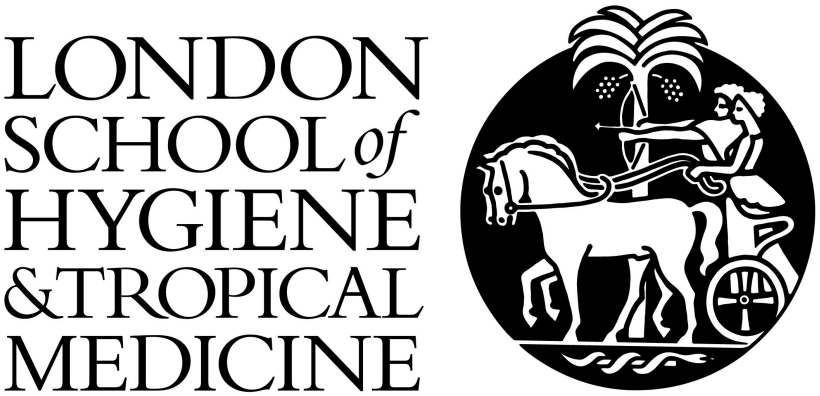


Using Machine Learning to Anticipate Antimicrobial Resistance in *Mycobacterium Tuberculosis*

Tanushree Tunstall

Thesis submitted in accordance with the requirements for the degree of

Doctor of Philosophy

University of London

2023

Department of Infection Biology

Faculty of Infectious and Tropical Diseases

LONDON SCHOOL OF HYGIENE & TROPICAL MEDICINE

Funded by BBSRC

Research group affiliation(s): Dr Nicholas Furnham
Professor Taane Clark

London School of Hygiene & Tropical Medicine
Keppel Street, London WC1E 7HT

T: +44 (0)20 7299 4646
F: +44 (0)20 7299 4656
www.lshtm.ac.uk

DECLARATION OF OWN WORK

All students are required to complete the following declaration when submitting their thesis.

Please note: Assessment misconduct includes any activity that compromises the integrity of your research or assessment of is will be considered under the Assessment Irregularity Policy. This includes plagiarism, cheating and failure to follow correct progression and examination procedures.

Please see the following documents for further guidance:

- Research Degrees Handbook
- Assessment Irregularities Policy

Supervisors should be consulted if there are any doubts about what is permissible.

1. STUDENT DETAILS

Student ID Number	LSH1806129	**Title**	Mrs
First Name(s)	Tanushree		
Surname/Family Name	Tunstall		
Programme of Study	Doctor of Philosophy		
LSHTM Email (if this is no longer active, please provide an alternative)	tanushree.tunstall@lshtm.ac.uk		

2. TITLE OF THESIS

Title of Thesis	Using Machine Learning to Anticipate Antimicrobial Resistance in Mycobacterium Tuberculosis

3. DECLARATION

I have read and understood the LSHTM's definition of plagiarism and cheating. I declare that this thesis is my own work, and that I have acknowledged all results and quotations from the published or unpublished work of other people.

I have read and understood the LSHTM's definition and policy on the use of third parties (either paid or unpaid) who have contributed to the preparation of this thesis by providing copy editing and, or, proof reading services. I declare that no changes to the intellectual content or substance of this thesis were made as a result of this advice, and, that I have fully acknowledged all such contributions.

I have exercised reasonable care to ensure that the work is original and does not to the best of my knowledge break any UK law or infringe any third party's copyright or other intellectual property right.

Student Signature	Tanushree Tunstall
Date	16/09/2022

Abstract

Antimicrobial resistance (AMR) continues to threaten public healthcare worldwide. Drug-resistant tuberculosis (DR-TB) is a major example of AMR, with resistance developing to multiple drugs, impeding treatment. Resistance in *Mycobacterium tuberculosis* (the bacterium causing human TB) is primarily mediated via mutations causing a single amino acid change at a specific position in a given protein, termed single amino acid variation (SAV). This thesis focusses on using computational methods to investigate the molecular consequences of SAVs on resistance development in the six main *M. tuberculosis* gene-drug targets: *alr*-cycloserine (DCS), *embB*-ethambutol (EMB), *gidB*-streptomycin (STR), *katG*-isoniazid (INH), *pncA*-pyrazinamide (PZA), and *rpoB*-rifampicin (RFP).

Mutation data was sourced from a genome-wide association study of over 35,000 clinical isolates. An analysis pipeline extracted over 4000 SAVs across all targets and calculated minor allele frequency, odds ratio and lineage contributions. Protein structure modelling and docking were performed to obtain the gene-drug complex in the absence of an experimentally determined structure. Multiple *in silico* estimators of mutational effects on protomer stability, molecular affinities, evolutionary conservation, and residue-level properties were calculated.

Initial analysis explored interrelationships between estimators for gene-targets. Visualisation tools, built to interactively inspect these relationships, aided the interpretation. Lineage effects on resistance were examined to understand the influence of epistasis. Together, these were used to build a supervised machine learning (ML) classification pipeline using multiple classifiers to predict resistance. ML models were built for individual and combined gene-drug targets, with the latter showing supervised ML classification could be used in a gene-agnostic manner to predict resistance.

Model performance was assessed using the Matthews Correlation Coefficient (MCC), with performance generally improving upon feature selection for most models. For individual gene-drug targets, ML prediction for predicting PZA resistance performed the best, with an MCC score of 0.52 achieved using the Multilayer Perceptron (MLP) classifier. This was followed by an MCC of 0.49 for RFP resistance prediction using the XGBoost model. EMB and INH resistance predictions followed equally with MCC scores of 0.42 using XGBoost for EMB, and both Linear Discriminant Analysis and Ridge classifiers for INH. For the combined model, PZA prediction was the highest with an MCC of 0.46 based on the Extra Tree classifier, followed by RFP prediction of 0.39 MCC using MLP. EMB resistance prediction was 0.34 MCC using the Random Forest classifier, and finally an MCC of 0.31 with Stochastic Descent for INH resistance prediction. INH resistance prediction was the lowest compared with other targets both in the individual and combined ML approaches, while DCS and STR resistance prediction results were inconclusive.

Exploiting a combined genomic and structural approach to understand mutational effects of resistance to anticipate resistance in a gene-agnostic manner would benefit clinical decision making and drug stewardship efforts. Future work could extend these methods to develop epistasis-informed ML models and apply transfer and unsupervised learning to other gene-targets in *M. tuberculosis*. The methods and pipelines developed can also be applied to other AMR pathogens.

Acknowledgements

To Papa

Mee

Shone

BBSRC, NPIF, and LIDo for funding this opportunity for me to make this humble contribution.

My supervisors Nick Furnham and Taane Clark for their time and support.

UQ colleagues and friends David Ascher, Stephanie Portelli, and Alex de Sa for hosting me and generously sharing their time and knowledge of ML.

LiDo comrades Ben Blundell and John Benest for ML brainstorming, Gary Napier for liberally sharing knowledge and wisdom, and Jody Phelan for always answering my questions with a smile.

My best friend Sunita Thakur for being there through thick and thin ...

Trevor Hansel for being a wonderful mentor, and inspiring it all...

My dear husband for his relentless support and optimism

... building a working home office through the pandemic

... keeping me sane, fed and entertained

... bringing me endless cups of tea

... keeping the household going

... contagious positivity

To paraphrase Isaac Asimov: "Education isn't something you can finish"

Table of Contents

List of Figures

List of Tables

Abbreviation List

aa	Amino Acid
ACT	Artemisin-based Combination Therapy
AMR	Antimicrobial Resistance
ART	Anti-Retroviral Therapy
CV	Cross Validation
DCS	Cycloserine
DM	Drug Mutations
DOPE	Discrete Optimised Protein Energy
DST	Drug Susceptibility Testing
EMB	Ethambutol
FDR	False Discovery Rate
FN	False Negative
FP	False Positive
GWAS	Genome-Wide Association Studies
HGT	Horizontal Gene Transfer
HIV	Human Immunodeficiency Virus
INDELs	INsertions/DELetions
INH	Isoniazid
KNN	K-Nearest Neighbours
LDA	Linear Discriminant Analysis
MAF	Minor Allele Frequency
MCC	Matthews Correlation Coefficient
mCSM	Mutation cut-off Scanning Matrix
MDR	Multi-Drug Resistant
MGIT	Mycobacteria Growth Indicator Tube
MIC	Minimum Inhibitory Concentration
ML	Machine Learning
MLP	Multi Layer Perceptron
mmCSM	Multiple Mutation cut-off Scanning Matrix
Mut	Mutant-type
NA	Nucleic Acid
NB	Naive Bayes
NMA	Normal Mode Analysis
nsSNP	Non-synonymous Single Nucleotide Polymorphism
nsSNV	Non-synonymous Single Nucleotide Variation
nt	Nucleotide
OM	Other Mutations
OR	Odds Ratio

PDB	Protein Data Bank
PLIP	Protein Ligand Interaction Profiler
PPI	Protein-Protein Interaction
PROVEAN	Protein Variant Effect Analyzer
PZA	Pyrazinamide
QDA	Quadratic Discriminant Analysis
RFECV	Recursive Feature Elimination with Cross Validation
RFP	Rifampicin
RMSD	Root Mean square Deviation
RNA	Ribonucleic Acid
RNAP	RNA polymerase complex
RRDR	Rifampicin Resistance Determining Region
SGD	Stochastic Gradient Descent
SMOTE	Synthetic Minority Oversampling
SNAP2	Screening for Non-Acceptable Polymorphisms v.2
SAV	Single Amino acid Variation
STR	Streptomycin
SVC	Support Vector Classifier
TB	Tuberculosis
TN	True Negative
TP	True Positive
WHO	World Health Organisation
Wt	Wild-type
XDR	Extensively-Drug Resistant
$\Delta\Delta$G	Change in Gibbs Free energy

Chapter 1

Introduction

1.1 Antimicrobial resistance

Antimicrobial resistance (AMR) is the ability of microorganisms including bacteria, viruses, fungi and parasites, to overcome the effects of drugs used to treat diseases caused by them. Drugs directed towards specific microbes such as antibiotics for bacteria, antivirals for viruses, antifungals for fungi, and anti-parasitics for parasites are collectively termed antimicrobials. AMR is an expected consequence of the Darwinian principle of survival of the fittest, where some microbes accumulate changes over successive generations to adapt and survive in the face of the pressure exerted by the drug. While a natural phenomenon, the widespread use, overuse, and misuse of antimicrobials in humans, animals and plant sectors has accelerated the emergence of drug resistant pathogens.[1] Infections caused by resistant pathogens have an adverse effect on human health, leading to prolonged hospital stays, poor disease outcome, less effective treatments, and potentially untreatable diseases.

