotic ligand is centred in the structure view. Features from the sequence panel can be mapped on the ligand binding site.

3.2.4 Case Study: Investigation of Bedaquiline-Resistant Variants on Siderophore Exporter MmpL5

Bedaquiline (BDQ) is a novel drug which was approved for the treatment of drug-resistant MTB in 2012, the first of its kind in 40 years[26]. It was classified as one of three core drugs for treating rifampicin-resistant MTB in 2018 by the WHO. With the increase in BDQ use, cases of treatment failures were reported soon after its introduction[27–29].

The WHO catalogue lists 6 resistance genes for BDQ. The first is the drug target, the ATP synthase subunit c atpE (Figure 3.2 in the Methods section) which forms a 9-mer alpha-helical transmembrane barrel in the ATP synthase heterocomplex. The next three targets which are attributed clinical relevance are three genes of the siderophore exporter MmpL5 together with the corresponding accessory protein MmpS5 and the transcriptional regulator MmpR5[30]. The last two candidates, the probable cytoplasmic peptidase PepQ and the uncharacterized transporter Rv1979c have only limited evidence linking them to BDQ resistance[31].

Understanding BDQ resistance is currently a crucial objective in the fight against MTB drug resistance. With the increased use of the drug, resistance variants associated with BDQ are expected to become more prevalent. If it would be possible to identify variants whose phenotype can be linked to clinically relevant drug resistance, BDQ resistance could

be diagnosed similarly to how Rapid Molecular Assays (see Chapter 1.1.2) are currently used to predict Rifampicin and Isoniazid resistance in patient samples today[32]. This requires the proper prioritisation of mutation candidates for further research. TBvar3D can be used to speed up the investigation of emerging variants in the context of their protein structures and to distinguish potentially impactful variants from less remarkable mutations.

In the case study we describe here[33], 291 isolates from South African BDQ-naive patients were screened for naturally occurring BDQ resistance by measuring the BDQ MIC and associating mutations in the 6 candidate genes to isolates with large BDQ MIC shifts.

The number of naturally occurring BDQ-resistant isolates was low (2/291 were BDQ resistant) and two variants in the siderophore exporter MmpL5 were associated with BDQ resistance: p.Thr794Ile and p.Asp767Asn. These associations are contradicted by the WHO catalogue which grades them both as susceptible to BDQ.

The extracellular transmembrane protein MmpL5 is required for the export and synthesis of siderophores which are small, high-affinity iron-binding compounds which help the organism to accumulate iron[34]. The overexpression of MmpL5 is described to mediate non-target-based resistance to azoles, clofazimine and BDQ[35–37].

In this study we used an AF2 model of MmpL5. The protein is predicted to be a transmembrane protein. The mutations are located at the beginning and end of a transmembrane alpha helix.

The mutation site of the first variant p.Thr794Ile (Figure 3.14) is slightly conserved with a Shannon entropy of 0.71 and a ConSurf score of 6. The site is rather buried with a relative solvent accessibility of 30.85%. The mutation site is predicted to be immersed in the transmembrane region. The mutation from a threonine to a more hydrophobic but still similar amino acid isoleucine should not be problematic in the expected hydrophobic environment. The PROVEAN score is far above the deleteriousness threshold with 3.65 denoting that the mutation is predicted to be well-tolerated by the protein. The data in TBvar3D supports the assessment of this variant made by the WHO.

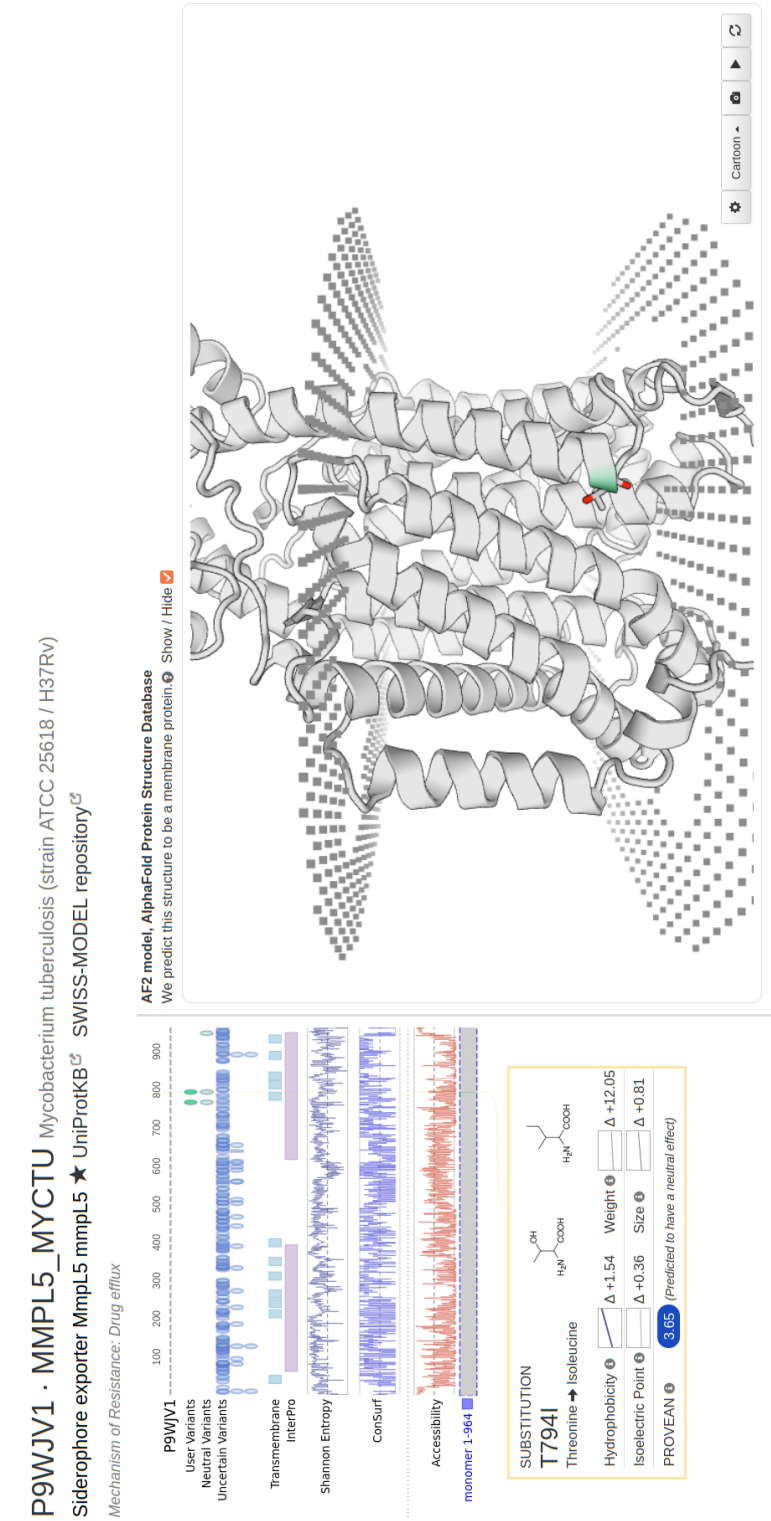

Figure 3.14: Close look at p.Thr794Ile (green) in TBvar3D

The other mutation p.Asp767Asn (Figure 3.15) on the other hand is on a strongly conserved site (Shannon 0.15, ConSurf 9). The residue site itself is also buried with a relative solvent accessibility of 20.65%. The residue site itself is predicted to be located outside the membrane region. The transition from an aspartate to an asparagine introduces a large difference in the isoelectric point of the amino acid (difference of 2.64) while every other physicochemical property remains very similar. The PROVEAN score predicts this mutation to be deleterious (-4.9). The WHO catalogue reports a different mutation with the "uncertain" grading at the same position, p.Asp767Ala, whose PROVEAN score also designates this to be a deleterious mutation (-7.9).

The first mutation p.Thr794Ile is likely not impactful from a structural perspective. No annotation besides the relative conservation of the region is remarkable. The mutation to another hydrophobic amino acid is unlikely to impact the protein structure. The mutation p.Asp767Asn however might have a potential impact on MmpL5: the mutation site is quite conserved with the PROVEAN score grading the mutation as having a deleterious effect. The site itself is rather buried. The mutation changes the isoelectric point of the site. The mutation p.Asp767Ala is also located at the same site and is predicted to be even more deleterious than p.Asp767Asn.

It has been shown for azoles[37] that MmpL5 is linked to active transport of the compound out of the cell, with a similar mechanism being postulated for Clofazimine and Bedaquiline. The mutations on p.Asp767 could potentially be linked to this mechanism, but further computational predictions and annotations would be necessary and this falls outside the scope of TBvar3D. Nevertheless, we could demonstrate how p.Asp767Asn is more likely to have an impact on the protein function than p.Thr794Ile given the available information.

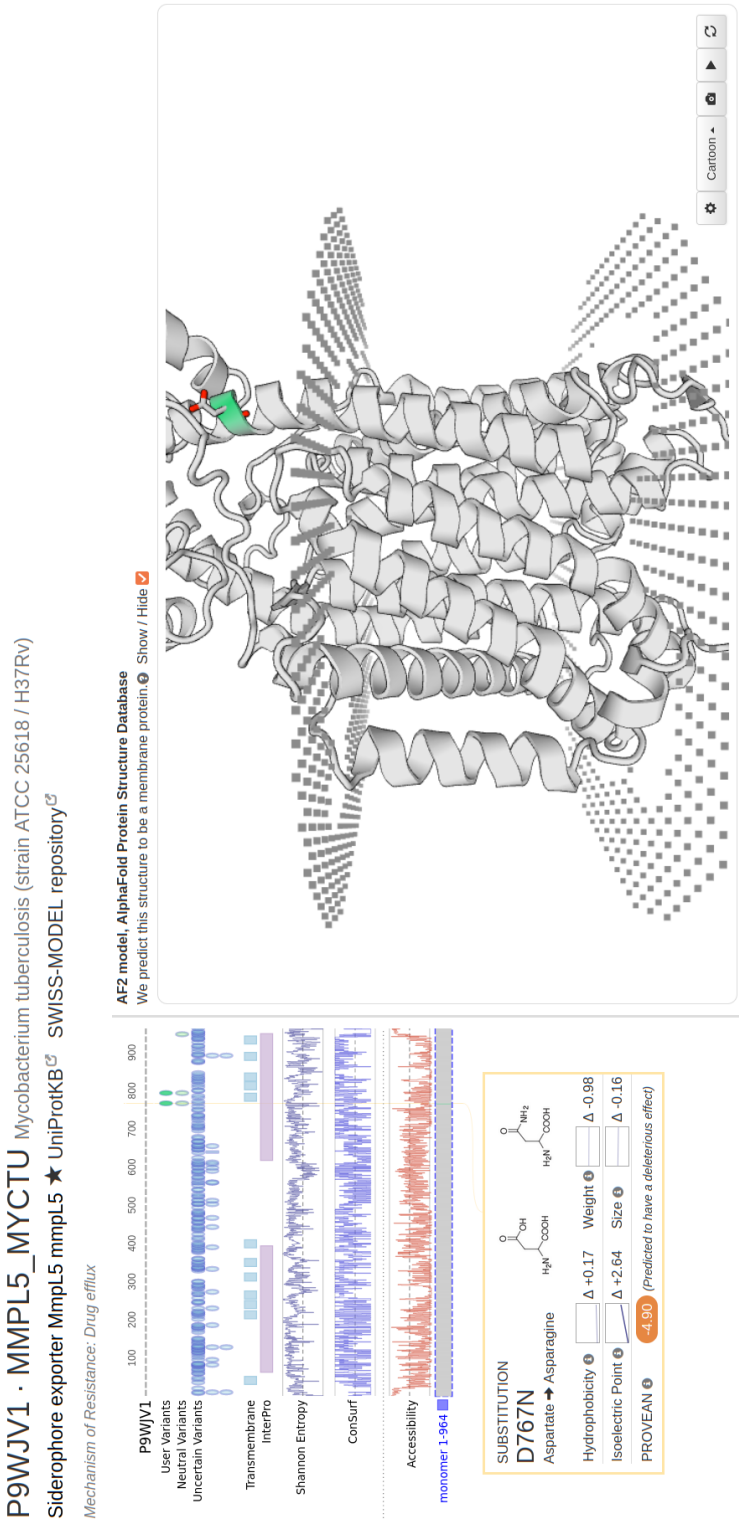

Figure 3.15: Close look at p.Asp767Asn (green) in TBvar3D

3.2.5 Case Study: Compensatory Mutation on ahpE for Isoniazid Resistance

Isoniazid is an antitubercular drug which was first created in 1952[38]. The combination of Isoniazid with Rifampicin, Ethambutol and Pyrazinamide is still used as the standard of care for the treatment of drug-susceptible tuberculosis[39]. The drug resistance mechanism against Isoniazid usually involves deleterious mutations in the prodrug activator catalase-peroxidase katG. The distribution of variants per target in Figure 3.5 showed the large number of known resistance variants in katG which includes many impactful mutations like frameshifts and indels.

Mutations whose phenotype is not linked directly to the resistance mechanisms can also be quite relevant for the study of drug resistance. Borell et al.[40] show the importance of two other phenotypes of bacterial strains for the successful propagation of drug-resistant MTB strains: their virulence and their relative fitness.

Fitness describes the ability of an organism to reproduce successfully. Drug resistance mutations are known to decrease the relative fitness when compared to their wild type[41]. For a resistant MTB cell to thrive, it needs to further obtain mutations which will increase its relative fitness while still maintaining the drug resistance phenotype. The alleviation of the handicap caused by resistance mutations allows the phenotype of drug resistance to increase their reproduction success rate and to be more likely to successfully spread from one patient to another. This makes it important to understand how the mechanism of fitness functions and to identify mutations that increase fitness significantly.

In this case study we show how TBvar3D can also be used to profile variants the resistance targets in the WHO mutation catalogue. Here we characterise the potential compensatory mutation for Isoniazid resistance p.Pro135Gln located in the Alkyl hydroperoxide reductase E enzyme (ahpE). Mutations in this enzyme were described to mitigate the initial fitness cost caused by katG mutations[42].

The result page (Figure 3.16) shows a SWISS-MODEL homo-2-mer homology model with an average model confidence of 0.93 (QMEANDisCo score). The mutation site is a highly exposed non-conserved proline localised on the tip of a protein loop. The RSA is 120.28% (improper refer-

ence values for normalisation can lead to RSA values above 100%[22]) and the ConSurf score and relative entropy score are 3 and 0.85 respectively. The mutation to glutamine would introduce an amino acid with profound physicochemical differences from the wild type. The PROVEAN impact score (-5.04) is also rating the mutation to be deleterious.

The mutation site is located in proximity to the active site of the protein (Figure 3.17). The site p.Cys45 is annotated as a conserved redox-active cysteine residue which performs the nucleophilic attack on its peroxide substrate. The mutation site lies on the binding site of the substrate[43].