1.1.1 Burden

The societal and economic consequences from AMR associated morbidity and mortality is predicted to be a staggering 100 trillion USD per year by 2050.[2] In February 2022 a comprehensive systematic review published in the Lancet medical journal by the Antimicrobial Resistance Collaborators[3] estimated a figure of 1.27 million deaths globally in 2019 attributable to bacterial AMR alone. This is far in excess of the 700,000 estimated global annual deaths reported by the 2016 O'Neill report and makes the predicted 10 million deaths from bacterial AMR by 2050,[2] far more imminent. The 2022 AMR review is an extensive study from 471 million isolates spanning 204 countries, which estimated this burden accounting for two alternative scenarios, highlighting 4.95 million preventable deaths in 2019 if all drug resistant infections were replaced by no infection (associated with resistance), and 1.27 million preventable deaths if all drug resistant infections were replaced by drug susceptible infections (attributable to resistance).[3] Nearly all major pathogenic diseases are affected by either prevailing or emerging resistance, with AMR being one of the foremost public health priorities.[1]

In 2019, the top six bacterial pathogens responsible for nearly a million deaths were *Escherichia coli* followed by *Staphylococcus aureus*, *Klebsiella pneumoniae*, *Streptococcus pneumoniae*, *Acinetobacter baumannii*, and *Pseudomonas aeruginosa*.[3] Extending this to the top six drug-pathogen combinations, Methicillin-resistant *Staphylococcus aureus* was responsible for over 100,000 deaths alone followed by multidrug resistant *M. tuberculosis*, third-generation cephalosporin-resistant *Escherichia coli*, carbapenem-resistant *Acinetobacter baumannii*, fluoroquinolone-resistant *Escherichia coli*, carbapenem-resistant *Klebsiella pneumoniae*, and third-generation cephalosporin-resistant *Kleb-*

siella pneumoniae[3] **(Figure 1)**. Uptake of antiretroviral therapy (ART) to treat Human Immunodeficiency Virus (HIV) has been considered a huge success with 26 million people receiving ART at the end of June 2020.[4] However, emergence of drug-resistant HIV has compromised most antiretroviral drugs, including newer ones, which are now at risk of becoming unusable due to resistance development. More than 10 developing countries had existing drug-resistant HIV.[1] Similarly, the globally emerging drug-resistant yeast (unicellular fungi), *Candida auris*, with known outbreaks in healthcare settings, has already been reported to have widespread resistance to fluconazole, amphotericin B and voriconazole, with emerging resistance to caspofungin.[1] Additionally, malaria is a life-threatening disease, with 241 million cases and 627,000 deaths reported worldwide in 2020.[5] Resistance to artemisin-based combination therapy (ACT), the principal first-line treatment for *Plasmodium falciparum* (one of the main species that causes malaria) has been confirmed in the Greater Mekong Region from studies conducted between 2001 and 2019 **(Figure 1)**.[1]

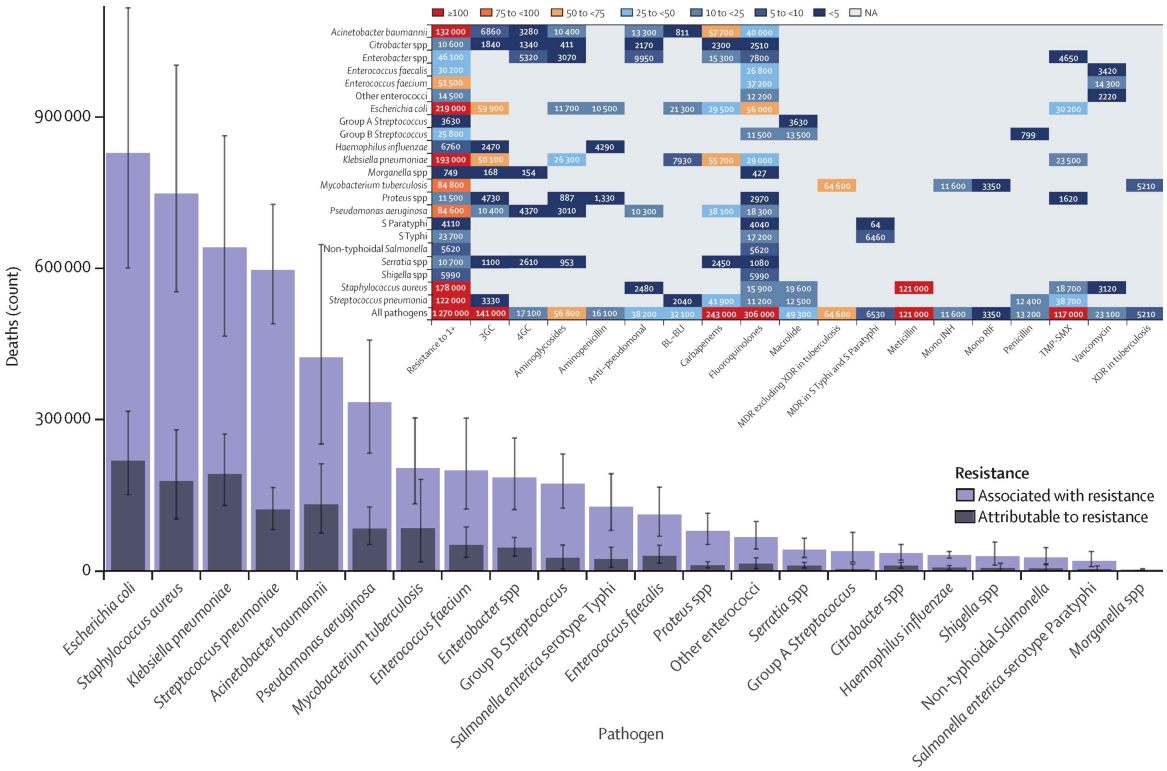
In recognition of this multi-faceted and multi-sectorial problem, the 2020 World Antimicrobial Awareness Week (held every year from 18$^{\text{th}}$ to 24$^{\text{th}}$ November since 2015) was renamed to reflect its expanded scope from antibiotics to antimicrobials. This highlights the broad reach of AMR and the need to address it from a One Health perspective.[9] To put it succinctly, AMR is present and rapidly spreading in every country and affects everyone.

1.1.2 Drivers

The disease burden of AMR has been accelerated by the overuse and misuse of antimicrobials in human health, animal husbandry, and agricultural industries. Poor diagnostic and prescribing practices (antibiotics for viral infections),[10,11] and patient non-compliance to treatment further contribute to this problem. Furthermore, inadequate infection prevention and control measures, especially in resource limited settings with poor sanitation and access to clean water further aggravate the spread and emergence of drug-resistance. This burden is further compounded by a lack of market incentives for antimicrobial drug development due to high costs with poor commercial returns.[2]

The biological drivers of AMR for pathogens can be intrinsic and acquired. While intrinsic mechanisms are comprised mainly of natural barriers present in microbes such as the lipid rich, hydrophobic cell wall of *Mycobacterium* making it naturally resistant to a wide array of antibiotics,[12] the presence of an additional outer membrane in Gram negative bacteria making these naturally resistant to antibiotics like vancomycin targeting cell wall synthesis,[13] the co-evolution of environmental microbes in the presence of a wide variety of variable compounds also contributes to this route of resistance.[14] Acquired drug resistance, however, is predominantly driven by genetic mutations: missense point mutations or

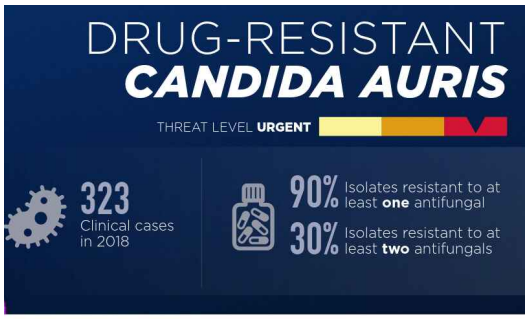
A



B

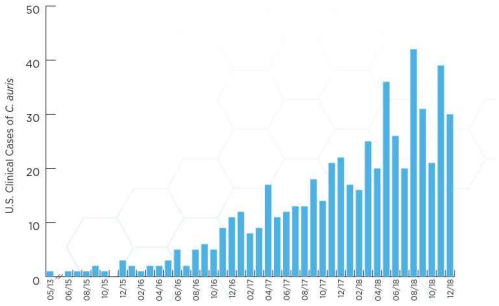
	AIDS deaths		New HIV infections		ART costs	
	2016-2020	2016-2030	2016-2020	2016-2030	2016-2020	2016-2030
Amount attributable to HIVDR	135 000	890 000	105 000	450 000	US$ 0.65 billion	US$ 6.5 billion
Percentage attributable to HIVDR	5.7%	16%	3.5%	8.7%	2.0%	7.7%

C



CASES OVER TIME

C. auris began spreading in the United States in 2015. Reported cases increased 318% in 2018 when compared to the average number of cases reported in 2015 to 2017.



D

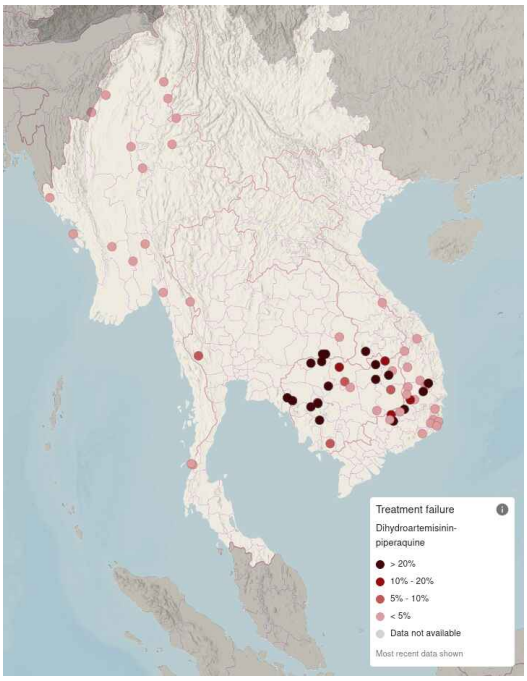


Figure 1: Global Antimicrobial Resistance (AMR) burden
A) Global number of deaths by pathogen in 2019 highlighting *Escherichia coli* as the leading pathogen linked to deaths, with inset showing global deaths by pathogen-drug combination attributed to AMR, highlighting Methicillin resistant *Staphylococcus aureus* linked to over 100,000 deaths. Figure adapted from the 2021 Lancet review on bacterial AMR,[3] **B)** Estimated impact of HIV drug resistance on AIDS deaths, new HIV infections and Antiretroviral therapy in sub-Saharan Africa highlighting pre-treatment HIV drug resistance of more than 10%. Figure adapted from the WHO HIV Drug Resistance report 2021,[6] **C)** Number of clinical cases of *Candida auris* in 2018 in the United States highlighting the steep increase in that year, as well as growing resistance to one or more antifungals. Figure adapted from the 2019 CDC fact sheet on Drug-Resistant Candida species[7] and, **D)** Treatment failure in *Plasmodium falciparum* malaria due to artemisin-based combination therapy between 2010-2020 in the Greater Mekong region. Figure generated from the WHO Malaria threat map.[8]

non-synonymous single nucleotide variations (nsSNVs), insertions/deletions (INDELs), and frameshift mutations. The two different routes of AMR are elaborated in the review published as part of this project in 2020 (included below in full). Please note that SNVs may be referred to as single nucleotide polymorphisms (SNPs), though these terms are strictly speaking not interchangeable. SNP is a type of substitution mutation/variation that must be present in at least 1% of the population to qualify, while a SNV a variation in a single nucleotide without any limitations of frequency.

1.1.3 Computational approaches in studying AMR

The increased use of rapid molecular and genetic testing methods in clinical care is facilitated by computational approaches utilising the wealth of data generated using large scale sequencing technologies.[15–20] In the context of drug resistance, whole genome sequencing followed by Genome Wide Association Studies (GWAS) help identify mutations associated with resistance, in specific gene loci or pan-genome. This contributes directly towards understanding the emergence, development, and spread of AMR. As such, they have become important decision support tools in medicine and public health, where they help bridge gaps in existing knowledge and inform future clinical research.[20–23] Additionally, the abundance of data has opened avenues for novel applications of artificial intelligence and machine learning (AI/ML) methods to combat AMR.[24–28]

Mutations play a fundamental role in evolution and in creating the diversity we see around us. The most common in bacteria, and of particular interest are nsSNVs, resulting in a single amino acid variation (SAV) in the encoded protein sequence. These in turn lead to changes in the protein's three-dimensional structure (3D), influencing local conformational changes and associated interactions.

1.1.3.1 3D structure to understand AMR

Protein structure modelling enables investigation of the biophysical effects of polymorphisms. Biophysical effects include changes in: protein stability, molecular binding affinity of ligands, and protein-

protein interactions upon mutation. Considerable advancement has been made in its application through the use of molecular dynamics simulations[29] and structure-based ML approaches to understand and predict resistance.[25,30] Protein structure is stabilised by physical interactions. The impact of SAVs inevitably alter this stability.[31] The impact of protein stability has been studied by site directed mutagenesis followed by thermodynamic measurements and structure determination.[31,32] Predicting the stability impact of SAVs computationally has numerous advantages over conducting mutagenesis experiments. Scaling up computational capacity is rapid, affordable, and does not take up valuable lab time. Computational investigations are also repeatable and allow alteration of multiple variables.

Using thermodynamic modelling, mutational effects can be assessed through quantitative measurements, reflecting changes made to the thermal stability of a two-state protein. These measurements calculate the difference in Gibbs free energy between the concentration of the unfolded (G_u) and the concentration of the folded (G_f) states, where $\Delta G = G_u - G_f$, as per the equation:

$$\Delta G = -RTln\frac{[folded]}{[unfolded]}$$

Where ΔG is change in Gibbs free energy, R is the gas constant (1.987 cal K^{-1} mol^{-1}, T is the temperature in Kelvin, and [folded] and [unfolded] refer to concentrations of the two forms of protein.

A higher concentration of the protein in the folded form relates to a more stable protein, i.e. a more negative ΔG, as thermodynamically a negative ΔG indicates release of energy by a system to achieve a more stable state. The impact of mutations on protein stability is then calculated as a free energy difference: $\Delta\Delta G = \Delta G_w - \Delta G_m$, where ΔG_w and ΔG_m refer to the free energy change (ΔG) between the unfolded and folded states of the wild-type and mutant proteins respectively. In this manner, a negative $\Delta\Delta G$ indicates that the mutation has destabilising effect, while conversely, a positive $\Delta\Delta G$ indicates a stabilising effect **(Figure 2)**.[33] There is currently no consensus in the literature regarding the calculation of $\Delta\Delta G$, as either $\Delta G_w - \Delta G_m$ or $\Delta G_m - \Delta G_w$ may be used leading to computational tools varying in their approaches for classifying mutational impact. However, many computational tools appear to prefer the use of a negative $\Delta\Delta G$ to indicate a destabilising mutational effect, and as such, in my thesis, I have followed this convention.

Attempts have also been made to predict the impact on affinity of protein-protein, protein-ligand and protein-nucleic acid interactions.[34–38] Ready availability and access to thermodynamic databases for proteins, mutants and interactions like ProTherm and ProNIT,[39] SKEMPI,[40] and Platinum[41] have

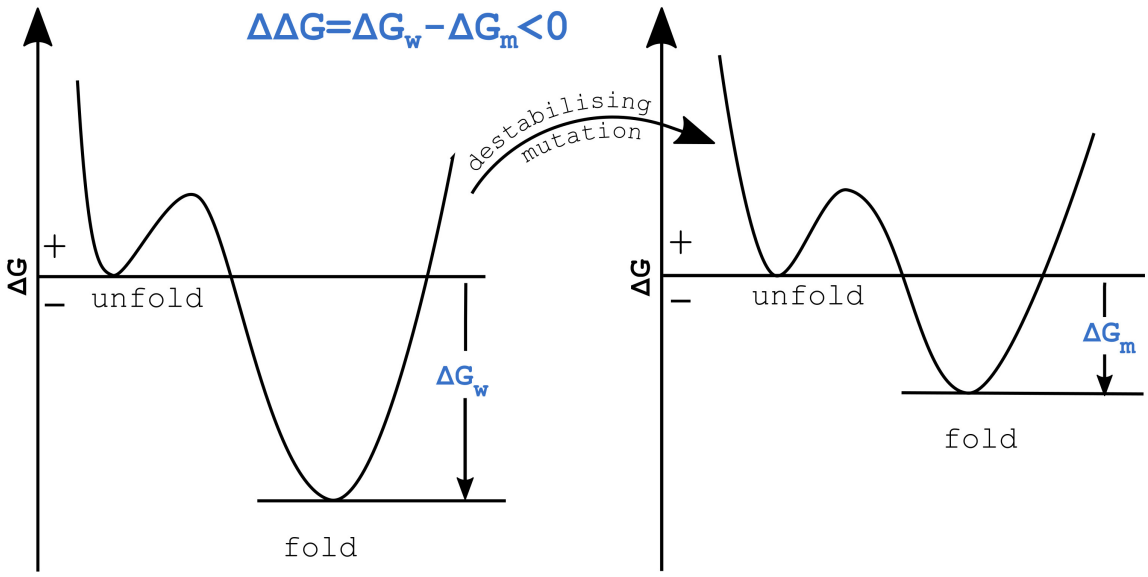


Figure 2: Stability change upon mutation in terms of Gibbs free energy (ΔG)
Figure adapted from Quan, *et. al.*.[33]

been critical in such analyses.

ML combined with structure-based methods has proven powerful in predicting disease related mutations,[42] novel resistance mutations,[25] and predicting stability changes from sequence or structure descriptors.[43,44]

1.1.3.2 Combining 3D structure and genomics to understand AMR

Combining genomic analyses with the biophysical effects of mutations can help reveal the molecular basis and consequences of resistance development. In the review paper, published as part of this project[45] (available in full at the end of this chapter) the application of such a mechanistic understanding of drug resistance to limit the impact of AMR is described. The paper also provides a review of available computational tools to investigate the effects of SAVs on protein structure and function.

As highlighted earlier, mutations leading to drug resistance can occur both outside (e.g. post-translational modifications, regulatory elements, etc.) and in the protein coding region of a pathogen genome. The latter, as a major route to resistance is the focus of this project. A prominent example where SAV driven resistance development is particularly extensive is *Mycobacterium tuberculosis* (*M. tuberculosis*) due to the absence of horizontal gene transfer (HGT). *M. tuberculosis* is the causative agent of human tuberculosis (TB) disease which continues to remain a global concern due to widespread resistance development. The availability of genomics data related to clinical isolates, together with computational tools and 3D protein structures used to investigate mutational impact, motivated a combined genomics, protein structure and ML driven approach to improve understanding

of resistance development in *M. tuberculosis*.

1.2 Tuberculosis

Tuberculosis (TB) is an ancient, communicable respiratory disease caused by the bacterium *M. tuberculosis*. TB most frequently affects the lungs (pulmonary TB), but it can also spread to other parts of the body such as lymph nodes and brain. The TB causing bacteria can remain dormant in humans for weeks to years before becoming active and causing infection. Only a small proportion of those infected with TB develop an active disease during their lifetime, though the risk is greatly increased for those living with HIV, diabetes, and other risk factors like under nutrition, smoking, and alcohol consumption. TB is a global disease and occurs in all age groups, though the majority of those infected are adult males.[46] TB was the leading cause of death from a single infectious agent ranking above HIV up until the COVID-19 pandemic. Worldwide there were approximately 10 million incident TB cases in 2020 with 1.5 million deaths from TB including 214,000 people co-infected with HIV **(Figures 3A** and **3B)**.[46] The burden of TB is disproportionately high in low- and middle-income countries which account for 98% of all reported TB cases **(Figure 3C)**. Despite an 11% decline in global TB incidence between 2015-2020, and TB being a preventable and curable disease, TB treatment is suffering from widespread drug resistance.[46]

1.2.1 Diagnosis

TB diagnosis can be a time consuming and challenging process as *M. tuberculosis* is an extremely slow growing bacteria doubling roughly once per day compared with every 20 minutes in the case of *Escherichia coli*. While smear microscopy and microbiological culture remain the reference standard for TB diagnosis, rapid molecular testing endorsed initially in 2010 by the WHO has become the recommended initial diagnostic test in people with suspected TB. These genetic diagnostic tests have considerable time advantage and higher sensitivity and specificity proving crucial in the early detection of the disease and drug resistance.

1.2.2 Treatment

As with its diagnosis, TB treatment is difficult with long treatment regimens resulting in problems with patient treatment compliance. Without treatment TB mortality remains high, 70% of people with sputum smear positive, and 20% of those with culture positive smear negative pulmonary TB died from TB within 10 years of diagnosis.[47] Anti-TB drugs are classified into several groups. Treatment regimens comprise of standardised fixed dose therapy with a combination of antibiotics from

different drug groups for several months. The broad classification of anti-TB treatment is: First-line drugs isoniazid, rifampicin, ethambutol and pyrazinamide; second-line drugs streptomycin, kanamycin, amikacin, cycloserine, and capreomycin; orals and/or injectables; and further add on drugs like fluroquinolones. Patients with drug-susceptible TB currently undergo a regimen of four first-line drugs for at least 6 months.[46] Classification of anti-TB drugs is being continually evaluated and revised to reflect priority management of drug resistant cases. While the original guidelines for this date back to the 1996[48] with subsequent revisions, the fundamentals describing therapeutic regimens to manage drug resistant cases was published in the 2006 and 2008 guidelines.[49] Up until 2011, the WHO recognised classification of anti-TB drugs was from group 1-5 in a step-down manner based on class, potency, efficacy and clinical experience.[50] The 2016 WHO guidelines however listed drugs in groups A-D in a hierarchical manner[51] with group 1 drugs from the 2011 classification losing priority and being assigned to group D1 with further re-classifications being proposed in light of growing evidence.[52,53] Furthermore, a more up-to-date classification is provided in the most recent WHO consolidated guidelines[54] **(Table 1A)**.