The annotations in TBvar3D show that the mutation might impact the catalytic reaction of the enzyme. Increased efficacy of the hydroperoxide reductase might decrease the concentration of radicals which would be beneficial for the cell. A definite characterization would need further research. The information aggregated by TBvar3D suggests that further investigation of this mutation could be worthwhile and provides testable hypotheses for experimental validation.

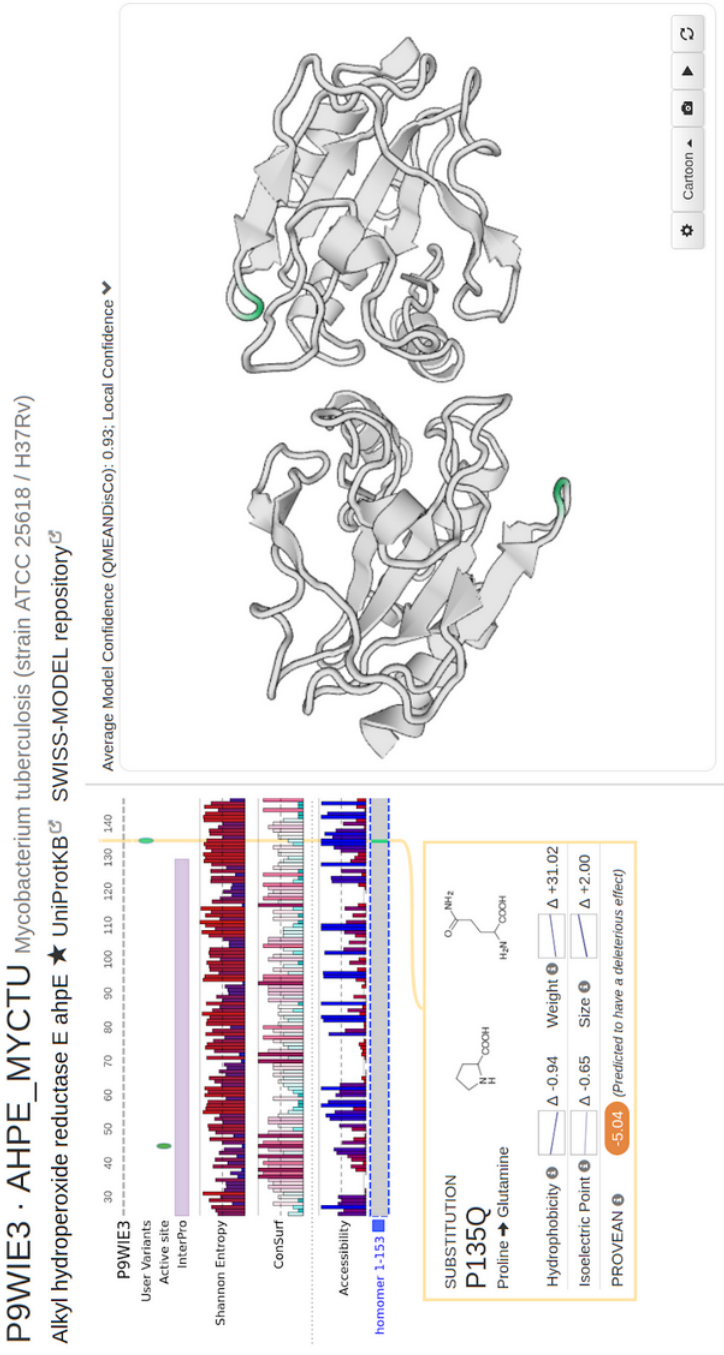

Figure 3.16: Results page for the protein ahpE and the compensatory mutation p.Pro135Gln

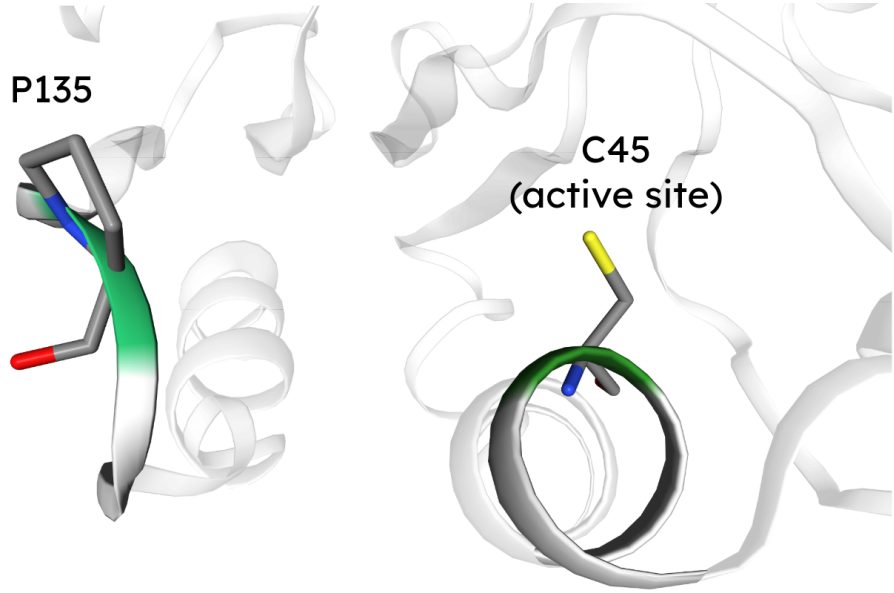

Figure 3.17: The mutation site of p.Pro135Gln is located close to the active site of ahpE, the redox-active cysteine residue p.Cys45.

3.3 Discussion

With TBvar3D, we created a web server for mapping and analysing Mycobacterium tuberculosis variants in the context of protein structures and the comprehensive resistance variant catalogue provided by the WHO. The web server is incorporated in the SWISS-MODEL technology stack which ensures a rigorous update cycle and high-quality protein structures and biologically relevant 3D models, providing the user with an up-to-date web-based environment that streamlines data integration, analysis and hypothesis generation on the role of a given variant or set of variants of interest in drug resistance.

The combined display of information helps to identify and distinguish potentially high-impact variants through a combination of annotations on a sequence, structure and variant level and the visual inspection of the mutation site on protein structure models. TBvar3D streamlines an otherwise time-intensive process of manual generation of annotations and the mapping and display of variants on structure models for a broad audience of MTB scientists and makes it a valuable tool to assist the user to formulate compelling hypotheses on the impact of variants across

the MTB proteome.

3.3.1 Limitations

Due to performance limitations, certain time-consuming annotations are not suitable for automated large-scale analyses within the TBvar3D pipeline. For example, free energy calculations based on molecular dynamics simulations to estimate stability change upon mutations and calculations of binding affinity changes would be valuable to characterise drug resistance variants. Attempts to include such approaches within this work were not successful due to challenges in automating such workflows.

TBvar3D does not provide structure predictions of the mutated protein. To our knowledge, reliable methods to model structural changes upon mutations are not available at the moment.

The curated structure database in its current form will require continuous maintenance to reflect the current knowledge about the target proteins. Newly released experimental structures will have to be integrated for proteins which are currently represented by predictions. While advancements in protein structure prediction and oligomeric predictions may allow to automatise the generation of the structure itself, predictions of ligand-target complexes are not yet fully reliable and will need curation.

Since TBvar3D focuses on the integration of variant data and protein structures, genetic resistance variants lying on non-protein-coding regions are not represented and not analysed in TBvar3D. The shift of genetic expressions induced by genetic mutations lying on promoter regions is a crucial mechanism of resistance in MTB and other bacteria but the proper analysis of these mutations lies out of the scope of the TBvar3D web server.

3.3.2 Future Work

There are three promising avenues for the further development of TBvar3D: i) expansion to encompass a wider range of pathogens, ii) the integration of more sophisticated tools for variant analysis, and iii) tools for supporting the interpretation and predicting phenotypic effects.

TBvar3D could be expanded to include a wider spectrum of resistance

variants, e.g. pathogens belonging to the clinically highly relevant ESKAPE group (*Enterococcus faecium*, *Staphylococcus aureus*, *Klebsiella pneumoniae*, *Acinetobacter baumannii*, *Pseudomonas aeruginosa*, and various *Enterobacter* species), which have been also identified as a significant global health threat in the context of drug resistance[44]. Additionally, the application of TBvar3D's structure-based variant interpretation can be extended to other areas of interest, such as human cancer mutations[45], and viral strains such as SARS-CoV-2.

The development and/or incorporation of new tools to estimate the stability change and drug binding affinity change upon mutation would be valuable additions to the repertoire of TBvar3D and should greatly enhance its capabilities.

The modular buildup of the Var3D pipeline also enables an effortless expansion and adaption of the variant analysis pipeline. With the advent of neural networks in bioinformatics, methods for the estimation of stability change upon mutation and the binding affinity change estimation upon mutations which have a low computational runtime and are reliable seem to be within reach. These new methodologies can be made available to any researcher worldwide without any prior computational knowledge thanks to TBvar3D.

The development or incorporation of new tools to estimate the stability change and drug binding affinity change upon mutation would be valuable additions to the repertoire of TBvar3D and would greatly enhance its capabilities.

Finally, a larger body of data for a broad spectrum of drug resistance in a variety of organisms allows for the development of tools which support the interpretation and prediction of phenotypic effects. Methods based on artificial intelligence can be tailored to specific resistance mechanisms and could be applied to predict the molecular phenotype of variants. Network approaches such as boolean network modelling could be explored to integrate data of individual protein variants into the prediction of a global phenotype. In our case we could explain which antibiotics are likely to be effective on a pathogen, given a list of variants in a clinical isolate. The TBvar3D system today relies on some knowledge of protein structure and function for efficient use. Future versions which include automated artificial intelligence-based interpretation functionalities might find their

use in a clinical setting to translate biological research findings to support treatment decisions for antibiotics.

3.4 Supplementary

Table 3.1: TBvar3D Structure Database

Gene	Drug name	Method structure	Method ligand	Oligomeric state	Associated Mechanism of Resistance	Ref
katG	Isoniazid	Exp	Trans	homo-2-mer	abrogated_prodrug_activation	46
rplC	Linezolid	Exp	Sup	ribosomal protein	drug_target_alteration	47
rpsJ	Linezolid	Exp		ribosomal protein	unclear	48
aftA	Ethambutol	AFDB		monomer	unclear	49
fbiB	Delamanid	Exp		homo-2-mer	abrogated_prodrug_activation	47
embA	Ethambutol	Exp	Exp	hetero-1-1-1-mer	drug_target_alteration	50
glf	Ethambutol	Exp		homo-2-mer	unclear	46
Rv3237c	Pyrazinamide	AFDB		monomer	unclear	51
embR	Ethambutol	Exp		monomer	gene_regulation	52
eis	Amikacin	Exp		homo-6-mer	abrogation_of drug_inactivating_enzyme	53
eis	Kanamycin	Exp		homo-6-mer	abrogation_of drug_inactivating_enzyme	53
tap	Isoniazid	AFDB		monomer	drug_efflux	54
tap	Streptomycin	AFDB		monomer	drug_efflux	54
tap	Pyrazinamide	AFDB		monomer	drug_efflux	54
Rv3806c	Amikacin	AFDB		monomer	unclear	55
Rv3806c	Capreomycin	AFDB		monomer	unclear	55
Rv3806c	Ethambutol	AFDB		monomer	unclear	55
Rv3788	Ethambutol	AFDB		monomer	unclear	56
ahpC	Isoniazid	Exp		homo-12-mer	unclear	57
rpoB	Rifampicin	Exp	Exp	hetero-2-1-1-1-1-mer	drug_target_alteration	47
mabA	Isoniazid	Exp		homo-4-mer	unclear	58
mabA	Ethionamide	Exp		homo-4-mer	unclear	58
gyrB	Moxifloxacin	AFDB	Trans	hetero-2-2-mer	drug_target_alteration	47
gyrB	Levofloxacin	AFDB	Trans	hetero-2-2-mer	drug_target_alteration	47
Rv1979c	Clofazimine	AFDB		monomer	drug_efflux	59
Rv1979c	Bedaquiline	AFDB		monomer	drug_efflux	59
whiB6	Amikacin	AFDB		monomer	gene_regulation	60
whiB6	Capreomycin	AFDB		monomer	gene_regulation	60
whiB6	Streptomycin	AFDB		monomer	gene_regulation	60
Rv0681	Streptomycin	AlphaFold-Multimer		homo-2-mer	gene_regulation	61
pncA	Pyrazinamide	Exp	Dock	monomer	abrogated_prodrug_activation	62
fgd1	Delamanid	Exp		homo-2-mer	abrogated_prodrug_activation	47
furA	Isoniazid	AlphaFold-Multimer		homo-2-mer	gene_regulation	63
mpt64	Clofazimine	Exp		monomer	unclear	
mpt64	Bedaquiline	Exp		monomer	unclear	
fprA	Amikacin	Exp		homo-2-mer	unclear	64
fprA	Capreomycin	Exp		homo-2-mer	unclear	64
mymA	Ethionamide	AFDB		monomer	unclear	65
panD	Pyrazinamide	Exp	Exp	hetero-4-4-mer	drug_target_alteration	66
PPE35	Pyrazinamide	AFDB		monomer	unclear	67
inhA	Isoniazid	Exp	Exp	homo-4-mer	drug_target_alteration	46
inhA	Ethionamide	Exp	Exp	homo-4-mer	drug_target_alteration	46
embC	Ethambutol	AFDB	Trans	homo-2-mer	drug_target_alteration	50
rpsD	Rifampicin	Exp		ribosomal protein	unclear	68
fbiD	Delamanid	Exp		monomer	unclear	69
Rv1692	Capreomycin	Exp		homo-2-mer	unclear	70
rsmG	Streptomycin	Exp		monomer	abrogation_of drug_target_methylation	71
ndh	Isoniazid	AlphaFold-Multimer		homo-2-mer	overabundance_of drug_target_substrate	46
ndh	Ethionamide	AlphaFold-Multimer		homo-2-mer	overabundance_of drug_target_substrate	46
rpoA	Rifampicin	Exp	Exp	hetero-2-1-1-1-1-mer	unclear	58
fbiA	Delamanid	AFDB		monomer	unclear	47
ethA	Ethionamide	AFDB	Dock	monomer	abrogated_prodrug_activation	46