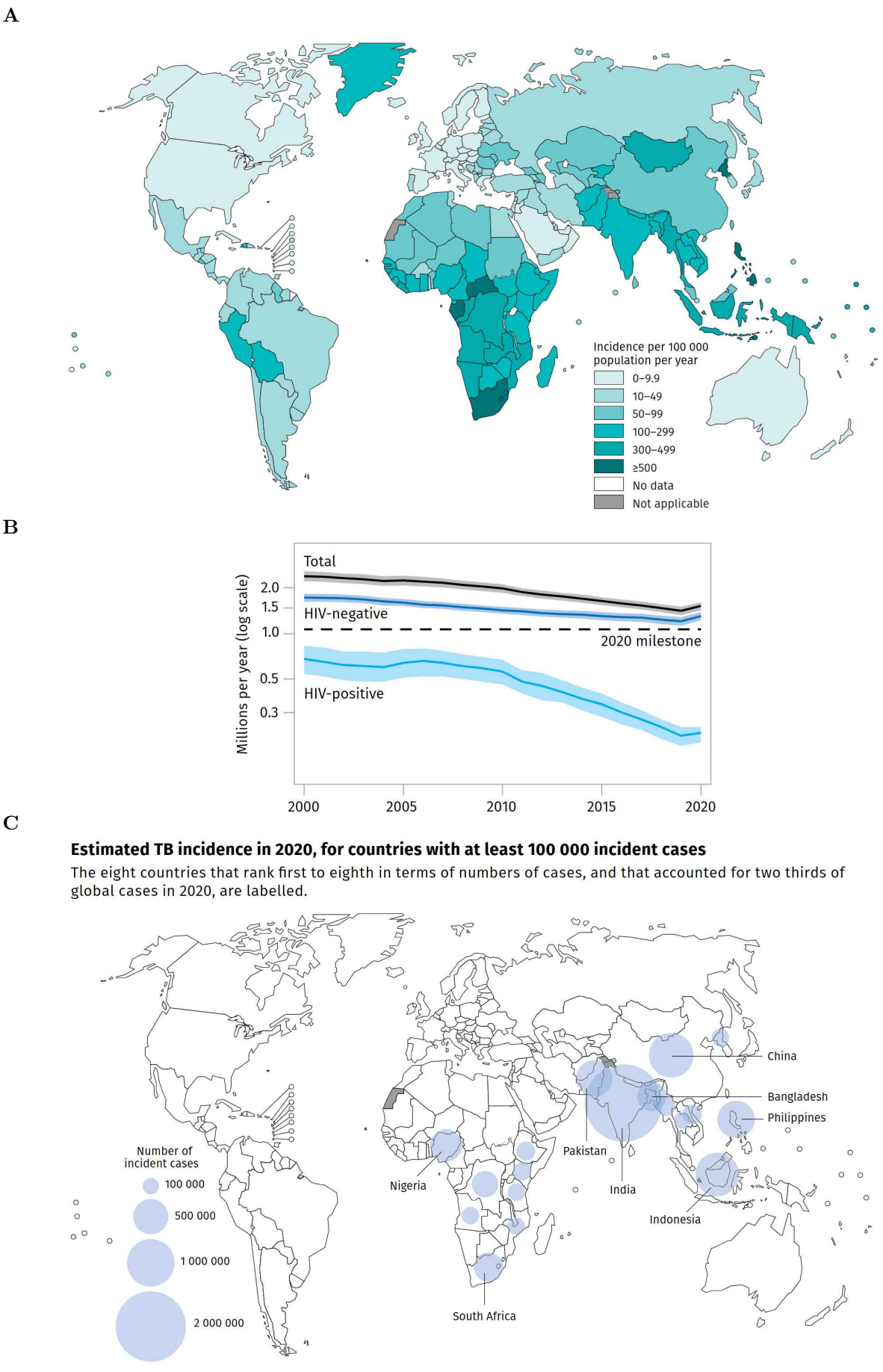


Figure 3: TB incidence, mortality and high burden countries in 2020
A) Estimated TB incidence, **B)** Estimated global trends in overall TB mortality including those with and without HIV, and **C)** Top eight ranked countries with at least 100,000 incident TB cases. Figure adapted from the Global Tuberculosis report 2021.[46]

A: 2011 WHO TB drugs classification

Group	Medicine
Group 1 First-Line oral anti-TB drugs	Isoniazid
	Rifampicin
	Ethambutol
	Pyrazinamide
Group 2 Injectable anti-TB drugs (injectable or parenteral agents)	Streptomycin
	Kanamycin
	Amikacin
	Capreomycin
Group 3 Fluoroquinolones	Levofloxacin
	Moxifloxacin
	Gatifloxacin
	Ofloxacin
Group 4 Oral bacteriostatic second-line anti-TB drugs	Ethionamide/prothionamide
	Cycloserine/terizidone
	p-Aminosalicylic acid
Group 5 Anti-TB drugs with limited data on efficacy and long term safety in the treatment of drug-resistant TB	Linezolid
	Clofazimine
	Amoxicillin/clavulanate
	Imipenem/cilastatin
	Meropenem
	High-dose isoniazid
	Thioacetazone
	Clarithromycin

B: 2016 WHO TB drugs classification for MDR-TB

Group	Classification	Medicine
A	Fluoroquinolones	Levofloxacin
		Moxifloxacin
		Gatifloxacin
B	Second-line injectable agents	Amikacin
		Capreomycin
		Kanamycin
		Streptomycin
C	Other core second-line agents	Ethionamide or prothionamide
		Cycloserine or terizidone
		Linezolid
		Clofazimine
D	Add-on agents	
D1		Pyrazinamide
		Ethambutol
		High-dose isoniazid
D2		Bedaquiline
		Delamanid
D3		*Para*-Aminosalicylic acid
		Imipenem/cilastatin
		Meropenem
		Amoxicillin/clavulanate

C: 2019 WHO TB drugs classification for MDR-TB

Group	Medicine	Step
A	Levofloxacin or Moxifloxacin	Include all three medicines (unless they cannot be used)
	Bedaquiline	
	Linezolid	
B	Clofazimine	Add one or both medicines (unless they cannot be used)
	Cycloserine or terizidone	
C	Ethambutol	Add to complete a four-to-five drug regimen when medicines from groups A and B cannot be used
	Delamanid	
	Pyrazinamide	
	Imipenem/cilastatin or meropenem	
	Amikacin or streptomycin	
	Ethionamide or prothionamide	
	Para-Aminosalicylic acid	

Table 1: Summary and comparison of WHO TB drug classification from 2011, 2016 and 2019
Anti-TB drug classification from (A) 2011,[50] (B) 2016,[51] and (C) 2019,[54] highlighting the shift in treatment focus from managing all cases of TB to managing drug resistant TB cases. Tables adapted from WHO references. MDR-TB refers to multidrug resistant TB, defined as resistance to both isoniazid and rifampicin.

1.2.3 Drug Resistant TB

Treatment for drug resistant TB (DR-TB) is challenging for patients due to long treatment times, and the economy, due to costs and lost productivity. Treatment for DR-TB is expensive, typically costing more than 1000 USD per person.[46] Patients also suffer greater side effects compared with the first line treatments used for drug susceptible TB. DR-TB takes several forms: Pre-MDR, MDR, Pre-XDR and XDR TB.

Multidrug-resistant tuberculosis (MDR-TB) is defined as resistance to two first line anti-TB drugs: isoniazid and rifampicin, where pre-MDR TB refers to resistance to isoniazid or rifampicin, the latter being referred to as rifampicin resistance TB (RR-TB). Both MDR and RR TB require treatment with second line drugs. Although MDR-TB can be treated with second-line drugs, these options are often limited and require prolonged treatment times (up to 2 years) with associated health and economic effects. In some cases, extended resistance to additional drugs referred to as extensively drug resistant TB (XDR-TB) can lead to more severe forms of the disease, further exacerbating the situation. The definition of XDR-TB up until 2021 was MDR/RR-TB strains that were further resistant to second-line-injectables (amikacin, capreomycin or kanamycin) and any fluoroquinolones (such as levofloxacin or moxifloxacin), while pre XDR-TB was defined as the MDR/RR-TB strains which were resistant to second-line-injectables or fluoroquinolones[55] **(Table 1B)**.

The revised 2021 definition of XDR-TB is MDR/RR-TB strains with resistance to any fluoroquinolones and at least one additional Group A drug (Group A drugs are the most potent group of drugs in the ranking of second-line medicines in TB treatment), while pre-XDR TB refers to MDR/RR-TB strains that are resistant to any fluoroquinolone. These revised definitions reflect the severity of disease progression with resistance to additional medicines, thus further limiting available treatment options **(Table 1C)**. Since the data for this project predates the revised definition of XDR-TB, the old definition of XDR-TB is used in this project.

Globally in 2020, roughly 70% of people with confirmed pulmonary TB were tested for RR-TB. A total of 157,903 cases of DR-TB were reported, with 132,222 cases of MDR/RR-TB, and 25,681 cases of pre-XDR- or XDR-TB.[46] In 2019 there were an estimated 465,000 incident cases of RR-TB worldwide, with 78% resulting in an estimated 182,000 deaths.[56] An online dashboard showing TB profile data globally and by region is available at: `https://worldhealthorg.shinyapps.io/tb_profiles`

In 2020, WHO recommended a new shorter (9-11 months) and fully-oral regimen for treating MDR-TB in an effort to improve patient compliance. This is only suitable with exclusion of prior resistance to fluoroquinolones. By the end of 2020, 65 countries were using the shorter MDR-TB treatment recommendation, and 109 countries were using bedaquiline in order to treat MDR-TB **(Figures 4A** and **4B)**.

The estimates in 2020 for TB were based on new methods due a sharp decline of 18% with TB in the preceding year. This was solely due to the COVID-19 pandemic and its effects on access and delivery of TB treatment.[46]

1.2.4 Drivers of TB resistance and evolution

The *M. tuberculosis* bacterium is part of the larger *Mycobacterium tuberculosis* complex, consisting of genetically related *Mycobacterium* species responsible for causing human and animal tuberculosis disease. The bacterium is an intracellular pathogen, unique in its lipid rich cell wall consisting of mycolic acid and glycolipids that play fundamental roles in its virulence.[58] The *M. tuberculosis* genome is remarkably conserved with no horizontal gene transfer (HGT) as observed in other organisms.[59,60] The size of the genome (H37Rv strain) is 4.4 Mb, with a high (65%) main lineages spread globally, L1: Indo-Oceanic, L2: East Asian, L3: East-Africa-Indian, and L4: Euro-American.[22] The lineages are further classified into ancient (L1, L56), modern (L24), and intermediate (L7) strains, with evidence of L2 being particularly mobile due to its recent spread to Europe and Africa from Asia.[22] As well as being globally diverse, *M. tuberculosis* lineages also differ in their virulence, tendency to acquire drug resistance, and biological fitness.[61–63]

TB treatment, though effective, has suffered due to HIV co-infection, immigration, and drug resistance leading to disease re-emergence.[64–67] DR-TB directly threatens disease control and outcome since diagnosis of DR-TB is difficult. Microbiological culture, which takes several weeks to grow remains the gold standard for confirmatory TB diagnosis, making treatment empirical (due to the need to start treatment based on clinical experience, best practice and clinical guidelines before confirmatory results). This adds potential for misdiagnosis, delays, and sub optimal treatment, which all contribute to DR TB.[16,68]

Resistance development in *M. tuberculosis* is an interplay of intrinsic and extrinsic factors. While epigenetic changes and post transcriptional modifications (PTMs) drive the phenotypic route to resistance in *M. tuberculosis*,[69,70] the genetic route to resistance is chiefly acquired via accumulation of mutations in the absence of HGT.