Rv3789	Ethambutol	AFDB		monomer	unclear	72
prfB	Amikacin	AFDB		monomer	unclear	73
prfB	Capreomycin	AFDB		monomer	unclear	73
rnj	Rifampicin	AlphaFold-Multimer		homo-4-mer	unclear	74
rnj	Isoniazid	AlphaFold-Multimer		homo-4-mer	unclear	74
ddn	Delamanid	AFDB	Dock	monomer	abrogated_prodrug_activation	47
fbiC	Delamanid	AFDB		monomer	abrogated_prodrug_activation	47
embB	Ethambutol	Exp	Exp	hetero-1-1-1-mer	drug_target_alteration	47
murA	Linezolid	AlphaFold-Multimer		homo-4-mer	unclear	75
murA	Capreomycin	AlphaFold-Multimer		homo-4-mer	unclear	75
murA	Streptomycin	AlphaFold-Multimer		homo-4-mer	unclear	75
murA	Amikacin	AlphaFold-Multimer		homo-4-mer	unclear	75
murA	Kanamycin	AlphaFold-Multimer		homo-4-mer	unclear	75
Rv3236c	Pyrazinamide	AlphaFold-Multimer		homo-2-mer	drug_efflux	67
whiB7	Amikacin	Exp		hetero-2-1-1-1-1-1-1-1-mer	unclear	76
whiB7	Kanamycin	Exp		hetero-2-1-1-1-1-1-1-1-mer	gene_regulation	76
whiB7	Streptomycin	Exp		hetero-2-1-1-1-1-1-1-1-mer	gene_regulation	76
aftB	Amikacin	AFDB		monomer	unclear	67
aftB	Capreomycin	AFDB		monomer	unclear	67
Rv2044c	Pyrazinamide	AFDB		monomer	unclear	77
ccsA	Amikacin	AFDB		monomer	unclear	67
ccsA	Capreomycin	AFDB		monomer	unclear	67
rpoC	Rifampicin	Exp	Exp	hetero-2-1-1-1-1-mer	unclear	57
mshA	Isoniazid	AlphaFold-Multimer		homo-2-mer	abrogated_prodrug_activation	78
mshA	Ethionamide	AlphaFold-Multimer		homo-2-mer	abrogated_prodrug_activation	78
mmpL5	Clofazimine	AFDB		monomer	drug_efflux	37
mmpL5	Bedaquiline	AFDB		monomer	drug_efflux	37
mmpS5	Clofazimine	AFDB		monomer	drug_efflux	37
mmpS5	Bedaquiline	AFDB		monomer	drug_efflux	37
mmpR5	Clofazimine	Exp		homo-4-mer	gene_regulation	37
mmpR5	Bedaquiline	Exp		homo-4-mer	gene_regulation	37
tlyA	Capreomycin	AFDB		monomer	abrogation_of drug_target_methylation	47
Rv0528	Amikacin	AFDB		monomer	unclear	
Rv0528	Capreomycin	AFDB		monomer	unclear	
gyrA	Moxifloxacin	AFDB	Trans	hetero-2-2-mer	drug_target_alteration	47
gyrA	Levofloxacin	AFDB	Trans	hetero-2-2-mer	drug_target_alteration	47
ethR	Ethionamide	Exp		homo-2-mer	gene_regulation	57
atpB	Bedaquiline	Hom	SM	hetero-1-3-1-1-1-1-1-9-mer	drug_target_alteration	79
clpC1	Pyrazinamide	Hom		homo-6-mer	unclear	66
Rv0485	Isoniazid	AlphaFold-Multimer		homo-2-mer	gene_regulation	80
Rv0485	Ethionamide	AlphaFold-Multimer		homo-2-mer	gene_regulation	80
atpE	Bedaquiline	Hom	SM	hetero-1-3-1-1-1-1-1-9-mer	drug_target_alteration	47
rpsL	Streptomycin	Exp		ribosomal protein	unclear	47
pepQ	Clofazimine	AlphaFold-Multimer		homo-2-mer	unclear	81
pepQ	Bedaquiline	AlphaFold-Multimer		homo-2-mer	unclear	81
Rv1693	Capreomycin	AFDB		monomer	unclear	

References

[1] Church, N. A., & McKillip, J. L. (2021). Antibiotic resistance crisis: Challenges and imperatives. *Biologia*, *76*, 1535–1550.

[2] World Health Organization. Global antimicrobial resistance surveillance system (glass) report: early implementation 2017-2018.

[3] Walker, T. M., Miotto, P., Köser, C. U., Fowler, P. W., Knaggs, J., Iqbal, Z., Hunt, M., Chindelevitch, L., Farhat, M. R., Cirillo, D. M. et al. (2022). The 2021 who catalogue of mycobacterium tuberculosis complex mutations associated with drug resistance: a genotypic analysis. *The Lancet Microbe*, *3*, e265–e273.

[4] Kapopoulou, A., Lew, J. M., & Cole, S. T. (2011). The mycobrowser portal: a comprehensive and manually annotated resource for mycobacterial genomes. *Tuberculosis*, *91*, 8–13.

[5] Biasini, M., Schmidt, T., Bienert, S., Mariani, V., Studer, G., Haas, J., Johner, N., Schenk, A. D., Philippsen, A., & Schwede, T. (2013). Openstructure: an integrated software framework for computational structural biology. *Acta Crystallographica Section D: Biological Crystallography*, *69*, 701–709.

[6] Cole, S., & Barrell, B. (1998). Analysis of the genome of mycobacterium tuberculosis h37rv. In *Genetics and Tuberculosis: Novartis Foundation Symposium 217* (pp. 160–177). Wiley Online Library.

[7] Bienert, S., Waterhouse, A., De Beer, T. A., Tauriello, G., Studer, G., Bordoli, L., & Schwede, T. (2017). The swiss-model repository—new features and functionality. *Nucleic acids research*, *45*, D313–D319.

[8] Bertoni, M., Kiefer, F., Biasini, M., Bordoli, L., & Schwede, T. (2017). Modeling protein quaternary structure of homo-and hetero-oligomers beyond binary interactions by homology. *Scientific reports*, *7*, 1–15.

[9] Varadi, M., Anyango, S., Deshpande, M., Nair, S., Natassia, C., Yordanova, G., Yuan, D., Stroe, O., Wood, G., Laydon, A. et al. (2022). Alphafold protein structure database: massively expanding the structural coverage of protein-sequence space with high-accuracy models. *Nucleic acids research*, *50*, D439–D444.

[10] Evans, R., O'Neill, M., Pritzel, A., Antropova, N., Senior, A., Green, T., Žídek, A., Bates, R., Blackwell, S., Yim, J. et al. (2022). Protein complex prediction with alphafold-multimer. *BioRxiv*, (pp. 2021–10).

[11] Waterhouse, A., Bertoni, M., Bienert, S., Studer, G., Tauriello, G., Gumienny, R., Heer, F. T., de Beer, T. A. P., Rempfer, C., Bordoli, L. et al. (2018). Swiss-model: homology modelling of protein structures and complexes. *Nucleic acids research*, *46*, W296–W303.

[12] Studer, G., Rempfer, C., Waterhouse, A. M., Gumienny, R., Haas, J., & Schwede, T. (2020). Qmeandisco—distance constraints applied on model quality estimation. *Bioinformatics*, *36*, 1765–1771.

[13] Blower, T. R., Williamson, B. H., Kerns, R. J., & Berger, J. M. (2016). Crystal structure and stability of gyrase–fluoroquinolone cleaved complexes from mycobacterium tuberculosis. *Proceedings of the National Academy of Sciences*, *113*, 1706–1713.

[14] Pettersen, E. F., Goddard, T. D., Huang, C. C., Couch, G. S., Greenblatt, D. M., Meng, E. C., & Ferrin, T. E. (2004). Ucsf chimera—a visualization system for exploratory research and analysis. *Journal of computational chemistry*, *25*, 1605–1612.

[15] Yang, K., Chang, J.-Y., Cui, Z., Li, X., Meng, R., Duan, L., Thongchol, J., Jakana, J., Huwe, C. M., Sacchettini, J. C. et al. (2017). Structural insights into species-specific features of the ribosome from the human pathogen mycobacterium tuberculosis. *Nucleic acids research*, *45*, 10884–10894.

[16] Kim, S., Chen, J., Cheng, T., Gindulyte, A., He, J., He, S., Li, Q., Shoemaker, B. A., Thiessen, P. A., Yu, B. et al. (2019). Pubchem 2019 update: improved access to chemical data. *Nucleic acids research*, *47*, D1102–D1109.

[17] Cheeseright, T., Mackey, M., Rose, S., & Vinter, A. (2006). Molecular field extrema as descriptors of biological activity: definition and validation. *Journal of chemical information and modeling*, *46*, 665–676.

[18] Eberhardt, J., Santos-Martins, D., Tillack, A. F., & Forli, S. (2021). Autodock vina 1.2. 0: New docking methods, expanded force field, and python bindings. *Journal of Chemical Information and Modeling*, *61*, 3891–3898.

[19] Fischer, A., Smiesko, M., Sellner, M., & Lill, M. A. (2021). Decision making in structure-based drug discovery: visual inspection of docking results. *Journal of Medicinal Chemistry*, *64*, 2489–2500.

[20] Holcomb, M., Chang, Y.-T., Goodsell, D. S., & Forli, S. (2022). Evaluation of alphafold2 structures as docking targets. *Protein Science*, (p. e4530).

[21] Scardino, V., Di Filippo, J. I., & Cavasotto, C. N. (2022). How good are alphafold models for docking-based virtual screening? *iScience*, (p. 105920).

[22] Tien, M. Z., Meyer, A. G., Sydykova, D. K., Spielman, S. J., & Wilke, C. O. (2013). Maximum allowed solvent accessibilites of residues in proteins. *PloS one*, *8*, e80635.

[23] Rose, A. S., Bradley, A. R., Valasatava, Y., Duarte, J. M., Prlić, A., & Rose, P. W. (2016). Web-based molecular graphics for large complexes. In *Proceedings of the 21st international conference on Web3D technology* (pp. 185–186).

[24] Biasini, M. (2014). Pv-webgl-based protein viewer. *Zenodo*, .

[25] Mariani, V., Biasini, M., Barbato, A., & Schwede, T. (2013). lddt: a local superposition-free score for comparing protein structures and models using distance difference tests. *Bioinformatics*, *29*, 2722–2728.

[26] Mahajan, R. Bedaquiline: first fda-approved tuberculosis drug in 40 years.

[27] Zimenkov, D. V., Nosova, E. Y., Kulagina, E. V., Antonova, O. V., Arslanbaeva, L. R., Isakova, A. I., Krylova, L. Y., Peretokina, I. V., Makarova, M. V., Safonova, S. G. et al. (2017). Examination of bedaquiline-and linezolid-resistant mycobacterium tuberculosis isolates from the moscow region. *Journal of Antimicrobial Chemotherapy*, *72*, 1901–1906.

[28] Somoskovi, A., Bruderer, V., Hömke, R., Bloemberg, G. V., & Böttger, E. C. (2015). A mutation associated with clofazimine and bedaquiline cross-resistance in mdr-tb following bedaquiline treatment. *European Respiratory Journal*, *45*, 554–557.

[29] Bloemberg, G. V., Keller, P. M., Stucki, D., Trauner, A., Borrell, S., Latshang, T., Coscolla, M., Rothe, T., Hömke, R., Ritter, C. et al. (2015). Acquired re-

sistance to bedaquiline and delamanid in therapy for tuberculosis. *New England Journal of Medicine*, *373*, 1986–1988.

[30] Briffotaux, J., Huang, W., Wang, X., & Gicquel, B. (2017). Mmps5/mmpl5 as an efflux pump in mycobacterium species. *Tuberculosis*, *107*, 13–19.

[31] Ismail, N., Rivière, E., Limberis, J., Huo, S., Metcalfe, J. Z., Warren, R. M., & Van Rie, A. (2021). Genetic variants and their association with phenotypic resistance to bedaquiline in mycobacterium tuberculosis: a systematic review and individual isolate data analysis. *The Lancet Microbe*, *2*, e604–e616.

[32] Eddabra, R., & Ait Benhassou, H. (2018). Rapid molecular assays for detection of tuberculosis. *Pneumonia*, *10*, 1–12.

[33] Rivière, E., Verboven, L., Dippenaar, A., Goossens, S., De Vos, E., Streicher, E., Cuypers, B., Laukens, K., Ben-Rached, F., Rodwell, T. C. et al. (2022). Variants in bedaquiline-candidate-resistance genes: prevalence in bedaquiline-naive patients, effect on mic, and association with mycobacterium tuberculosis lineage. *Antimicrobial Agents and Chemotherapy*, *66*, e00322–22.

[34] Wells, R. M., Jones, C. M., Xi, Z., Speer, A., Danilchanka, O., Doornbos, K. S., Sun, P., Wu, F., Tian, C., & Niederweis, M. (2013). Discovery of a siderophore export system essential for virulence of mycobacterium tuberculosis. *PLoS pathogens*, *9*, e1003120.

[35] Hartkoorn, R. C., Uplekar, S., & Cole, S. T. (2014). Cross-resistance between clofazimine and bedaquiline through upregulation of mmpl5 in mycobacterium tuberculosis. *Antimicrobial agents and chemotherapy*, *58*, 2979–2981.

[36] Andries, K., Villellas, C., Coeck, N., Thys, K., Gevers, T., Vranckx, L., Lounis, N., de Jong, B. C., & Koul, A. (2014). Acquired resistance of mycobacterium tuberculosis to bedaquiline. *PLOS one*, *9*, e102135.

[37] Milano, A., Pasca, M. R., Provvedi, R., Lucarelli, A. P., Manina, G., Ribeiro, A. L. d. J. L., Manganelli, R., & Riccardi, G. (2009). Azole resistance in mycobacterium tuberculosis is mediated by the mmps5–mmpl5 efflux system. *Tuberculosis*, *89*, 84–90.

[38] Walker, S., & Walker, S. R. (1988). *Trends and changes in drug research and development*. Springer.

[39] Stuart, M. C., Kouimtzi, M., & Hill, S. R. (2009). *WHO model formulary 2008*. World Health Organization.

[40] Borrell, S., & Gagneux, S. (2009). Infectiousness, reproductive fitness and evolution of drug-resistant mycobacterium tuberculosis [state of the art]. *The International Journal of Tuberculosis and Lung Disease*, *13*, 1456–1466.

[41] Gagneux, S. (2009). Fitness cost of drug resistance in mycobacterium tuberculosis. *Clinical Microbiology and Infection*, *15*, 66–68.

[42] Rinder, H., Thomschke, A., Rüsch-Gerdes, S., Bretzel, G., Feldmann, K., Rifai, M., & Löscher, T. (1998). Significance ofahpc promoter mutations for the prediction of isoniazid resistance inmycobacterium tuberculosis. *European Journal of Clinical Microbiology and Infectious Diseases*, *17*, 508–511.

[43] Kumar, A., Balakrishna, A. M., Nartey, W., Manimekalai, M. S. S., & Grüber, G. (2016). Redox chemistry of mycobacterium tuberculosis alkylhydroperoxide reductase e (ahpe): structural and mechanistic insight into a mycoredoxin-1

independent reductive pathway of ahpe via mycothiol. *Free Radical Biology and Medicine*, *97*, 588–601.

[44] Murray, C. J., Ikuta, K. S., Sharara, F., Swetschinski, L., Aguilar, G. R., Gray, A., Han, C., Bisignano, C., Rao, P., Wool, E. et al. (2022). Global burden of bacterial antimicrobial resistance in 2019: a systematic analysis. *The Lancet*, *399*, 629–655.

[45] Krebs, F. S., Zoete, V., Trottet, M., Pouchon, T., Bovigny, C., & Michielin, O. (2021). Swiss-po: a new tool to analyze the impact of mutations on protein three-dimensional structures for precision oncology. *NPJ precision oncology*, *5*, 1–9.

[46] Vilchèze, C., & Jacobs Jr, W. R. (2014). Resistance to isoniazid and ethionamide in mycobacterium tuberculosis: genes, mutations, and causalities. *Microbiology spectrum*, *2*, 2–4.

[47] Gygli, S. M., Borrell, S., Trauner, A., & Gagneux, S. (2017). Antimicrobial resistance in mycobacterium tuberculosis: mechanistic and evolutionary perspectives. *FEMS microbiology reviews*, *41*, 354–373.

[48] Sekyere, J. O., & Asante, J. (2018). Emerging mechanisms of antimicrobial resistance in bacteria and fungi: advances in the era of genomics. *Future microbiology*, *13*, 241–262.

[49] Alderwick, L. J., Seidel, M., Sahm, H., Besra, G. S., & Eggeling, L. (2006). Identification of a novel arabinofuranosyltransferase (afta) involved in cell wall arabinan biosynthesis in mycobacterium tuberculosis. *Journal of Biological Chemistry*, *281*, 15653–15661.

[50] Cui, Z., Li, Y., Cheng, S., Yang, H., Lu, J., Hu, Z., & Ge, B. (2014). Mutations in the embc-emba intergenic region contribute to mycobacterium tuberculosis resistance to ethambutol. *Antimicrobial agents and chemotherapy*, *58*, 6837–6843.

[51] Dubnau, E., Fontán, P., Manganelli, R., Soares-Appel, S., & Smith, I. (2002). Mycobacterium tuberculosis genes induced during infection of human macrophages. *Infection and immunity*, *70*, 2787–2795.

[52] Brossier, F., Sougakoff, W., Bernard, C., Petrou, M., Adeyema, K., Pham, A., Amy de la Breteque, D., Vallet, M., Jarlier, V., Sola, C. et al. (2015). Molecular analysis of the embcab locus and embr gene involved in ethambutol resistance in clinical isolates of mycobacterium tuberculosis in france. *Antimicrobial Agents and Chemotherapy*, *59*, 4800–4808.

[53] Gikalo, M. B., Nosova, E. Y., Krylova, L. Y., & Moroz, A. M. (2012). The role of eis mutations in the development of kanamycin resistance in mycobacterium tuberculosis isolates from the moscow region. *Journal of Antimicrobial Chemotherapy*, *67*, 2107–2109.

[54] Liu, J., Shi, W., Zhang, S., Hao, X., Maslov, D. A., Shur, K. V., Bekker, O. B., Danilenko, V. N., & Zhang, Y. (2019). Mutations in efflux pump rv1258c (tap) cause resistance to pyrazinamide, isoniazid, and streptomycin in m. tuberculosis. *Frontiers in Microbiology*, *10*, 216.

[55] He, L., Wang, X., Cui, P., Jin, J., Chen, J., Zhang, W., & Zhang, Y. (2015). ubia (rv3806c) encoding dppr synthase involved in cell wall synthesis is associated

with ethambutol resistance in mycobacterium tuberculosis. *Tuberculosis*, *95*, 149–154.

[56] Islam, M. M., Hameed, H. A., Mugweru, J., Chhotaray, C., Wang, C., Tan, Y., Liu, J., Li, X., Tan, S., Ojima, I. et al. (2017). Drug resistance mechanisms and novel drug targets for tuberculosis therapy. *Journal of genetics and genomics*, *44*, 21–37.

[57] Khawbung, J. L., Nath, D., & Chakraborty, S. (2021). Drug resistant tuberculosis: a review. *Comparative Immunology, Microbiology and Infectious Diseases*, *74*, 101574.

[58] Veyron-Churlet, R., Zanella-Cléon, I., Cohen-Gonsaud, M., Molle, V., & Kremer, L. (2010). Phosphorylation of the mycobacterium tuberculosis β-ketoacyl-acyl carrier protein reductase maba regulates mycolic acid biosynthesis. *Journal of Biological Chemistry*, *285*, 12714–12725.

[59] Zhang, S., Chen, J., Cui, P., Shi, W., Zhang, W., & Zhang, Y. (2015). Identification of novel mutations associated with clofazimine resistance in mycobacterium tuberculosis. *Journal of Antimicrobial Chemotherapy*, *70*, 2507–2510.

[60] Rodríguez-Castillo, J. G., Pino, C., Niño, L. F., Rozo, J. C., Llerena-Polo, C., Parra-López, C. A., Tauch, A., & Murcia-Aranguren, M. I. (2017). Comparative genomic analysis of mycobacterium tuberculosis beijing-like strains revealed specific genetic variations associated with virulence and drug resistance. *Infection, Genetics and Evolution*, *54*, 314–323.

[61] Zheng, X. (2007). *Serine/threonine phosphorylation in Mycobacterium tuberculosis: substrates of PknH kinase*. Ph.D. thesis University of British Columbia.

[62] Rajendran, V., & Sethumadhavan, R. (2014). Drug resistance mechanism of pnca in mycobacterium tuberculosis. *Journal of Biomolecular Structure and Dynamics*, *32*, 209–221.

[63] Lucarelli, D., Vasil, M. L., Meyer-Klaucke, W., & Pohl, E. (2008). The metal-dependent regulators fura and furb from mycobacterium tuberculosis. *International journal of molecular sciences*, *9*, 1548–1560.

[64] McLEAN, K. J., Scrutton, N. S., & Munro, A. W. (2003). Kinetic, spectroscopic and thermodynamic characterization of the mycobacterium tuberculosis adrenodoxin reductase homologue fpra. *Biochemical Journal*, *372*, 317–327.

[65] Grant, S. S., Wellington, S., Kawate, T., Desjardins, C. A., Silvis, M. R., Wivagg, C., Thompson, M., Gordon, K., Kazyanskaya, E., Nietupski, R. et al. (2016). Baeyer-villiger monooxygenases etha and myma are required for activation of replicating and non-replicating mycobacterium tuberculosis inhibitors. *Cell chemical biology*, *23*, 666–677.

[66] Anthony, R., den Hertog, A., & van Soolingen, D. (2018). 'happy the man, who, studying nature's laws, thro'known effects can trace the secret cause.'do we have enough pieces to solve the pyrazinamide puzzle? *Journal of Antimicrobial Chemotherapy*, *73*, 1750–1754.

[67] Farhat, M. R., Freschi, L., Calderon, R., Ioerger, T., Snyder, M., Meehan, C. J., de Jong, B., Rigouts, L., Sloutsky, A., Kaur, D. et al. (2019). Gwas for quantitative resistance phenotypes in mycobacterium tuberculosis reveals resistance genes and regulatory regions. *Nature communications*, *10*, 1–11.

[68] Rocha, D. M., Viveiros, M., Saraiva, M., & Osório, N. S. (2021). The neglected contribution of streptomycin to the tuberculosis drug resistance problem. *Genes*, *12*, 2003.

[69] Rifat, D., Li, S.-Y., Ioerger, T., Shah, K., Lanoix, J.-P., Lee, J., Bashiri, G., Sacchettini, J., & Nuermberger, E. (2020). Mutations in fbid (rv2983) as a novel determinant of resistance to pretomanid and delamanid in mycobacterium tuberculosis. *Antimicrobial agents and chemotherapy*, *65*, e01948–20.

[70] Larrouy-Maumus, G., Biswas, T., Hunt, D. M., Kelly, G., Tsodikov, O. V., & de Carvalho, L. P. S. (2013). Discovery of a glycerol 3-phosphate phosphatase reveals glycerophospholipid polar head recycling in mycobacterium tuberculosis. *Proceedings of the National Academy of Sciences*, *110*, 11320–11325.

[71] Nishimura, K., Hosaka, T., Tokuyama, S., Okamoto, S., & Ochi, K. (2007). Mutations in rsmg, encoding a 16s rrna methyltransferase, result in low-level streptomycin resistance and antibiotic overproduction in streptomyces coelicolor a3 (2). *Journal of bacteriology*, *189*, 3876–3883.

[72] Kolly, G. S., Mukherjee, R., Kilacsková, E., Abriata, L. A., Raccaud, M., Blaško, J., Sala, C., Dal Peraro, M., Mikušová, K., & Cole, S. T. (2015). Gtra protein rv3789 is required for arabinosylation of arabinogalactan in mycobacterium tuberculosis. *Journal of bacteriology*, *197*, 3686–3697.

[73] Zeng, J., Platig, J., Cheng, T.-Y., Ahmed, S., Skaf, Y., Potluri, L.-P., Schwartz, D., Steen, H., Moody, D. B., & Husson, R. N. (2020). Protein kinases pkna and pknb independently and coordinately regulate essential mycobacterium tuberculosis physiologies and antimicrobial susceptibility. *PLoS Pathogens*, *16*, e1008452.

[74] Martini, M. C., Hicks, N. D., Xiao, J., Alonso, M. N., Barbier, T., Sixsmith, J., Fortune, S. M., & Shell, S. S. (2022). Loss of rnase j leads to multi-drug tolerance and accumulation of highly structured mrna fragments in mycobacterium tuberculosis. *PLoS Pathogens*, *18*, e1010705.

[75] Kavvas, E. S., Catoiu, E., Mih, N., Yurkovich, J. T., Seif, Y., Dillon, N., Heckmann, D., Anand, A., Yang, L., Nizet, V. et al. (2018). Machine learning and structural analysis of mycobacterium tuberculosis pan-genome identifies genetic signatures of antibiotic resistance. *Nature communications*, *9*, 1–9.

[76] Burian, J., Ramon-Garcia, S., Howes, C. G., & Thompson, C. J. (2012). Whib7, a transcriptional activator that coordinates physiology with intrinsic drug resistance in mycobacterium tuberculosis. *Expert review of anti-infective therapy*, *10*, 1037–1047.

[77] Hameed, H. A., Tan, Y., Islam, M. M., Lu, Z., Chhotaray, C., Wang, S., Liu, Z., Fang, C., Tan, S., Yew, W. W. et al. (2020). Detection of novel gene mutations associated with pyrazinamide resistance in multidrug-resistant mycobacterium tuberculosis clinical isolates in southern china. *Infection and Drug Resistance*, *13*, 217.

[78] Feuerriegel, S., Köser, C. U., & Niemann, S. (2014). Phylogenetic polymorphisms in antibiotic resistance genes of the mycobacterium tuberculosis complex. *Journal of Antimicrobial Chemotherapy*, *69*, 1205–1210.

[79] Huitric, E., Verhasselt, P., Koul, A., Andries, K., Hoffner, S., & Andersson, D. I. (2010). Rates and mechanisms of resistance development in mycobacterium

tuberculosis to a novel diarylquinoline atp synthase inhibitor. *Antimicrobial agents and chemotherapy*, *54*, 1022–1028.

[80] Goldstone, R. M., Goonesekera, S. D., Bloom, B. R., & Sampson, S. L. (2009). The transcriptional regulator rv0485 modulates the expression of a pe and ppe gene pair and is required for mycobacterium tuberculosis virulence. *Infection and immunity*, *77*, 4654–4667.

[81] Almeida, D., Ioerger, T., Tyagi, S., Li, S.-Y., Mdluli, K., Andries, K., Grosset, J., Sacchettini, J., & Nuermberger, E. (2016). Mutations in pepq confer low-level resistance to bedaquiline and clofazimine in mycobacterium tuberculosis. *Antimicrobial agents and chemotherapy*, *60*, 4590–4599.

CHAPTER 4

Impact of Natural Polymorphisms in Antibody-Antigen Interfaces

The work described in this chapter has been a collaborative effort between Erblin Asllanaj and Rosalba Lepore.

Contributions: RL designed and supervised the study. EA parsed the structure data and variant data, annotated the therapeutics and implemented the analysis pipeline. EA and RL identified a subset of targets for further experimental characterization (equal contributions). RL organised collaboration to perform experimental validation (will be concluded after submission of this dissertation).

The phenomenon of human polymorphisms on the epitopes of antibody therapeutics which lower their efficacy has been observed in single cases[1,2] but to our knowledge has never been studied comprehensively. With the high relevance antibody therapeutics have for the development of new medical drugs[3] it is important to understand the effect that human diversity might have on their efficacy.

The project described in this chapter aims to identify natural polymorphisms in the interfaces between human drug targets and their respective antibody therapeutics. Using the available structural information of drug target complexes we aim to capture the extent to which polymorphisms are located on these interfaces and to assess the potential impact on antibody binding, and therefore on the efficacy of the antibody therapeutic.

The first objective is to assemble and annotate the variant data set of naturally occurring human polymorphisms which are located in the epitopes of therapeutic antibodies. This requires the mapping of the polymorphisms on quality-controlled protein structures of antigen therapeutic complexes. Mutations which are located in the interface between antibody and antigen would constitute the variant data set on which we focus our attention. Variants in this data set are then further annotated with features related to structural properties and their frequency of occurrence in various human subpopulations. We then use these annotations to select a subset of targets for which we plan to measure the impact of the discovered variants on the interaction between antibody and antigen experimentally.

4.1 Methods

4.1.1 Overview

The used methodology can be broken down into four steps (Figure 4.1): (A) The selection and annotation of therapeutic mABs with structural information (B) Quality control of the mAB complex structures resulting in a curated structure database (C) The aggregation of human polymorphisms located on the antigens in the curated database and (D) the automatic analysis and annotation of structure and variant data with Var3D. The result is a list of annotated variants located on the interfaces of therapeutic mAB and their antigen target.