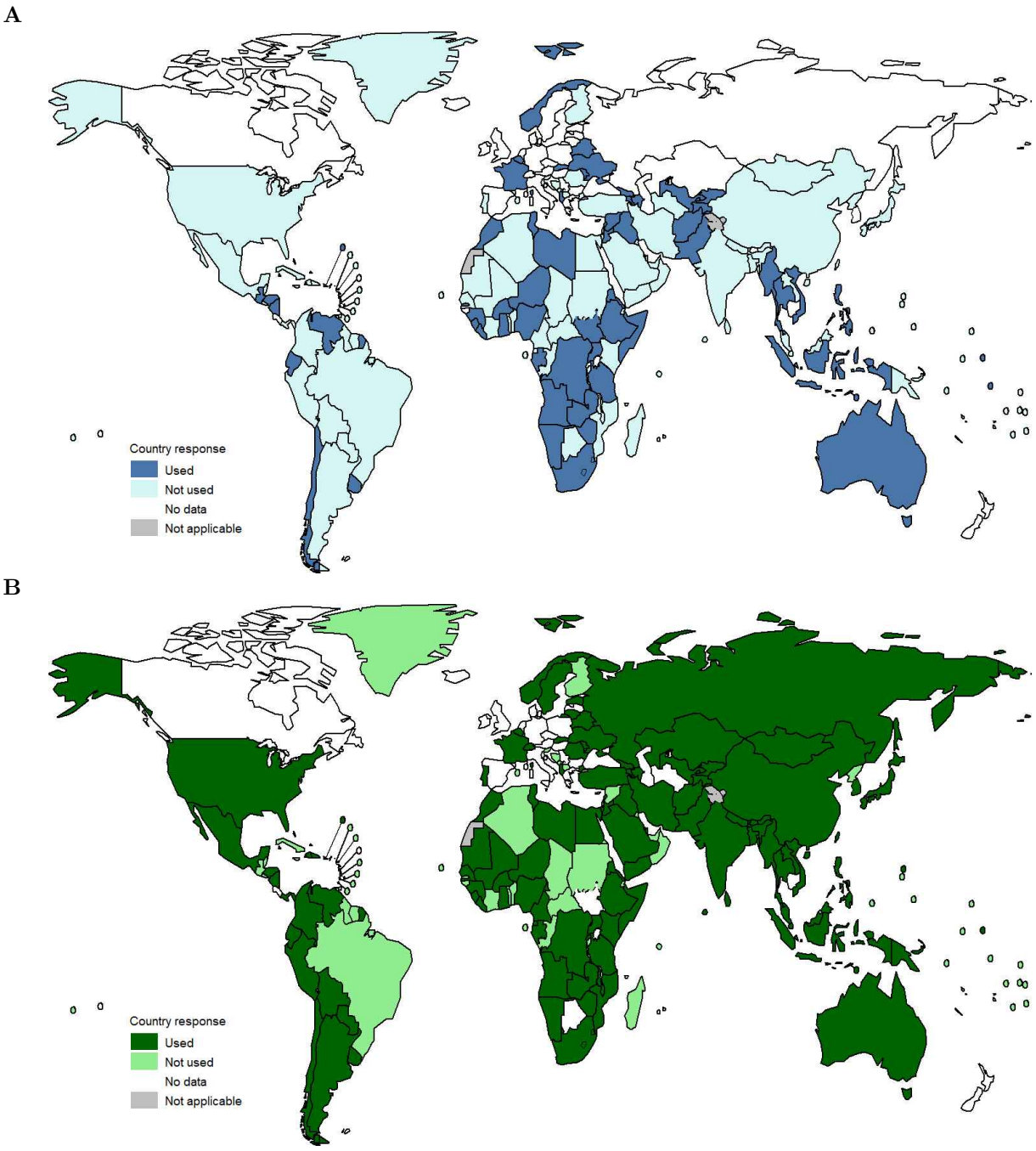


Figure 4: WHO Drug Resistant TB treatment coverage for Multi-Drug Resistant TB (MDR)/Extensively Drug Resistant (XDR) TB in 2020. **A)** countries that used all-oral shorter MDR-TB treatment regimens, and **B)** Countries which used bedaquiline for the treatment of MDR/XDR-TB. Figure adapted from the digital publication section of the Global Tuberculosis report 2021.[57]

Extrinsic factors are commonly social: patient non-compliance owing to long treatment regimens, drug toxicity effects, as well as lack of, and access to, new therapies.[64,66] The intrinsic aspect of resistance manifests in various forms like reduction in cell wall permeability, loss of porins, and the type and number of active efflux pumps.[61,71,72] Intrinsic routes can be innate mechanisms in the organism, driven with or without the genetic route. For example, the low permeability of the lipid rich cell wall in *M. tuberculosis* acts as an innate natural barrier to many antibiotics,[73] while specific enzymes are involved in altering cell wall permeability for certain antibiotics.[74,75] Expression of a

second class of transpeptidases forming non-classical linkages between peptides in the *M. tuberculosis* cell wall confer resistance to β-lactam antibiotics such as amoxicillin and carbapenems.[76,77] The loss of porins in mycobacteria have shown to confer resistance to hydrophilic antibiotics which use these channels to permeate the outer membrane due to the limited permeability offered by their lipid rich membrane. Mutational changes affecting efflux pump activity alter their ability to transport antibiotics out of the cell, and as such have been responsible for the emergence of resistance[78–80] particularly to isoniazid, ethambutol and streptomycin.[61] Methylation of drug targets is yet another route for intrinsic resistance, and has been observed in macrolide resistance,[81] resistance to capreomycin and viomycin.[82]

The intrinsic route, however, is largely driven genetically through mutations such as nsSNVs (leading to SAVs) and INDELs accumulating in the genes coding for drug targets, drug activating enzymes, or including efflux pump activities. Resistance-associated point mutations, specifically SAVs, have been described for all first-line drugs, and for several second-line and newer drugs (fluoroquinolones, bedaquiline).[83–86]

While resistance mutations may bear a fitness cost to the bacterium, putative compensatory mutations allow resistance mutations to become fixed in a population. A classic example of this was demonstrated by whole genome sequence analysis which revealed that mutations in the *rpoA* and *rpoC* gene in rifampin resistant isolates were acting in a compensatory manner to mitigate the fitness loss induced by mutations in the *rpoB* gene in these isolates.[71,87,88]

While resistance development in *M. tuberculosis* has largely been a single step process of acquiring chromosomal mutations, there is now evidence supporting a stepwise accumulation and fixation of mutations in *M. tuberculosis*.[86,89,90] To this effect, the order in which multiple interacting mutations get fixed in a population, leading to a gradual increase in resistance, becomes an important contributing factor towards understanding resistance development. The phenomenon of interaction between genes that influences a phenotype is defined as epistasis, where the effect of a gene mutation depends on the presence or absence of other mutations in one or more genes. This implies that the effect of a mutation then becomes dependent on the genetic background in which it appears,[91] where epistatic mutations result in a different outcome (e.g. resistance) when occurring independently or together. Indeed, there are known conserved patterns like the *katG* S315T mutation for isoniazid resistance, commonly preceding rifampicin resistance irrespective of lineage, geography and time.[92] Epistasis can be positive or negative and can be linked to fitness where positive epistasis would indicate a smaller fitness cost due to multiple interacting genes or mutations.[93] Both forms have been observed

in *M. tuberculosis* between resistance-linked and compensatory mutations.[62,94] As such, epistasis can determine evolutionary trajectories for resistance acquisition.

There are also sub-populations of *M. tuberculosis* that can become phenotypically tolerant to anti-TB drugs without acquiring genetic mutations, termed persisters.[95] These are antibiotic tolerant cells that exhibits arrested growth and low metabolic activity in order to increase drug tolerance, thereby contributing to resistance. Mechanisms of *M. tuberculosis* persistence are not fully understood, with several factors including metabolic traits and physiological states being linked to *M. tuberculosis* persistence.[96,97]

1.3 Project structure

The overall aim of the project is to investigate the mutational impact on protein structures of six genes in *M. tuberculosis*, relating these to the genomics measures including lineage, and using this interdisciplinary approach to develop a gene-agnostic ML-driven resistance prediction tool. The six structural genes analysed in the project are listed below:

1. *alr*-cycloserine (DCS)
2. *embB*-ethambutol (EMB)
3. *gidB*-streptomycin (STR)
4. *katG*-isoniazid (INH)
5. *pncA*-pyrazinamide (PZA)
6. *rpoB*-rifampicin (RFP)

The thesis is divided into 12 chapters, with chapters 3-8 exploring each of the six genes individually with chapter 9 summarising these findings. An overview of the chapters is provided below:

Chapter 1: Introduction includes the review paper published as part of this thesis in October 2020.

Tunstall T, Portelli S, Phelan J, Clark TG, Ascher DB, Furnham N. Combining structure and genomics to understand antimicrobial resistance. Comput Struct Biotechnol J. 2020 Oct 29;18:3377-3394. doi: 10.1016/j.csbj.2020.10.017. PMID: 33294134; PMCID: PMC7683289.

Chapter 2: Methods detailing the dataset used and the *in silico* framework developed.

Chapter 3: Explores the structural and genomic consequences of mutations in the *M. tuberculosis* gene-drug target: *pncA*-PZA. This is a published manuscript from July 2021.

Tunstall T, Phelan J, Eccleston C, Clark TG, Furnham N. Structural and Genomic Insights Into Pyrazinamide Resistance in Mycobacterium tuberculosis Underlie Differences Between Ancient and Modern Lineages. Front Mol Biosci. 2021 Jul 23;8:619403. doi: 10.3389/fmolb.2021.619403. PMID: 34422898; PMCID: PMC8372558.

Chapter 4: Details the structural and genomic relationship for *M. tuberculosis* gene-drug target: *embB*-EMB.

Chapter 5: Details the structural and genomic relationship for *M. tuberculosis* gene-drug target: *gidB*-STR.

Chapter 6-8: Covers three *M. tuberculosis* gene-drug targets: *alr*-DCS, *katG*-INH, and *rpoB*-RFP which updates using our genomic data and analysis tools to what has been previously reported on, with additional genomic and lineage interactions.

Chapter 9: Integrated summary of all six gene-drug targets, which details and discusses the notable findings from chapters 3-8.

Chapter 10: Focusses on mutations that display differing drug susceptibility profiles to assess their relevance in understanding resistance development in *M. tuberculosis*.

Chapter 11: Describes machine learning to anticipate resistance in a gene-target, and using a gene-agnostic approach.

Chapter 12: Discussion, Conclusion and Future work.

London School of Hygiene & Tropical Medicine
Keppel Street, London WC1E 7HT

T: +44 (0)20 7299 4646
F: +44 (0)20 7299 4656
www.lshtm.ac.uk

RESEARCH PAPER COVER SHEET

Please note that a cover sheet must be completed for each research paper included within a thesis.

SECTION A – Student Details

Student ID Number	LSH1806129	**Title**	Mrs
First Name(s)	Tanushree		
Surname/Family Name	Tunstall		
Thesis Title	Using Machine Learning to Anticipate Antimicrobial Resistance in Mycobacterium Tuberculosis		
Primary Supervisor	Nicholas Furnham		

If the Research Paper has previously been published please complete Section B, if not please move to Section C.

SECTION B – Paper already published

Where was the work published?	Computational and Structural Biotechnology Journal		
When was the work published?	October 2020		
If the work was published prior to registration for your research degree, give a brief rationale for its inclusion			
Have you retained the copyright for the work?*	Yes	Was the work subject to academic peer review?	Yes

*If yes, please attach evidence of retention. If no, or if the work is being included in its published format, please attach evidence of permission from the copyright holder (publisher or other author) to include this work.

SECTION C – Prepared for publication, but not yet published

Where is the work intended to be published?	
Please list the paper's authors in the intended authorship order:	
Stage of publication	Choose an item.

SECTION D – Multi-authored work

For multi-authored work, give full details of your role in the research included in the paper and in the preparation of the paper. (Attach a further sheet if necessary)	I wrote the original draft of the manuscript, with initial input on the computational tools from our collaborators (Ascher lab, Prof David Ascher, University of Queensland). However, due to major revisions required as part of the publication, I extended all sections within the manuscript substantially until the final published version. I conducted the analysis within the review, from the GWAS dataset received from Prof Taane Clark's lab (my secondary supervisor), using custom Python scripts to extract, process, and analyse gene-specific data. The protein structure data was provided by our collaborators, but all figures and analysis was performed by myself. I coordinated with co-authors on the response to reviewer comments, as well as any subsequent journal queries, revisions and the final proof.

SECTION E

Student Signature	Tanushree Tunstall
Date	19/07/2022

Supervisor Signature	Nicholas Furnham
Date	15/08/2022

Computational and Structural Biotechnology Journal 18 (2020) 3377–3394

COMPUTATIONAL AND STRUCTURAL BIOTECHNOLOGY JOURNAL

journal homepage: www.elsevier.com/locate/csbj

Combining structure and genomics to understand antimicrobial resistance

Tanushree Tunstall [a], Stephanie Portelli [b,c], Jody Phelan [a], Taane G. Clark [a,d], David B. Ascher [b,c], Nicholas Furnham [a,*]

[a] Department of Infection Biology, London School of Hygiene and Tropical Medicine, Keppel Street, London WC1E 7HT, UK
[b] Computational Biology and Clinical Informatics, Baker Heart and Diabetes Institute, Australia
[c] Structural Biology and Bioinformatics, Department of Biochemistry and Molecular Biology, Bio21 Institute, University of Melbourne, Australia
[d] Department of Infectious Disease Epidemiology, London School of Hygiene and Tropical Medicine, Keppel Street, London WC1E 7HT, UK

ARTICLE INFO

Article history:
Received 18 May 2020
Received in revised form 15 October 2020
Accepted 17 October 2020
Available online 29 October 2020

Keywords:
Structural bioinformatics
Machine learning
Antimicrobial resistance
Tuberculosis
Genome wide association studies
Pathogen surveillance

ABSTRACT

Antimicrobials against bacterial, viral and parasitic pathogens have transformed human and animal health. Nevertheless, their widespread use (and misuse) has led to the emergence of antimicrobial resistance (AMR) which poses a potentially catastrophic threat to public health and animal husbandry. There are several routes, both intrinsic and acquired, by which AMR can develop. One major route is through non-synonymous single nucleotide polymorphisms (nsSNPs) in coding regions. Large scale genomic studies using high-throughput sequencing data have provided powerful new ways to rapidly detect and respond to such genetic mutations linked to AMR. However, these studies are limited in their mechanistic insight. Computational tools can rapidly and inexpensively evaluate the effect of mutations on protein function and evolution. Subsequent insights can then inform experimental studies, and direct existing or new computational methods. Here we review a range of sequence and structure-based computational tools, focussing on tools successfully used to investigate mutational effect on drug targets in clinically important pathogens, particularly *Mycobacterium tuberculosis*. Combining genomic results with the biophysical effects of mutations can help reveal the molecular basis and consequences of resistance development. Furthermore, we summarise how the application of such a mechanistic understanding of drug resistance can be applied to limit the impact of AMR.

Contents

* Corresponding author.
E-mail address: Nick.Furnham@lshtm.ac.uk (N. Furnham).

https://doi.org/10.1016/j.csbj.2020.10.017

1. Introduction

1.1. Antimicrobial resistance (AMR)

Drugs against bacterial, viral and parasitic pathogens have truly revolutionised modern medicine, transforming human health and saving millions of lives. This transformation, however, is under threat due to emerging and widespread resistance to these drugs [1]. This threat is termed antimicrobial resistance (AMR), and is a natural and expected consequence of the Darwinian principle of "survival of the fittest". Almost all antimicrobial drugs have seen resistance arise within 5–10 years of their introduction [2]. The consequences of AMR pose a catastrophic public health threat, responsible for over 700,000 annual deaths [3], prolonged hospital stays, poor disease outcome, less effective treatments, and potentially untreatable diseases. Considering antibiotic resistance alone, the toll is predicted to rise above 10 million deaths per year by 2050 if left unchecked. The associated global economic burden is estimated at 100 trillion USD [3].

The disease burden of AMR has been accelerated by the overuse and misuse of antimicrobials in health, animal and agricultural industries. This burden is further compounded by a lack of market incentives for antimicrobial drug development [3]. Nearly all major infectious diseases are affected by either prevailing or emerging resistance. For example, it is estimated that people with MRSA (Methicillin-Resistant *Staphylococcus aureus*) are 64% more likely to die than people with a non-resistant form of the infection [1]. Similarly, resistance to artemisinin-based combination therapy, the first-line treatment for malaria caused by *Plasmodium falciparum (P. falciparum)*, has been confirmed in 5 countries in the Greater Mekong Region in 2016 [1]. Likewise, in 2010, an estimated 7–15% patients starting antiretroviral therapy (ART) in developing countries had drug-resistant HIV, with up to 40% resistance observed in patients re-starting treatment [1].

Tuberculosis (TB), caused by *Mycobacterium tuberculosis* (*Mtb*), is a major global health problem, with increasing drug resistance making disease control difficult [4]. In 2017, 558,000 cases of rifampicin resistant TB were reported, among which 82% had additional resistance to isoniazid, leading to multidrug-resistant TB (MDR-TB). Among these MDR cases, ~9% cases were further resistant to one fluoroquinolone and one injectable 2nd line drug, leading to extensively drug resistant TB (XDR-TB) [5,6].

Resistance is attributed to multiple factors including selective pressure on *Mtb* from repeated exposure to the same antibiotic, a lack of access to new therapies, and patient non-compliance due to long treatment regimens and drug toxicity effects [7,8]. Both phenotypic and genotypic routes are involved in the development of *Mtb* resistance. While epigenetic changes and post transcriptional modifications drive the phenotypic route to resistance [9,10], the genetic route is chiefly acquired via accumulation of mutations in the absence of horizontal gene transfer. Resistance-associated point mutations have been described across all anti-TB drugs, including newer ones (fluoroquinolones, bedaquiline) [11,12].

1.2. Drivers of AMR

The drivers of AMR can be both intrinsic or acquired. Intrinsic resistance refers to the innate mechanisms present within microbes to combat the action of drugs, and is considered to be independent of previous drug exposure. Intrinsic mechanisms include:

(i) the presence of an additional impermeable outer membrane in Gram negative bacteria making them naturally resistant to antibiotics that target cell wall synthesis such as vancomycin [13].
(ii) the presence of enzymes that either prevent drug binding within an organism, or destroy the drug. An example of the former is the low affinity binding by Gram positive bacteria of penicillin-binding proteins (PBPs) required for the synthesis of peptidoglycan in the cell wall, thus making them naturally resistant to the β-lactam antibiotic aztreonam. An example of the latter is the production of β-lactamase by Gram negative bacteria which destroy β-lactam antibiotics before they can reach their PBP targets [14].
(iii) the presence of multi-drug efflux pumps, which are complex bacterial molecular machines capable of removing drugs and toxic compounds out of the cell. For example, efflux mediated drug resistance in tetracycline is mediated by the Tet efflux pumps which use proton exchange as its energy source to expel the antibiotic [15].
(iv) the lack of enzymes or metabolic pathways in aerobic bacteria to chemically reduce the drug metronidazole to its active form [13].
(v) the co-evolution of microbes with their surroundings containing a variety of toxic and benign molecules and compounds, which is commonly observed in environmental microbes. For example, the soil bacteria actinomycetes harbours an intrinsic 'resistome' to the many antibiotics it produces [16,17].
(vi) the phenomenon of bacterial persistence, notably observed in asymptomatic and chronic infections such as typhoid and TB. Persisters are a sub population of antibiotic tolerant cells that exhibit low metabolic activity and arrested growth, contributing to increased drug tolerance and resistance [18].

Acquired drug resistance is typically driven by genetic variation including point mutations (missense mutations or non-synonymous single nucleotide polymorphisms; nsSNPs) and insertions/deletions (INDELs) such as frameshift mutations. Such muta-

tions can alter drug activation, binding affinity and permeability, efflux pump activity, and biofilm formation [19]. Furthermore, a common and prominent mechanism called horizontal gene transfer (HGT) or lateral gene transfer (LGT) has been a significant cause of widespread drug resistance. HGT/LGT is found almost exclusively in bacteria where resistance conferring genes are transferred between bacterial species [20,21].

Despite the two distinct routes of resistance, intrinsic mechanisms may be driven by adaptive/acquired routes. For example the efficacy of drug efflux pumps in *Mtb* are modulated by SNP mutations [22,23]. The drivers of AMR and the various mechanisms beyond point mutations (which forms the focus of this review) have been extensively reviewed elsewhere: antibiotic resistance [13,14], antifungal resistance [24–26], antiviral resistance [27,28] and antiparasitic drug resistance [29–31].