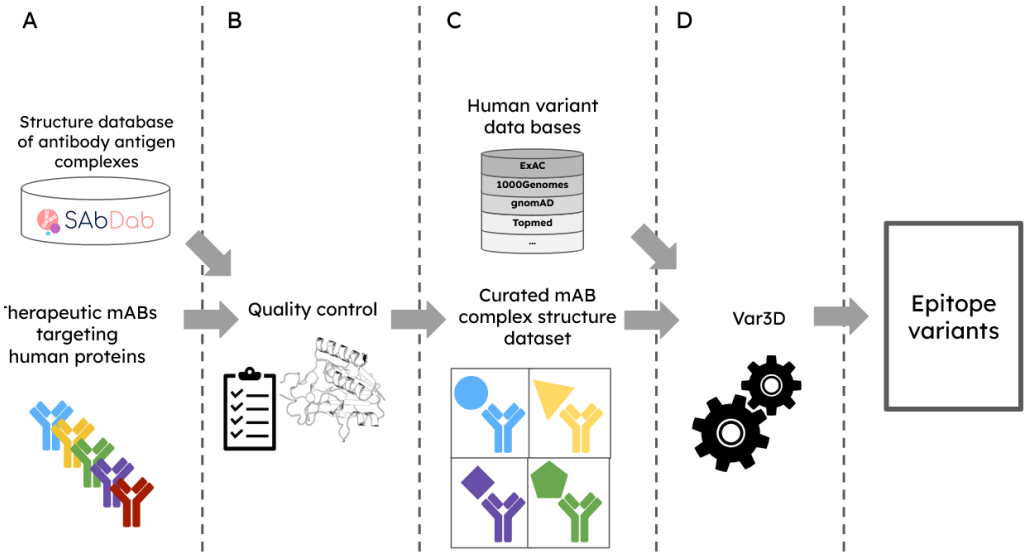

Figure 4.1: Pipeline to obtain human polymorphisms on the interface of therapeutic mABs and their antigen. The steps can be separated into (A) selection and annotation of therapeutic antibodies (B) Quality control of antibody-antigen complex structures (C) Identification of antigen proteins and mapping of human polymorphisms (D) Annotation of variants in their structural context with Var3D. The result is a data set of epitope variants.

4.1.2 Selection and Annotation of Therapeutic Antibodies

We compiled a list of clinically relevant antibodies for which structural information of their complex with their molecular targets was available, with a specific focus on human targets. We used the Therapeutic Structural Antibody Database (Thera-SAbDab)[4] as a reference for the search. This database links antibody therapeutics with protein structure data from the Protein Data Bank (PDB).

For this set of therapeutics we researched their clinical indications, estimates of sales of approved therapeutics and most common brand names. The clinical indications were obtained from DrugBank[5], a knowledge base with comprehensive molecular information on drugs, their mechanisms, their interactions and their targets.

The most prominent brand name and the numbers on the sales of an approved therapeutic in the year 2021 were inferred from various articles in economic and pharmaceutical newsletters and publicly available yearly sales reports of companies selling the approved therapeutics.

4.1.3 Quality Control of Antibody-Antigen Complex Structures

The structure files of the therapeutics were obtained from the TheraSabDab web page. The obtained set of structures required further refinement to only select the protein structures which contained the complex of the therapeutic antibody together with its designated antigen target. The following three conditions were used for the selection: 1.) The presence of the antigen protein chain in the structure 2.) The antigen is in contact with the antibody CDR 3.) The amino acid sequence of the antibody protein chain is identical to the sequence of the therapeutic mAB.

The structure metadata of TheraSabDab indicates the structures of the complexes as opposed to structures of the mAB alone (Figure 4.2, left). In the next step, the complex structures were manually curated to select only the antibody-antigen complexes representing a biologically relevant state. Complexes in which the antigen is not in contact with the CDR were discarded (Figure 4.2, right).

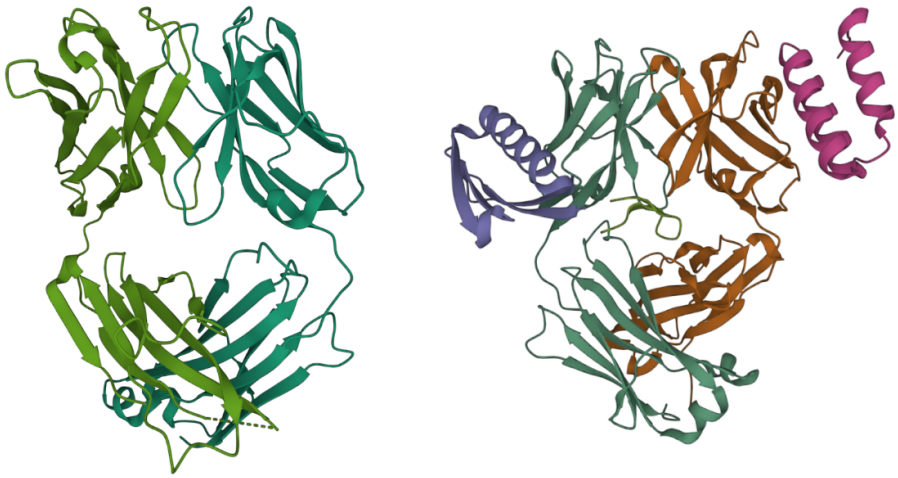

Figure 4.2: Examples of discarded structures: Left: The crystal structure of adalimumab FAB fragment (PDB ID 4NYL) does not contain an antigen protein chain. Right: Trastuzumab Fab v3 in complex with 5-phenyl meditope variant (PDB ID 6B9Y) has protein chains in contact with the antibody, but they are not interacting with the CDR located on top of the antibody and the proteins are not the designated target of Trastuzumab (Receptor tyrosine-protein kinase erbB-2).

We used the OpenStructure software framework[6] to align the sequence of the light and heavy chains in the structures to the therapeutic se-

quence using a modified version of the Needleman/Wunsch algorithm (in OpenStructure: ost.seq.SemiGlobalAlign). Any structure which had the sequences of their antibody chains not match the respective therapeutic mAB sequence in the data set was discarded.

4.1.4 Identification of Antigen Proteins and Mapping of Human Polymorphisms

Using the API of the universal protein knowledge base UniProt[7] we obtained the UniProt Accession code for all antigen protein chains present in our structure data set. The accession code was used to acquire all available human polymorphisms occurring on this set of proteins using the Search API of the European Bioinformatics Institute[8]. We obtained variant data from various large-scale population studies like ExAC[9], the genome aggregation database gnoMAD[10], the 1000 Genomes project[11], TopMed[12], the Exome Sequencing Project (ESP)[13], ClinGen[14], ClinVar[15] and others. The variant data are furthermore annotated with their somatic status, and their association with disease and also contain sequence-based variant impact predictions using PolyPhen[16] and SIFT[17].

Variant Allele frequencies in various human subpopulations were retrieved from the NCBI database of genetic variation dbSNP[18]. We identify the respective genetic allele for every protein variant in our data set and obtain the frequencies from the following studies: ExAC, gnomAD and 1000Genomes. The frequency data is clustered according to the major populations of the earth (African, American, Asian, European, East Asian, Ashkenazi Jew) with the addition of the "Global" category accounting for all population groups together and the "Other" category for individuals which do not unambiguously cluster with any of the major populations.

4.1.5 Annotation of Variants in their Structural Context with Var3D

The Var3D pipeline (Schema in Figure 4.3) implemented for this project contains a new epitope detector and adjustments to the calculations of the relative solvent accessibility (RSA) and the free energy estimation.

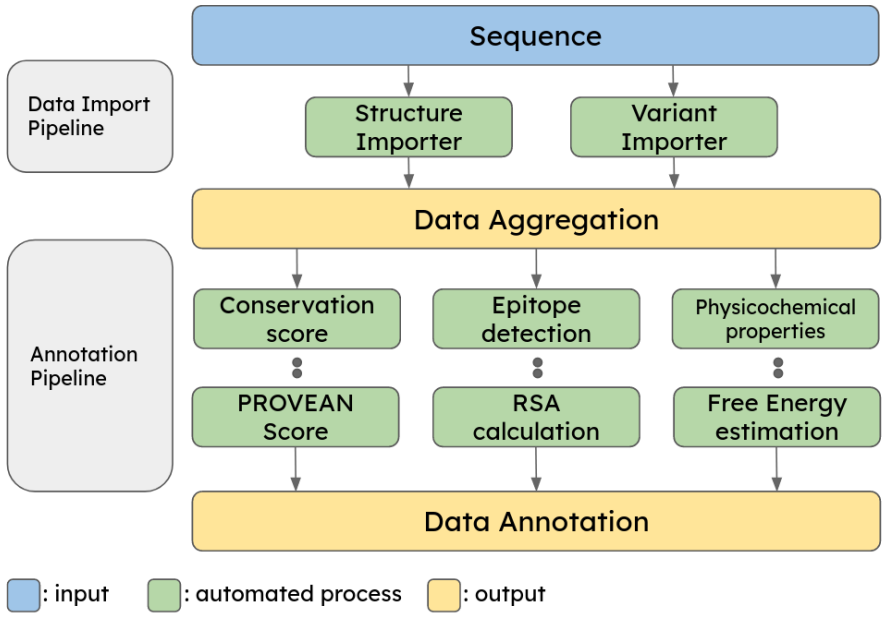

Figure 4.3: Schematic overview of the Var3D pipeline used for the analysis of the antigen variant data. The pipeline first creates a sequence-centric data structure which links variants and structures to a specific protein sequence. In the second step, various subprocesses annotate the data aggregation.

An epitope was defined as the set of residues in the antigen protein chains which have a distance lower than 5 Ångstrom to any residue in the antibody. Variants lying on these residues are considered to be epitope variants.

The calculation of the per-residue RSA was adjusted to include the calculation and the difference of the RSA between the bound and unbound state of the antigen (Figure 4.4). Similar adjustments were made to the free energy estimator Figure 4.5).

Variants which show a high energy difference upon mutation in the complex and a low energy difference in the apo form are of interest to us

because they show that the mutation is primarily altering chemical interactions of the mutated residue with antibody residues. We were interested in mutations for which a large difference between the estimated free energy of the complex structure and the apo form was observed. Additionally, we can use the free energy difference in the apo form to detect mutations which are generally deleterious to the protein. Ideally, a variant we would consider critical would show a high free energy difference in the complex and inconspicuous energy estimates in the apo form of the protein.

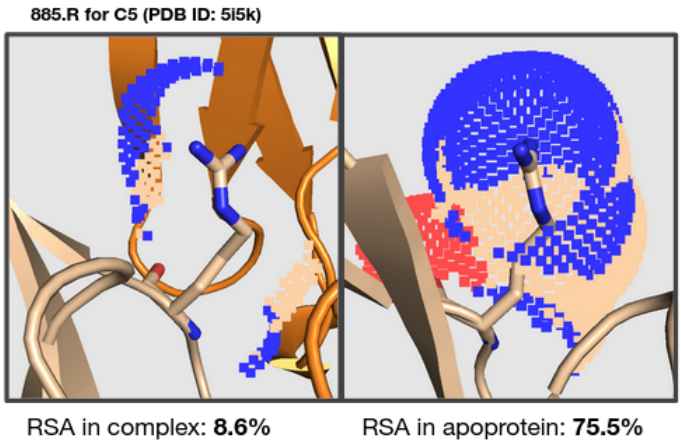

Figure 4.4: RSA calculation with 885.R of C5 as an example[1]. The Arginine is deeply buried in the complex while being very accessible in the apoprotein.

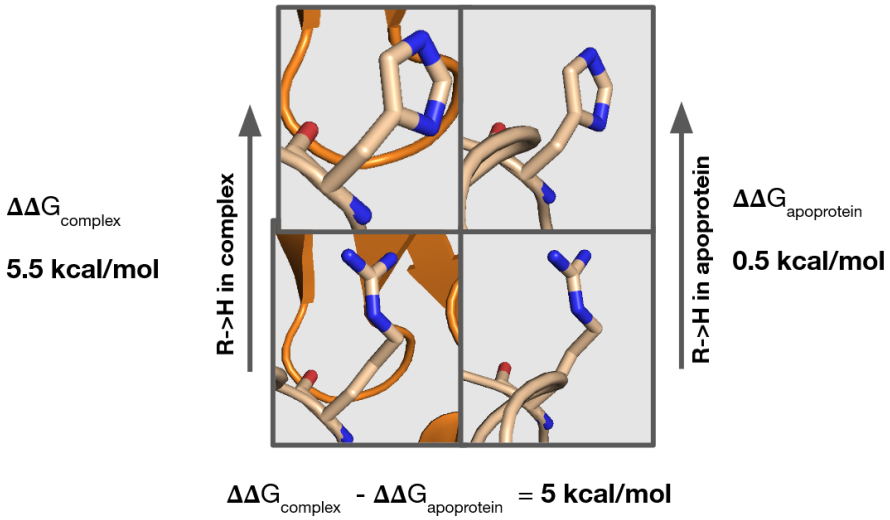

Figure 4.5: Energy difference estimation of R885H in C5 as an example[1]. The mutation of the arginine to a histidine is estimated to cause a free energy increase of 5.5 kcal/mol in the complex while the same mutation only causes an increase of 0.5 kcal/mol in the apo form of the antigen. This indicates that the mutation is disruptive in the complex and benign in the apo form.

4.2 Results

4.2.1 Identification of Antibody Therapeutics with Structural Information

As of July 2022, there were 743 clinically relevant monoclonal antibody therapeutics present in Thera-SAbDab. 169 therapeutics were represented by at least one experimental structure of the full therapeutic antigen complex and 136 of these complexes have a human protein as their antigen (Figure 4.6).

This set of 136 therapeutics interacts with 73 different antigen proteins through 152 antibody-antigen interfaces. Some therapeutics are able to target multiple antigens, but we observe more frequently that multiple therapeutics target the same antigen.

94 of the therapeutics (69%) are still actively maintained or developed. Among the actively maintained therapeutics, 32 are approved and in active clinical use and 7 awaited the completion of the approval process. 54 therapeutics are still being investigated in clinical trials (20 in Phase

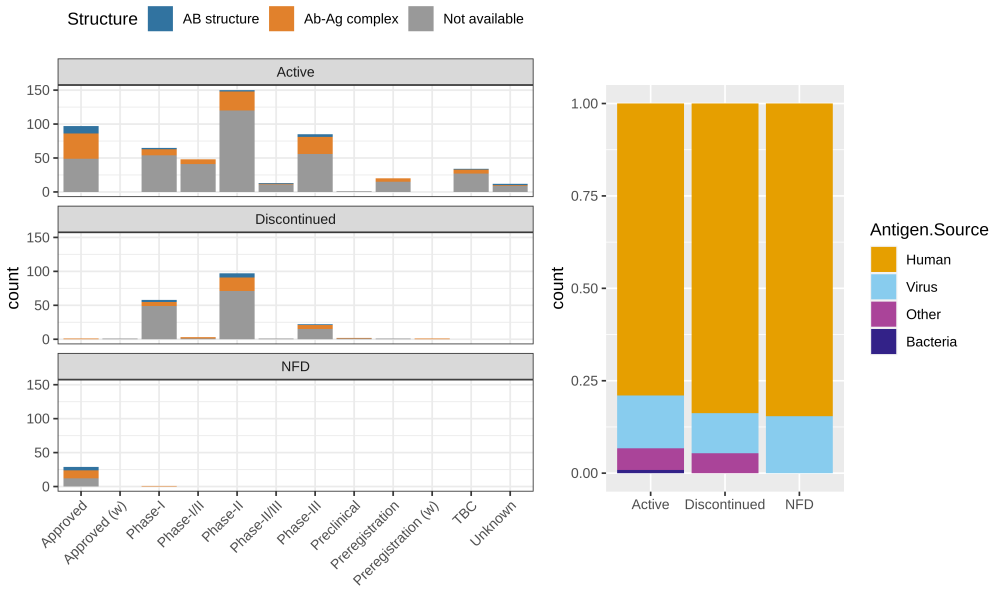

Figure 4.6: Overview over the entries in the therapeutic structural antibody database. The left panel shows the clinical stage and the current status of the 743 therapeutics. The right panel shows the organism in which the antigen target is expressed. (NFD: not further developed (no longer marketed or distributed), TBC: to be confirmed)

III, 21 in Phase II, 5 in Phase I/II and 8 in Phase I). The clinical stage of one therapeutic antibody was unknown.