1.3. Point mutations linked to AMR

A major route to AMR is driven by point mutations. For example, in *Mtb,* mutations in several genes have been associated with resistance to rifampicin (*rpoB),* isoniazid (`katG,` *inhA* and *ahpC),* streptomycin *(gidB, rrs* and *rpsL),* pyrazinamide (*pncA),* ethambutol (*embB)* and fluroquinolone (*gyrA* and *gyrB).* More generally, mutations within *gyrA* confer low level fluroquinolone resistance in Gram negative bacteria, `while` additional mutations in *parC* and *gyrB* are responsible for high level resistance [32]. Ribosomal mutations affecting ribosome assembly are particularly problematic since these lead to large scale transcriptomic and proteomic changes. In *Mycobacterium smegmatis,* such mutations have led to downregulation of KatG catalase (activating enzyme for the drug isoniazid) and upregulation of the transcription factor WhiB7 involved in innate antibiotic resistance. Further, the fitness cost of these mutations is alleviated in a multi-drug environment which promotes the evolution of high-level, target-based resistance [33].

Antiviral resistance is mainly an adaptive process, chiefly driven by mutations [27]. In the case of antiretrovirals used in HIV treatment, the primary mechanism of resistance to most Nucleoside Reverse Transcriptase Inhibitors (NRTI) is through accumulation of mutations near the drug binding site [34]. In Hepatitis B virus, multiple missense point mutations have been linked to several drugs, along with cross resistance observed between drugs [35]. Point mutations in the preS/S region are associated with vaccine failure, immune escape, occult HBV infection and the occurrence of hepatocellular carcinoma (HCC). Similarly, nsSNPS in the preC/C region are related to HBeAg negativity, immune escape, and persistent hepatitis, while those in the X region are implicated in promoting HCC [36]. Likewise, antifungal resistance in *Aspergillus fumigatus* is also primarily driven by mutations in the azole target cyp51A gene [37], while resistance to artemisinin in *P. falciparum* malaria is driven by multiple mutations in the Kelch 13 (K13) propeller protein.

1.4. Genomics to identify point mutations linked to AMR

High throughput genomic platforms methods of next generation sequencing (NGS) technologies such as whole genome sequencing (WGS) and genotyping arrays have enabled large scale investigations of AMR for identifying resistance determining genetic variants such as SNPs, INDELs, copy number variation, and frameshift mutations [38–43]. The role of genetic variants, in particular SNPs, have been implicated in drug resistance by several studies [44–47]. Building on human complex disease applications [48–50], genome-wide association studies (GWASs) have been applied to reveal genotype - AMR phenotype associations, at a locus or variant level. Furthermore, GWAS regression models allow the estimation of mutation or genotype effect sizes (e.g. odds ratios). Examples of GWAS analysis in the context of AMR include for *Burkholderia multivorans* [51], *Mtb* [11,52,53], severe malaria [50] and fungal pathogens [54].

Bioinformatic approaches exploiting output from WGS technologies and GWAS analyses have enabled AMR prediction and surveillance. Leveraging this wealth of information has enabled novel applications of artificial intelligence and machine learning (AI/ML) in the pan-genome identification of resistance genes, pathways, mechanisms [55–58], as well as resistance prediction [59–61]. Bioinformatic approaches have also been used to identify novel drug targets like Inositol-3-phosphate synthase (I3PS) in *Mtb*, opening up new avenues in TB drug discovery [62].

Despite the immense utility provided by genomic analysis, these methods lack the mechanistic underpinning required to develop robust prediction tools [63] necessitating follow-up functional studies [64]. In order to strengthen genomic analysis, it is important to supplement genomic associations with functional consequences of mutations on drug targets. One of the ways to achieve this is via biophysical assessment of mutations on drug-target structure and their interactions.

1.5. Biophysical consequences of point mutations on protein structure

The biophysical consequences of protein mutations are mainly studied by assessing thermodynamic stability, which is often used as a proxy for function [65]. This relationship has been clearly demonstrated in the evolution of influenza nucleoprotein which appears to be constrained to avoid low-stability sequences [66]. The synergy between the fields of protein biophysics and protein evolution helps contextualise and rationalise concepts of thermodynamic stability, mutational robustness, evolvability and epistasis in resistance development [67–69]. Missense mutations resulting in a change in the amino acid may disrupt downstream function by altering protein stability and its associated interactions [70]. For example, three missense point mutations within the *Mtb gidB* gene lead to gidB mutants with lower thermodynamic stability and higher flexibility, considered to be a major driving factor in the emergence of high-level streptomycin resistance [71]. Equally, structural insights into the stability-function relationship have highlighted the rationale for such a trade-off in the development of antibiotic resistance [72].

1.6. Using structure to understand impact of point mutations linked to AMR

Structural consequences of point mutations can provide functional insights for resistance phenotypes. For example, point mutations in the Penicillin-Binding Proteins confer resistance to β-lactam antibiotics by making the active site amenable to hydrolysis, or reducing binding affinity for the antibiotic [73]. Structure guided design demonstrated the potential of boronate-based PBP inhibitors to overcome β-lactam resistance in Gram positive organisms [74]. Similarly, missense mutations in the *Mtb gidB* gene (target for the antibiotic streptomycin) are responsible for drug resistance through distortion of the binding pocket affecting SAM (co-factor) binding [71]. Likewise, mutations in *Mtb pncA* gene (target for the pro-drug pyrazinamide) are responsible for the loss of enzyme activity [75]. The underlying mechanism of mutations in the *gidB* gene conferring low and high-level streptomycin resistance in *Mtb* were found to be associated with distortion in the active site morphology by proximal and distal residues affecting the overall structure [76]. Further, the prominent mutation H275Y within the neuraminidase enzyme of the H1N1 pandemic strain renders the drug oseltamivir ineffective due to distortion in the binding pose of the drug within the active site [77]. Structural analysis of C580Y and R539T mutations in the *K13 propeller*

gene (associated with artemisinin resistance) in *P. falciparum* malaria revealed local conformational disruption in the mutant and two solvent-exposed patches at conserved sites affecting protein–protein interactions [78].

Structural insights can aid in the absence of phenotypic data [79] as well as provide a physical basis to a more comprehensive understanding of mutational impact on the underlying biological mechanisms. Therefore, computational tools measuring the biophysical effects of resistance linked mutations can aid mechanistic understanding and inform functional studies. Understanding mutational consequences with respect to global (drug-target structure) and local (protein–ligand, protein–protein and protein-nucleic acid) stability effects [80] can be further extended to predict drug resistance for novel mutations [81,82].

Here, we review several of the principal computational tools and methods currently available for measuring mutational consequences, focusing on those tools which have been used to analyse variation within a pathogen genome and their application in the context of AMR. It is not meant to be an exhaustive list, with other tools available centred on important questions like assessing cancer variations and other human mutations. As such, these go beyond the scope of this review and have been extensively reviewed elsewhere [83–85].

2. Computational tools measuring the effect of mutations

While no general pre-emptive predictor for AMR has been developed, we and others have shown that computational tools for understanding the underlying molecular mechanisms of mutations can be used to identify likely resistant variants [79–82,86–95]. This insight has even been used to guide medicinal chemistry design of inhibitors less prone to resistance [96–99].

Different tools can be used to describe the effect of mutation on protein function, which may provide an explanation for the AMR phenotype. Some are primarily based on conservation or substitution matrices, and do not require a protein structure as input (Sequence-based methods). Others consider the local environment of the variant within the protein structure in their calculation (Structure-based methods). In the presence of a known AMR-related phenotype, these tools are useful as they provide mechanistic insight which may explain how resistance is brought about at the protein level. Therefore, when analysing specific proteins, it can be beneficial to use different methodologies, as different strategies may give complementary information. Summaries of the types of methods are given below and represent some of the principal tools currently available. Table 1 summarises the main features of some of the currently available tools for analysing effects of pathogen mutations.

2.1. Sequence-based methods

As these methods rely solely on the gene or protein sequence, they are often useful in the absence of a known protein structure or when homology modelling is not possible. The predictions from these tools are generally based on sequence alignments, predicted secondary structures and subsequent conservational trends. Most methods determine a score with cut-offs leading to functional classification of mutations into deleterious or neutral. This functional classification is not always applicable to AMR mutations, as variants may be 'deleterious' to protein conservation, but gain-of-function through survival in the presence of drug. For example, when analysing rifampicin resistant *Mtb* mutations we found that they tended to cluster within more conserved regions of the *rpoB* gene [80] (Portelli and Ascher, personal communication). Similar analysis carried out on pyrazinamide [82] and bedaquiline [81], revealed that known resistant *Mtb* mutations were more likely to lead to deleterious effects compared to susceptible variants in the same gene [100]. However, when measuring mutational tolerance [101], strong evidence of positive selection for resistant mutations was observed. Therefore, the utility of these tools in understanding AMR mechanisms lies in the actual scores, where a comparison of different scores across variants, accounting for their genetic position can uncover important underlying mechanisms and trends related to evolutionary conservation. We have previously shown that this sequence information is also complementary to structural information, particularly within the context of machine learning [102]. Several of the major methods which are applicable across pathogens and human genomes are:

a. SIFT

The SIFT (Sorting Intolerant From Tolerant) can be used to analyse missense mutations and INDELs. The SIFT scoring method combines sequence alignment with a position-specific scoring matrix (PSSM), which accounts for the likelihood of an amino acid to occur within a specific position. The amino acid chemical properties are also incorporated to determine a scaled probability of the mutation (SIFT score), on which the output (tolerated or deleterious) is based [100]. SIFT has been used to build the Variant Effect Predictor (VEP) tool developed as part of the Ensembl 2018 project [103].

b. PROVEAN

PROVEAN (Protein Variant Effect Analyzer) is able to account for (multiple) missense mutations and INDELs. It uses the BLOSUM62 substitution matrix as an amino acid probability matrix and combines this with differences in sequence similarity between wild-type and mutant sequences. The sequence context in which variation occurs is also considered, to represent environmental surroundings and effects. A numerical score is generated for each variant, which enables the functional classification into deleterious or neutral [104]. PROVEAN scores have provided the evolutionary basis for the recently deployed web-based tool SUSPECT-PZA [82] which predicts pyrazinamide (PZA) resistance mutations in the *Mtb pncA* gene.

c. SNAP2

SNAP2 (Screening for Non-Acceptable Polymorphisms v.2) characterises the effect of all possible missense mutations as either neutral or deleterious. It is a machine learning-based predictor trained on neural networks. It also accounts for amino acid position probabilities using position-specific independent counts, based on the BLOSUM62 matrix. This predictor considers other features such as protein fold (Pfam, PROSITE) and functional annotations (SWISS-PROT) during training, and as such is the tool that spans the most comprehensive feature space [105]. As well as forming part of the SUSPECT-PZA tool [82], SNAP2 scores have provided the evolutionary basis for a similar tool called SUSPECT-BDQ [81]. This tool predicts the effect of missense mutations on the anti-TB drug bedaquiline, reserved to treat MDR and XDR TB.

d. ConSurf

ConSurf estimates an evolutionary rate score for every position across the sequence, unlike the tools above which base functional classification on score thresholds. In the context of drug resistance, it can help identify sites which are likely to lead to resistance if mutated. The ConSurf score is based on a multiple sequence alignment, which generates probabilistic evolutionary models and phylogenetic links. Through this score, more conserved sites (having slower evolutionary rates), which have important functional and structural consequences are identified [106]. Consurf has been used to estimate and visualise conserved regions within SARS-CoV-2 [107], the SARS-CoV nsp12 polymerase domain [108], and

Table1
Sequence and structure-based tools that predict effect of pathogen missense mutations. The table is an up-to date list of currently available tools (as on 3rd August 2020). The type of method for each tool is specified using the following code; **S**: sequence-based, **St:** structure-based, **SA**: sequence alignment, **SS**: sequence and structure, **(St)**: structure if available. Other abbreviations used: MSA (Multiple Sequence alignment), EC (Evolutionary Conservation), NN (Neural Network), SVM (Support Vector Machine), ML (Machine learning), NMA (Normal Mode Analysis), ΔG: Gibbs free energy in Kcal/mol, ΔΔG: Change in Gibbs free energy in Kcal/mol, wt: wild-type, mt: mutant, K_{wt}: affinity of the wild-type protein-ligand complex, K_{mt}: affinity of the mutant complex, RSA: Relative Solvent Accessibility (%).