31 (23%) of the therapeutics were discontinued. While most of these therapeutics (29) were stopped during clinical trials, one was discontinued in preregistration and one was approved and discontinued later due to the age of the therapeutic (Okt3, the first monoclonal antibody to be approved for clinical use in humans in 1986).

11 (8%) therapeutics are not further developed (NFD), which means that the drug is no longer marketed or distributed. 10 of these therapeutics were approved and 7 of them are still of high medical relevance today. This group includes important drugs like Remicade against rheumatoid arthritis or Aimovig which are used to treat migraines.

Table 4.1 shows the number of therapeutics grouped by indication categories. The majority are used against various cancer types, against autoimmune diseases (Arthritis, Lupus, Crohn's disease, ...) and to a lesser extent therapeutics developed against degenerative disorders (age-related diseases like Alzheimer, Parkinson's, age-related macular degeneration, ...).

Indication group	Number Therapeutics	Indication group	Number Therapeutics
Cancer	47	Chirurgical Aid	2
Autoimmune	33	Cholesterol	1
Degenerative disorder	13	Infectious Disease	1
Rare diseases	5	Migraines	1
Blood Disorder	2		

Table 4.1: Indication groups of the 136 antibodies with human targets and with protein structures of the respective antibody-antigen complex

As an indirect measure of the clinical relevance of a therapeutic we collected the available sales data for 32 approved therapeutics in the dataset (Figure 4.7). The best-selling therapeutic was Humira, a medication against rheumatoid arthritis, which would have been the drug generating the highest sales in 2021 with 20.694 billion USD were it not for the Pfizer and BioNTech COVID-19 vaccine[19]. Sales were generated in the same range by Keytruda, a cancer immunotherapeutic, with 17.186 billion USD.

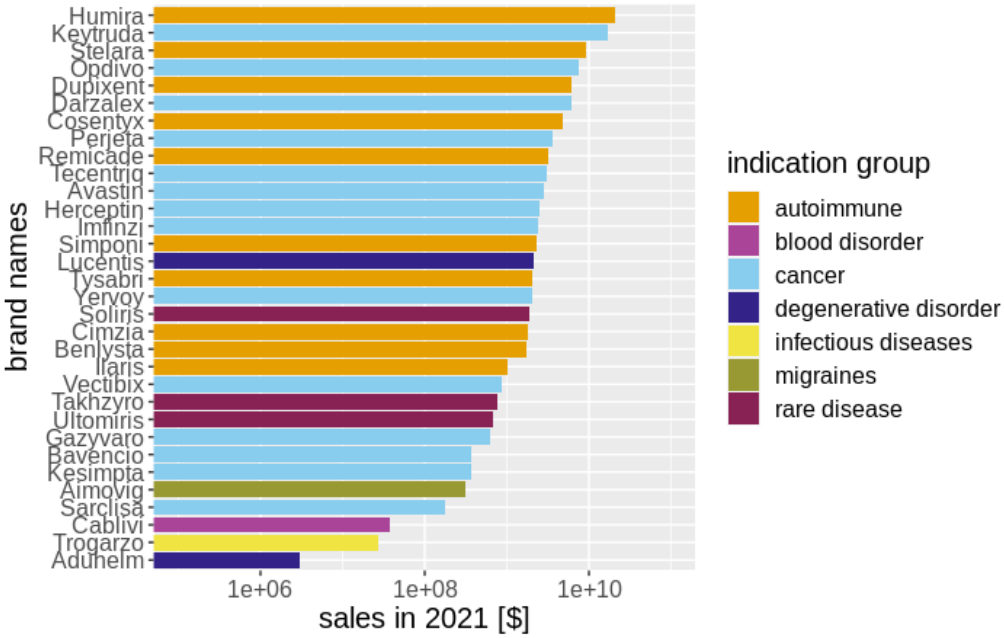

Figure 4.7: Sales numbers of approved therapeutic mABs for the year 2021 could be identified. Therapeutics are referred to by their most prominent brand name. The colours show the indication group of the therapeutic.

4.2.2 Annotation of Variants in their Structural Context

After the completion of the structure quality control on the set of 289 PDB entries obtained from Thera-SAbDab, we have a data set of 114 structures which represent 98 therapeutics, 62 antigens and 104 epitopes. 72% of the initial set of therapeutics could be retained.

Based on the sequence of the protein antigens in our dataset, we retrieved a total of 25'453 unique variants using EBI Search. Of these, 10' 352 variants were mapped onto the 3D structure of the antibody-antigen complexes using Var3D, and 1'389 were found to occur within the epitope region of 60 antigens interacting with 98 therapeutics through 102 different epitopes. The distribution of the variants across the antigens can be seen in Figure 4.9.

42 targets only interact with one therapeutic in our data set. 18 targets have interactions with multiple therapeutics with PDCD1 (programmed cell death protein 1) having the most interactions with 6 non-redundant therapeutics.

We find that every drug target has naturally occurring human polymorphisms which map to epitope regions. The epitope variants generally are spread relatively evenly across all antigens: the target with the lowest number of variants is Alpha-synuclein (SYUA) with 5 variants and the target with the highest number of variants is Interferon alpha-2 (IFNA2) with 58 variants. The mean and median of variant counts per target are 23.15 and 20.5 respectively.

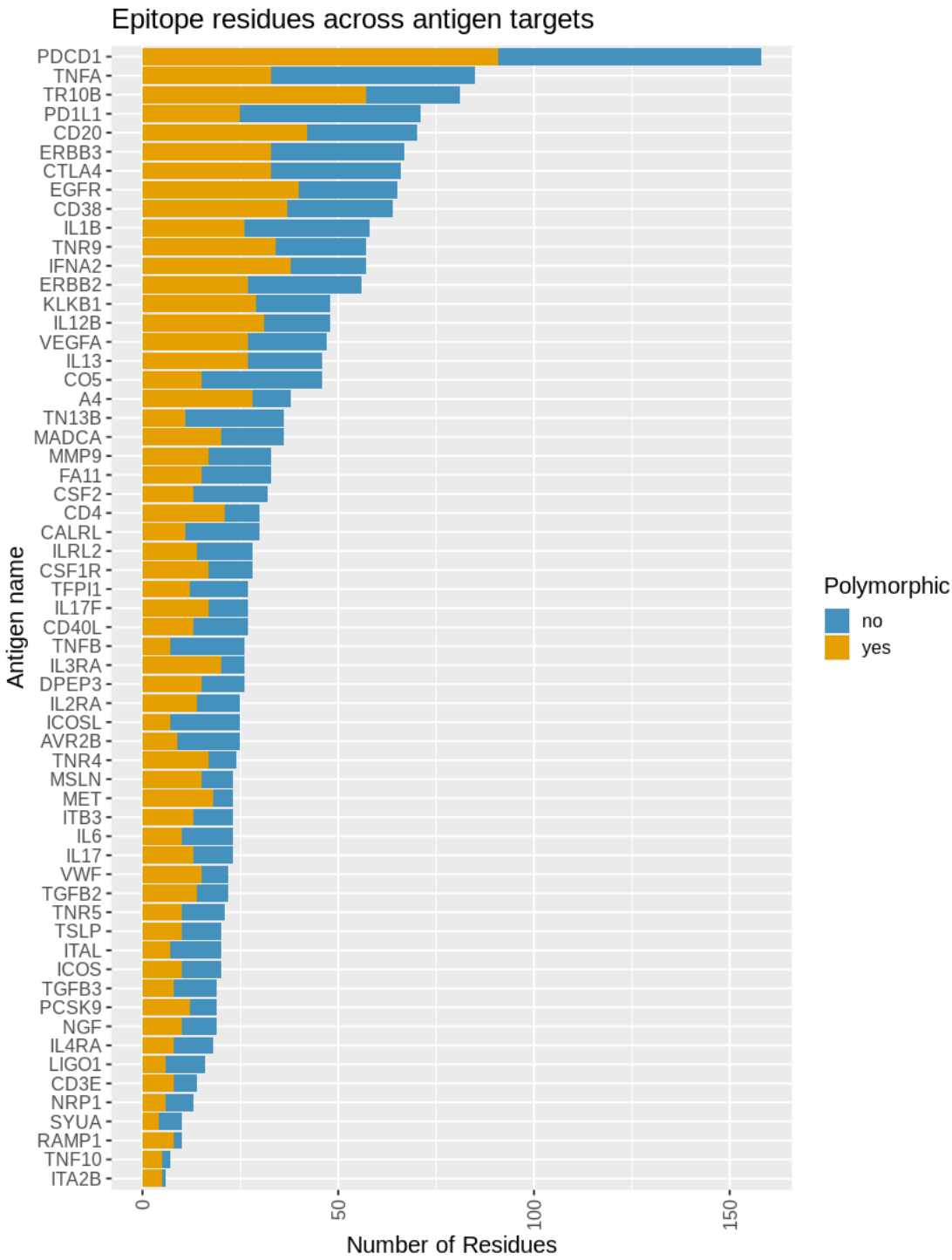

Figure 4.8: Overview over all epitopes and the number of human polymorphisms therein. On the y-axis are the investigated antigens. The x-axis represents the number of residues which are annotated as epitope residues. The number of sites in the epitope which are polymorphic are coloured orange.

4.2.3 Target Selection

We selected a subset of priority drug targets from our data set for which we will further investigate the degree of disruption of the interaction between therapeutic and target. We selected 10 antigen targets (Figure 4.10) using the following two criteria:

1. The therapeutics represented in this subset of targets should be of high clinical relevance. The data set mostly contains therapeutics against various types of cancers (7 targets, 20 therapeutics). The represented cancer therapeutics included are highly relevant and economically successful like Opdivo (Nivolumab), Darzalex (Daratumumab), Perjeta (Pertuzumab) and 7 more approved therapeutics. The therapeutic Dupixent (Dupilumab) is used to treat autoimmune diseases and Zinbryta (Daclizumab) was used to treat relapsing forms of multiple sclerosis. The two therapeutics Soliris (Eculizumab) and Crovalimab (Phase-III) are used to treat a rare disease.

2. Targets with non-overlapping epitopes (e.g. CD38, CO5, ERBB2, PD1L1 and PDCD1, see Figure 4.12) enable the comparison of a variant effect on the interaction with different antibody therapeutics which allows to investigate if a variant which prevents epitope recognition of one therapeutic can be avoided by using an alternative mAB which does not interact with the variant site.

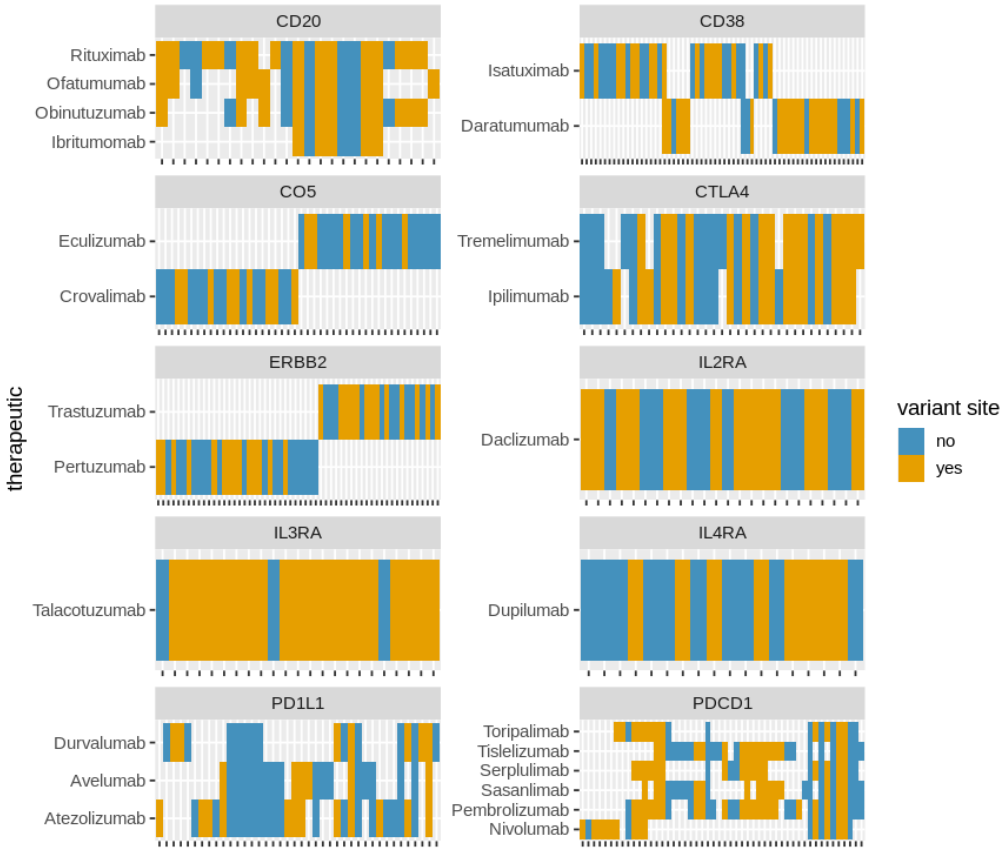

Figure 4.9: Overview of all epitopes and the localization of human polymorphisms therein for the targets selected for experimental validation. Every subplot represents an antigen target. On the y-axis are the therapeutics which are targeting the respective antigen. The x-axis represents all the residues of the antigen which are annotated as epitopes. If a residue is known to carry a human polymorphism, it is marked orange. If the residue is on an epitope, it is marked blue.

4.2.4 Selected Candidate Variants for Experimental Validation

Nivolumab (brand name Opdivo) is an important anticancer drug that targets a protein called PDCD1, which downregulates the immune system[20]. The mutation p.Pro28Leu in PDCD1 may interfere with the drug's effectiveness (Variant ID: rs56234260, Sources: ClinGen, 1000Genomes, ExAC, TOPMed, gnomAD). The highest frequency (0.0038) for this variant is reported in the "American" subpopulation of the ExAC dataset. First, the difference between the RSA value of the apo form and the complex form is very high (98.9% difference, 1.3% in complex, 100.2% in apo form) which indicates that the proline is not interacting with any other residue in the apo form but is deeply buried when interacting with the antibody. The mutation of the proline to the leucine is also estimated to be more deleterious in the complex (3.5 kcal/mol) when compared to the apo form (0.45 kcal/mol).

Upon further investigation of the structure of the complex of Nivolumab with PDCD1 (PDB ID: 5wt9), one can observe that the proline is located closely to p.Trp52 of the antibody chain (3.5 Ångstrom distance) (Figure 4.11), suggesting an interaction between the polarised C-H bonds in the antigen proline and the pi aromatic face of the antibody tryptophan.

When introducing a mutation to leucine using *in silico* mutagenesis[21], we observe clashes with the residues p.Trp52 and p.Val50 are unavoidable. The mutation is expected to abolish a favourable hydrophobic interaction and introduce repulsion between the antigen leucine and the antibody tryptophan. Therefore, we expect that the binding affinity of the AB should be significantly reduced for the variant protein target.

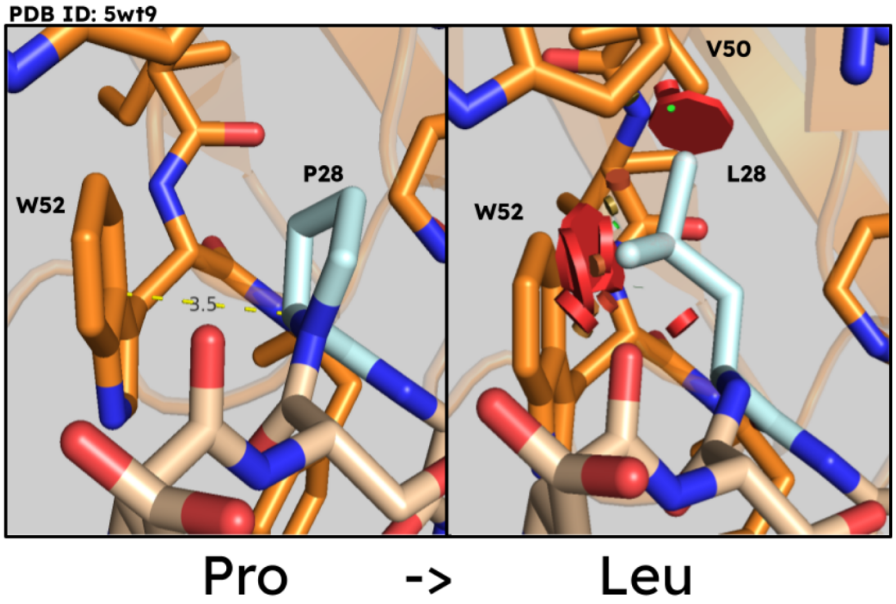

Figure 4.10: Potential critical variant p.Pro28Leu in PDCD1 to showcase the mechanism in which the mutation could disrupt the interaction with the cancer drug Nivolumab. Left: The mutation site p.Pro28 (cyan) on the antigen chain (wheat) is a proline which probably undergoes a hydrophobic interaction with the p.Trp52 of the heavy chain of Nivolumab. If the site is mutated to a Leucine, clashes with p.Trp52 (symbolised by red discs) and p.Val50 are unavoidable.

The former drug Daclizumab (brand name Zinbryta) is used to treat multiple sclerosis by targeting the interleukin IL2RA inhibiting their mediation in the activation of lymphocytes[22]. An interesting candidate variant is the polymorphism p.Gly173Arg (Variant ID: rs752423140, Sources: ClinGen, ClinVar, ExAC, TOPMed, dbSNP, gnomAD). The highest population frequency of 0.0012 for this mutation was reported for the African subpopulation in the "gnomAD - Exomes" dataset. The solvent accessibility difference between the apo form and the complex changes by 50%. The free energy difference of the mutation in the complex is estimated to be at 5.3 kcal/mol while the same difference in the apo form is estimated at 2.3 kcal/mol. The mutation is predicted to not be tolerable in the apo form, but it is even less tolerated in the complex.

Upon inspection of the mutation site in the drug target complex structure (Figure 4.12), it can be observed that the wild-type glycine is adjacent to an antibody tyrosine (p.Tyr48) and an antibody valine (p.Val102). The distance is too large to assume any meaningful chemical interaction.

The problem caused by the mutation becomes evident when trying to introduce an arginine into the limited space. The energetically most favourable rotamer of arginine which can be introduced into this site is still clashing strongly with the tyrosine in the immediate environment. Any potential rotamer of arginine is pointing into p.Tyr48, which means that without a conformational change of the antibody loop with the tyrosine clashes between the mutated arginine and the antibody are unavoidable. As a result, we anticipate that the AB will not exhibit a significant binding affinity to the variant protein.

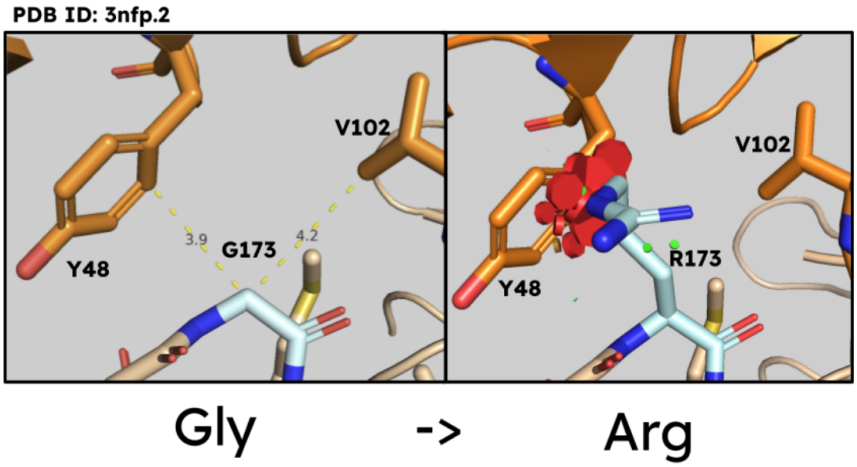

Figure 4.11: Potential critical variant p.Gly173Arg in IL2RA to showcase the mechanism in which the mutation could disrupt the interaction with the former MS drug Daclizumab. Left: The mutation site p.Gly173 (cyan) on the antigen chain (wheat) is a glycine which does not undergo any interaction with antibody residues. If the site is mutated to arginine, clashes with p.Tyr48 (symbolised by red discs) cannot be avoided.

The mAB Barecetamab is currently undergoing clinical trials as a treatment of lung and breast cancer by targeting the receptor tyrosine-protein kinase erbB-3. The overexpression of erbB-3 is thought to be a major cause of treatment failure due to its role in the activation of several biological pathways which increase the resilience of cancer cells[23]. The mutation p.Arg436Trp shows the potential of strongly disrupting the interaction between the drug and target (Variant ID: rs375932235, Sources: ClinGen, ESP, ExAC, TOPMed, gnomAD). The highest reported frequency for this mutation is 0.0003 for the Ashkenazi Jewish subpopulation in the "gnomAD - Genomes" dataset. The mutation site itself is deeply buried in the complex (RSA 2.7%) and relatively exposed (RSA 67.6%) in the apo form. Our free energy difference estimations predict a strong deleterious effect ($\Delta\Delta$G in the complex is 21 kcal/mol) while the mutation does not cause any significant energy shift in the apo form ($\Delta\Delta$G 0.3 kcal/mol).

In the complex structure (Figure 4.13), we can observe the importance of the residue p.Arg436 for the interaction between antibody and antigen: The arginine extends deeply into the antibody interface and has an electrostatic interaction with an aspartic acid of the light chain and a histidine of the heavy chain, connecting the residues in a line of complementary electrostatic interactions. This interaction is not only completely abolished upon the mutation to tryptophan, but the tryptophan is also predicted to clash with three residues of the heavy chain: p.Phe108, p.Tyr59 and p.His99. These clashes explain the predicted large value in the free energy estimation in the hypothetical complex. We anticipate that the mAB will not exhibit significant binding affinity to the variant protein.

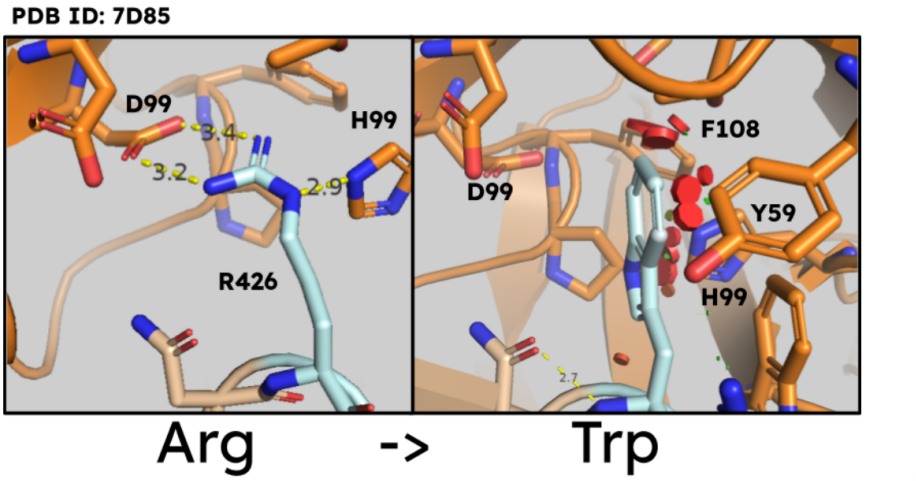

Figure 4.12: Potential critical variant p.Arg436Trp in ERBB3 to showcase the mechanism in which the mutation could disrupt the interaction with the anti-cancer drug Barecetamab. Left: The mutation site p.Arg436 (cyan) on the antigen chain (wheat) is an arginine which constitutes the central part of a system of electrostatic interactions between aspartic acid and histidine. If the site is mutated to tryptophan, the electrostatic interactions can not be maintained anymore and clashes with the 3 antibody amino acids p.Phe108, p.Tyr59 and p.His99 (symbolised by red discs) cannot be avoided.

4.3 Discussion

To our knowledge no overview of human polymorphisms on therapeutic antibody epitopes as comprehensive as the one presented here exists. Notably, our analysis shows that on this dataset all epitopes contained human polymorphisms, potentially affecting all therapeutics investigated in this project. Among those, our computational analysis shows the potential of some variants impacting the epitope recognition of antibody therapeutics. To confirm this hypothesis, we plan to conclude this project with the experimental validation of a subset of variants.

We could recover variants which are known to have a critical impact on the drug interaction in this project like p.Arg885His in CO5, showing the capability of our approach. We identified and showcased other potential candidate variants which could affect the drug efficacy in a similar way. Based on the work presented here, we plan to further refine the computational workflow and assess the capability to assist in the identification of further variant candidates. The currently available set of variants with experimental confirmation of their ability to disrupt epitope recognition is

currently too small to undertake this kind of project successfully. Further experimental profiling of the epitope variant data collected in this project could permit the computational prediction of critical epitope variants.

The implications of these findings on the clinical applications of antibody therapeutics is not yet clear. However, the amount and spread of the variants across the targets of all therapeutics in this dataset justify their further investigation and characterization.

4.4 Future Work

4.4.1 Experimental Characterization of Natural Polymorphisms at Antibody-Antigen Interfaces

We plan to measure the impact of the epitope variants on the target subset (refer to "Target selection" in this chapter). We want to validate the assertions we made on the impact of various variants based on their structural features. The target subset contains 10 different antigens, 21 unique therapeutics and a total of 316 variants.

We aim to characterise around 20 variants which show the potential to impact the epitope recognition of the therapeutic. The "Selected candidate variants for experimental validation" showcases some examples we would like to experimentally validate.

The experiments will be conducted in collaboration with the research group of Prof. Dr. Lukas Jeker of the Department of Biomedicine of the University and University Hospital of Basel. The reagents The experiments are planned to begin in February 2023. The cloning and expression of recombinant wild-type human proteins of the 10 selected target proteins in the HEK293 cell line[24] are currently ongoing and the respective antibody therapeutics have been ordered. The antibody binding is going to be measured using a fluorescence-activated cell sorter (FACS) antibody assay, a method used to sort cells based on their physical and chemical properties. Results are expected to be obtained within a few months.

References

[1] Nishimura, J.-i., Yamamoto, M., Hayashi, S., Ohyashiki, K., Ando, K., Brodsky, A. L., Noji, H., Kitamura, K., Eto, T., Takahashi, T. et al. (2014). Genetic variants in c5 and poor response to eculizumab. *New England Journal of Medicine*, *370*, 632–639.

[2] De Weers, M., Tai, Y.-T., Van Der Veer, M. S., Bakker, J. M., Vink, T., Jacobs, D. C., Oomen, L. A., Peipp, M., Valerius, T., Slootstra, J. W. et al. (2011). Daratumumab, a novel therapeutic human cd38 monoclonal antibody, induces killing of multiple myeloma and other hematological tumors. *The Journal of Immunology*, *186*, 1840–1848.

[3] Lu, R.-M., Hwang, Y.-C., Liu, I.-J., Lee, C.-C., Tsai, H.-Z., Li, H.-J., & Wu, H.-C. (2020). Development of therapeutic antibodies for the treatment of diseases. *Journal of biomedical science*, *27*, 1–30.

[4] Raybould, M. I., Marks, C., Lewis, A. P., Shi, J., Bujotzek, A., Taddese, B., & Deane, C. M. (2020). Thera-sabdab: the therapeutic structural antibody database. *Nucleic acids research*, *48*, D383–D388.

[5] Wishart, D. S., Feunang, Y. D., Guo, A. C., Lo, E. J., Marcu, A., Grant, J. R., Sajed, T., Johnson, D., Li, C., Sayeeda, Z. et al. (2018). Drugbank 5.0: a major update to the drugbank database for 2018. *Nucleic acids research*, *46*, D1074–D1082.

[6] Biasini, M., Schmidt, T., Bienert, S., Mariani, V., Studer, G., Haas, J., Johner, N., Schenk, A. D., Philippsen, A., & Schwede, T. (2013). Openstructure: an integrated software framework for computational structural biology. *Acta Crystallographica Section D: Biological Crystallography*, *69*, 701–709.

[7] UniProt Consortium (2021). Uniprot: the universal protein knowledgebase in 2021. *Nucleic acids research*, *49*, D480–D489.

[8] Madeira, F., Pearce, M., Tivey, A. R., Basutkar, P., Lee, J., Edbali, O., Madhusoodanan, N., Kolesnikov, A., & Lopez, R. (2022). Search and sequence analysis tools services from embl-ebi in 2022. *Nucleic acids research*, *50*, W276–W279.

[9] Karczewski, K. J., Weisburd, B., Thomas, B., Solomonson, M., Ruderfer, D. M., Kavanagh, D., Hamamsy, T., Lek, M., Samocha, K. E., Cummings, B. B. et al. (2017). The exac browser: displaying reference data information from over 60 000 exomes. *Nucleic acids research*, *45*, D840–D845.

[10] Karczewski, K., & Francioli, L. (2017). The genome aggregation database (gnomad). *MacArthur Lab*, (pp. 1–10).

[11] Siva, N. (2008). 1000 genomes project. *Nature biotechnology*, *26*, 256–257.

[12] Burgess, D. J. (2021). The topmed genomic resource for human health. *Nature Reviews Genetics*, *22*, 200–200.

[13] NHLBI GO Exome Sequencing Project (ESP). Exome variant server. `https://evs.gs.washington.edu/EVS/`. [Accessed 22-Jan-2023].

[14] Rehm, H. L., Berg, J. S., Brooks, L. D., Bustamante, C. D., Evans, J. P., Landrum, M. J., Ledbetter, D. H., Maglott, D. R., Martin, C. L., Nussbaum, R. L. et al. (2015). Clingen—the clinical genome resource. *New England Journal of Medicine*, *372*, 2235–2242.

[15] Landrum, M. J., Lee, J. M., Benson, M., Brown, G., Chao, C., Chitipiralla, S., Gu, B., Hart, J., Hoffman, D., Hoover, J. et al. (2016). Clinvar: public archive of interpretations of clinically relevant variants. *Nucleic acids research*, *44*, D862–D868.

[16] Sunyaev, S., Ramensky, V., Koch, I., Lathe III, W., Kondrashov, A. S., & Bork, P. (2001). Prediction of deleterious human alleles. *Human molecular genetics*, *10*, 591–597.

[17] Ng, P. C., & Henikoff, S. (2003). Sift: Predicting amino acid changes that affect protein function. *Nucleic acids research*, *31*, 3812–3814.

[18] Sherry, S. T., Ward, M.-H., Kholodov, M., Baker, J., Phan, L., Smigielski, E. M., & Sirotkin, K. (2001). dbsnp: the ncbi database of genetic variation. *Nucleic acids research*, *29*, 308–311.

[19] Buntz, B. Pharma50: 50 of 2021's best-selling pharmaceuticals: Drug discovery & development. URL: `https://www.drugdiscoverytrends.com/50-of-2021s-best-selling-pharmaceuticals/`.

[20] Pardoll, D. M. (2012). The blockade of immune checkpoints in cancer immunotherapy. *Nature Reviews Cancer*, *12*, 252–264.

[21] Schrödinger. The pymol molecular graphics system, version 1.30.

[22] Alcina, A., Fedetz, M., Ndagire, D., Fernández, O., Leyva, L., Guerrero, M., Abad-Grau, M. M., Arnal, C., Delgado, C., Lucas, M. et al. (2009). Il2ra/cd25 gene polymorphisms: uneven association with multiple sclerosis (ms) and type 1 diabetes (t1d). *PloS one*, *4*, e4137.

[23] Ma, J., Lyu, H., Huang, J., & Liu, B. (2014). Targeting of erbb3 receptor to overcome resistance in cancer treatment. *Molecular cancer*, *13*, 1–9.

[24] Thomas, P., & Smart, T. G. (2005). Hek293 cell line: a vehicle for the expression of recombinant proteins. *Journal of pharmacological and toxicological methods*, *51*, 187–200.

General Discussion

4.5 Summary

This thesis aimed to advance the investigation of the impact of protein-coding variants on protein structures. Using 3D representations of proteins, we identify the respective mutation site and from the structural environment we deduce parameters to describe the effect a specific amino acid might have in this chemical environment. The interpretation of the variant is complemented by the aggregation of conservation-based features, physicochemical features, functional annotations, solvent accessibility calculations and the estimation of the free energy difference upon mutation. By combining these features with the visual inspection of the protein structure environment of the site, a well-grounded hypothesis can be constructed on the impact the variant might have on its immediate structural environment. This process was streamlined and upscaled by the creation of the Var3D variant analysis engine.

This analysis workflow was applied to facilitate the structure-based variant interpretation of antibiotic resistance variants in MTB. The release of a resistance variant catalogue by the WHO and the publication of the revolutionary structure prediction method AlphaFold2 made it possible to obtain protein structure models for all antibiotic resistance target proteins for a comprehensive view of the MTB resistome. The mutation catalogue showed that 90% of the variants were still annotated as "uncertain", demonstrating the need for computational tools to aid in their further characterization. Combining Var3D with the SWISS-MODEL technology stack and a manually curated structure data set of resistance targets, we were able to create the TBvar3D web server. The server facilitates the inspection of variants in a data-rich 3D context which would

otherwise require manual time-consuming structure modelling steps, data integration and visualisation. The results are displayed on the web server and do not require the specific computational expertise of the user. This makes TBvar3D a valuable tool to assist researchers worldwide to form compelling hypotheses on the impact of variants for MTB.

The research goal of the second project in this thesis is the identification of naturally occurring human polymorphisms in the interfaces of antibody therapeutics and their respective antigen targets which may impact antibody binding. Individual cases of natural variants preventing epitope recognition were reported in the literature, but a comprehensive investigation was never performed before. With more than 100 antibody therapeutics approved and far more therapeutics undergoing clinical trials, the investigation of the potential impact this phenomenon might have on the efficacy of antibody drugs becomes pertinent. Through the use of Var3D, it was possible to map and annotate around 25'000 human variants on over 100 structure models of drug target complexes representing 98 therapeutics. We identified around 1'400 naturally occurring polymorphisms distributed across every single epitope, showing the phenomenon of natural polymorphisms occurring on clinically relevant epitopes to be quite common. Among this variant data set, we are planning to experimentally characterise the impact variants located on 10 prolific antibody therapeutics.

4.6 Future Outlook

The results of the Critical Assessment of Structure Prediction (CASP14) in 2020 revolutionised the protein structure prediction field. The AlphaFold2 (AF2) algorithm of DeepMind performed so well in the task of predicting protein structures that the problem for single protein chains was considered to be solved. The high quality of the predictions made by AF2 is especially valuable for variant interpretation where the correct orientation of the amino acid side chains is a prerequisite for an accurate perspective of the mutation site.

Attempts to model the effect of mutations on the 3D structures of proteins were not as successful. AF2-based prediction methods do not account for changes introduced by small-scale mutations of the structure.

But the development of a methodology which is tailored to correctly predict molecular effects upon mutation with high accuracy was never more in reach in the history of structural biology. Having an accurate prediction of the effect of a mutation on protein folding, structure, and dynamics would be a crucial technology for the study of the structural and functional consequences of variations. We would not be relying on indirect measurements of variant impact anymore and could directly assess and analyse the structural alterations caused by a mutation.

The structure prediction revolution of AF2 opens up the discussion on the nature of protein structures and the models we have for them. A protein is not a rigid constellation of atoms as it is suggested by our static representations and visualisations, they are quite dynamic and their structure is better understood as an ensemble of different conformations. A protein is also constantly interacting with the surrounding solvents, small molecules, nucleic acids and other proteins. The cytoplasm of a cell is quite densely populated with biomolecules. The potential conformational changes have a big impact if one wants to accurately investigate small-scale mutations like single amino acid substitutions: if the full conformational space is not known, important interactions between a wild-type amino acid and its environment will not be considered. The incorporation of dynamics for variant interpretation is still a significant obstacle in the field. But with the solution of the "static" protein structure prediction problem in sight, incorporating dynamics into protein structure bioinformatics will be a logical next step for the field to advance, a step which will greatly benefit the interpretation of protein variants in their structural context.

4.7 Closing Remarks

The field of variant interpretation was exciting to work in. Connecting cutting-edge computational methods with clinically relevant projects was very rewarding. The development in recent years in structural bioinformatics presented new opportunities to automate the process of structure-based variant interpretation. With this thesis, I brought the capabilities of structural bioinformatics closer to a wide scientific audience through TBvar3D and systematically analysed the variant space of clinically relevant epitopes.

List of Publications

TBvar3D: Mapping antibiotic resistance variants in Mycobacterium tuberculosis on 3D protein structures; Erblin Asllanaj, Gabriel Studer, Andrew Waterhouse, Rosalba Lepore and Torsten Schwede; *Manuscript in preparation*

Impact of naturally occurring human polymorphisms at interfaces of antibody therapeutics and antigens; Erblin Asllanaj, Rosalba Lepore, Romina Matter-Marone, Lukas Jeker and Torsten Schwede; *Manuscript in preparation, author list may change*

Var3D Source Code:
https://git.scicore.unibas.ch/schwede/var3d

TBvar3D Web Server:
https://swissmodel.expasy.org/var3d/

Acknowledgements

I would like to express my deepest gratitude to my supervisor, Torsten Schwede, for his unwavering guidance and support over the past five years. He kept me on track when I strayed and ensured that my work was properly funded.

I would also like to extend my thanks to Timm Maier and Vincent Zoete for serving on my PhD committee and providing invaluable advice throughout the process. I want to thank Vincent additionally for his kind hospitality in hosting me in his research group for one month. I had the opportunity to gain many valuable insights from Vincent and the other members of his team.

I am particularly grateful to Rosalba Lepore for her mentorship and unwavering support throughout my journey. Her influence on my development as a scientist has been immense. I could always count on her for both scientific guidance and emotional support.

I am thankful to all of the not previously mentioned members, past and present, of the Schwede research group for creating a stimulating and friendly environment for scientific discovery. I would like to thank Leila Alexander, Stefan "Bienchen" Bienert, Marc Creuss, Janani "Jay" Durairaj, Jérôme Eberhardt, Rafal Gumienny, Jürgen Haas, Lorenzo Pantolini, Joana Pereira, Xavier Robin, Gabriel Studer, Gerardo Tauriello and Andrew Waterhouse.

I would also like to give special thanks to Alba, Jay, Jérôme, Joana and Gabriel for valuable feedback on this thesis.

I want to extend my thanks to the team of the Center of Scientific Computing (sciCORE) at the University of Basel for their tireless technical support and expertise.

I would like to express my deepest gratitude to Sébastien Gagneux and the TB Research Unit at the Swiss Tropical and Public Health Institute

for the opportunity to participate in their group meetings and events which were a constant source of inspiration and greatly contributed to my understanding of MTB.

I would like to express my appreciation for the support of Sarah Guethe, Rita Manohar, and Yvonne Steger-Bieli in navigating the administrative aspects of my PhD.

I would like to thank The Swiss Institute of Bioinformatics (SIB) for providing me with great opportunities to connect with bioinformaticians all over Switzerland.

Last but not least, I am forever grateful to my family for their unwavering support and love. My father Afrim, my mother Raza and my siblings Ermira and Erdrin have been my rock throughout this journey. Ju dua pergjithmone.

Cover design by Janani Durairaj

Dr. phil. Erblin Asllanaj

Bläsiring 42, 4057 Basel, CH
✆ (+41) 78 883 75 38 | ☒ erblin.asllanaj@unibas.ch | Swiss Citizen

Summary

Bioinformatician with a research focus in structure biology, drug resistance research and machine learning. Always excited to work on innovative projects; I am quick to adapt to new skills, solve tricky problems, communicate professionally and dedicate myself to a team.

Work Experience

Postdoctoral Research Associate – *Biozentrum, University Of Basel*

Apr 2023 – Sep 2023

- Completing and publishing pending projects from my doctoral research
- Supervisor: Prof. Dr. Torsten Schwede

Education

PhD. Computational Biology – *Biozentrum, University Of Basel*

Feb 2018 – Mar 2023

- Main Thesis Subject: Investigation of drug resistance variants in the context of protein structures
- Supervisor: Prof. Dr. Torsten Schwede

M.Sc. Systems Biology – *D-BSSE, ETH Zürich*

Sep 2015 – Sep 2017

- Master Thesis Subject: Random sampling of metabolic fluxes in microbial community models”
- Supervisor: Prof. Dr. Jörg Stelling

B.Sc. Computational Biology – *University Of Basel*

Sep 2012 – Aug 2015

PhD Projects Overview

- **Investigation of natural polymorphisms at antibody-antigen interfaces:** I systematically detected all naturally occurring human polymorphisms on therapeutic antibody epitopes and experimentally validated for a subset if the polymorphism has an effect on the therapeutics epitope recognition
- **SARS-CoV-2 Remdesivir Resistance:** In a collaboration with another research group, I provided a detailed structure-based analysis for resistance variants in SARS-CoV-2 resulting from their global genomic analysis.
- **TBvar3D: Antibiotic Resistance Variants and Protein Structures:** I worked on a web server which integrates recently published resistance variant data of MTB with the respective protein structure and accompanying automatic variant analysis. I developed this project and I am involved in the project planning, the software development and frontend design of this web server.

Technical Expertise

Languages

German: Native Speaker (C2), English: Fluent (C2), French: Intermediate (B2), Albanian: Native Speaker (C2)

Computational Skills

OS: LINUX, Windows, iOS

Programming Languages: Python (advanced), C/C++ (beginner)

Others: JavaScript, Latex, HTML, MongoDB, Singularity, Docker, GitHub, PyTorch, R, Matlab, Excel

Teaching

- Participated in the design and presentation of a group-internal course “Introduction and Practical Applications of Neural Networks”.
- Designed and presented a masters-level lecture on protein structures with exercises for the lecture series “Introduction To Bioinformatics” at the Swiss Tropical And Public Health Institute.

Publications

Mari, Alfredo, Tim Roloff, Madlen Stange, Kirstine K. Søgaard, **Erblin Asllana**j, Gerardo Tauriello, Leila Tamara Alexander et al. **"Global Genomic Analysis of SARS-CoV-2 RNA Dependent RNA Polymerase Evolution and Antiviral Drug Resistance**." *Microorganisms* 9, no. 5 (2021): 1094

Erblin Asllanaj, Andrew Waterhouse, Gabriel Studer, Rosalba Lepore, Torsten Schwede **“TBvar3D: Antibiotic Resistance Variants and Protein Structures**" *Manuscript in preparation*

Erblin Asllanaj, Rosalba Lepore, Romina Matter-Marone, Lukas Jeker and Torsten Schwede “**Impact of naturally occurring human polymorphisms at interfaces of antibody therapeutics and antigens**” *Manuscript in preparation*

Interests

- Exploring the extent to which neural networks can solve biological problems
- drug design, drug resistance mechanisms, antibody therapeutics
- application of protein structures
- communicating and teaching advanced scientific topics to any audience

References

Torsten Schwede	Vice President for Research, University Of Basel Professor for structural bioinformatics, Biozentrum, University of Basel	*torsten.schwede@unibas.ch*
Rosalba Lepore	Principal Scientist at Cimeio Therapeutics	*albalepore@gmail.com*
Mubera Krijezi	Roche Diagnostics Solutions - Global Clinical Program Lead	*mubera.krijezi@roche.com*