Name of tool	Type	Operating Principle	Availability	Summary	User input	Output	User Notes
SIFT: **S**orting **I**ntolerant **F**rom **T**olerant REF: [100]	S	EC	http://sift-dna.org Download: Yes	Calculates a normalised probability of substitution score from multiple alignments based on sequence homology using PSI-BLAST. Removes close homologous sequences to prevent over prediction of "tolerated" substitutions. Mutational effect on protein function is classified as damaging (<=0.05) or tolerated (>0.05).	Fasta sequence or aligned sequences SNP list	Per-SNP: 1) SIFT score 2) Binary mutation classification 3) Median sequence conservation	Predictions for submitted SNPs, as well as all possible SNPs (but without a score). Positions are weighted equally within an alignment. Alignments may be user defined. Sequence conservation score provides a useful estimate of whether the alignment contains sufficient variation to support classification.
PROVEAN: **Pro**tein **V**ariation **E**ffect **An**alyzer REF: [104]	S	EC	http://provean.jcvi.org/seq_submit.php Download: Yes	Related sequences are collected with BLAST (using CD-HIT) and clustered based on 75% global sequence identity. The top 30 clusters of closely related sequences form the supporting sequence set, used to generate the prediction. Delta alignment scores are computed for each supporting sequence and averaged within and across clusters to generate the final PROVEAN score. Predicted mutation effects are classified as either deleterious or neutral based on a predefined threshold (-2.5). Available as: PROVEAN Protein PROVEAN Protein Batch* PROVEAN Genome Variants* **Human and Mouse only*	Fasta sequence Mutation list (SNPs and INDELs)	Per-mutation: 1) PROVEAN score 2) Binary mutation classification	Predictions for submitted mutations only. Predict effects for both SNPs and INDELs, but not frameshift mutations. Batch processing of multiple organisms. The classification threshold is fixed in the online version. Stand-alone package only available for PROVEAN Protein.
SNAP2: **S**creening for **N**on-**A**cceptable **P**olymorphisms, v2 REF: [105]	S	NN	https://www.rostlab.org/services/SNAP/ Download: Yes	Combines evolutionary information with an expanded list of original SNAP features (amino acid properties) including features such as AA index, predicted binding residues and disordered regions, residue annotations from Pfam and PROSITE, etc. Mutations are classified as either neutral or effect based on predicted scores, between (-100 to 100) respectively. The prediction algorithm is based on a NN consisting of a feed-forward multi-layer perceptron. 10-fold cross-validation is used to create 10 models, each providing a single score for each output class (neutral/effect). The final score is calculated as the difference between the average scores for each output class.	Fasta sequence	For all possible substitutions: 1) Heatmap representing the predicted effect 2) Multi column table with Predicted Effect, Score and Accuracy.	Predictions for all possible substitutions. Prediction scores are accompanied by an "accuracy metric" to aid interpretation. Uses predicted structural features. Heatmap generated for visualisation of the predictions. Additional method (SNAP2noali) predicts functional effects without alignments. Automatic selection of best method (SNAP2 by default, and SNAP2noali for orphans) with notification to users.
ConSurf REF: [106]	S(St)	EC	https://consurf.tau.ac.il/ Download: No	Estimates evolutionary conservation rate of amino/nucleic acid positions based on the phylogenetic relations between homologous sequences. Homologous sequences are searched using CSI-BLAST, PSI-BLAST or BLAST, with closely related sequences removed using CD-HIT with multiple sequence alignments (MSA) generated by MAFFT by default.	Amino acid/ nucleotide sequence Structure (if available) MSA (if available) *Advanced options:*	Detailed output containing conservation scores, MSA and BLAST results. Estimates mapped onto sequence and structure.	Analysis at amino acid and nucleotide levels. Improved HMMER algorithm to search for homologous proteins. Results are accompanied by confidence intervals. Robust statistical approach to differentiate between apparent conservation (short evolutionary time) and genuine conservation (purifying selection). 'ConSeq' mode used in the absence of a

(*continued on next page*)

40

Table1 (*continued*)

Name of tool	Type	Operating Principle	Availability	Summary	User input	Output	User Notes
				MSA is used to construct phylogenetic relationships using the neighbour joining method. Position specific evolutionary rates are calculated using the empirical Bayesian or Maximum Likelihood methods. Scores graded 1 (variable) to 9 (conserved) for visualisation.	- homologue database - MSA methods - Phylogenetic tree - structural data - calculation method - evolutionary substitution model Optional: user defined MSA and phylogenetic tree.		structure. Site-specific predictions of the buried/exposed status of each position.
MAPP: **M**ultivariate **A**nalysis of **P**rotein **P**olymorphism REF: [110]	SA	EC	Download only: http://mendel.stanford.edu/SidowLab/downloads/MAPP/index.html	Combines MSA with 6 physicochemical properties for amino acids. Calculates a MAPP impact score for each position within the MSA. Sequences in the MSA are weighted to account for phylogenetic correlation. Physicochemical property scores for each column along with their mean and variances are calculated. The deviation of each property is calculated for every possible variant and converted to a single score.	Fasta format MSA Phylogenetic tree	Multicolumn table giving the physico-chemical characteristics of each position, MAPP impact score, and a listing of "good" and "bad" amino acids.	Predictions for every possible substitution, and median MAPP scores calculated for each position. Constructs a physiochemical profile rather than an amino acid profile. Demonstrates value of using only orthologous protein in creating a conservation profile. Scores are continuous and interpreted in a relative manner with higher MAPP scores indicating more conserved areas. Can be optimised for individual genes including MAPP impact score threshold for classification. Requires user defined MSA and phylogenetic tree.
PANTHER-PSEP: **P**rotein **AN**alysis **Th**rough **E**volutionary **R**elationships-**P**osition **S**pecific **E**volutionary **P**reservation REF: [149]	S	EC	http://www.pantherdb.org/tools/csnpScoreForm.jsp Download: Yes	Uses Hidden Markov Model (HMM) to align sequence to protein families and subfamilies in its database to calculate the evolutionary preservation metric. Uses variation over each alignment position to estimate the likelihood of a coding SNP to cause a functional impact on the protein. Score represents the time (in millions of years [my]) a given amino acid has been preserved in the lineage, directly corresponding to the likelihood of a functional impact. Score classified into: Probably damaging, Possibly damaging, Probably benign.	Fasta sequence SNP list *Other parameters:* Organism	Per-SNP: 1) Preservation Time: PANTHER PSEP score 2) Message: Classification of the PSEP score	Positions are weighted equally at all positions within an alignment. Profiles are subfamily specific if they substantially differ from entire family. User defined alignments are not possible since scores are derived from HMMs (PANTHER protein library) together with an ontology of protein function (PANTHER/X – a simplified form of GO) to make predictions.
FoldX suite REF: [113]	St	Empirical force field	Download only: http://foldxsuite.crg.eu/	Empirical force-field used for calculating mutational effects of stability, folding, and dynamics on proteins and nucleic acids ΔG (free energy of unfolding) is calculated using a combination of empirical terms. Empirical data (derived from protein engineering experiments) is used for weighting energy terms for stability calculations.	PDB file SNP list (including chain ID)	Multiple output files where requested. Main output is present in '*Dif_*" files, containing ΔG of wt and mt residues along with ΔΔG of mutation. Output also contains changes in the associated energy terms.	Command line interface. Creates mutant structure models. Can be used to analyse protein–protein and protein-DNA interactions. Calculates actual stabilities of wt and mt structures, as well as change in stability upon mutation (ΔΔG). Easily integrated into custom workflows.

Table1 (*continued*)

Name of tool	Type	Operating Principle	Availability	Summary	User input	Output	User Notes
				Foldx *BuildModel* command calculates stability changes upon mutation based on a full atomic description of the protein structure. Classification of ΔΔG: ΔΔG > 0: Destabilising ΔΔG < 0: Stabilising			Optimised energy function for faster calculations. Requires registration to download.
PoPMuSic (v2.1): **P**rediction **O**f **P**roteins **MU**tations **S**tab**I**lity **C**hanges REF: [115]	St	Physics-based and NN	https://soft.dezyme.com/ Download: No	Stability change upon mutation calculated using a linear combination of statistical energy potentials, accounting for variation in volume of the mutant residue. Predictive models include an optimised set of 52 parameters, whose values are estimated and optimised using a neural network. ΔΔG of point mutation is calculated by a linear combination of 16 terms: 13 statistical potentials, 2 terms for volume of wt and mut residues, and 1 independent term. Classification of ΔΔG: ΔΔG > 0: Destabilising ΔΔG < 0: Stabilising Additional "optimality" score is assigned for each position in the protein sequence. It indicates poorly optimised positions with potential functional consequences.	Only accepts currently available entries in the PDB SNP list in three input modes: 1) Systematic: all possible point mutations 2) Manual: single mutation 3) File: SNP list	Multi-column table containing secondary structure, solvent accessibility (%) and predicted ΔΔG of mutations.	Optimised to rapidly calculate stability changes of all possible mutations in mid-size proteins. Graphical output of sequence optimality scores. No option to upload user-defined PDB files. Requires registration to download.
I-Mutant (2.0) REF: [116]	S(St)	SVM	http://gpcr2.biocomp.unibo.it/cgi/predictors/I-Mutant3.0/I-Mutant3.0.cgi Download: No	Predicts stability effect of a point mutation, as a classification or regression task. The classification task predicts the direction of change, while the regression estimator predicts the ΔΔ*G*. Can be applied to both sequence and structure. RI value (Reliability Index) is computed from the output of the SVM model. Binary classification ΔΔG: ΔΔG < 0: Decrease Stability ΔΔG > 0: Increase Stability Ternary Classification ΔΔG: Large Decrease of Stability: ΔΔG < -0.5 Large Increase of Stability: ΔΔG > 0.5 Neutral Stability: 0.5<=ΔΔG<=0.5	Fasta sequence or PDB code/file Chain ID Single SNP Temperature PH Prediction request: Binary/ Ternary classification	Table containing: 1) RSA (%) of mt residue 2) RI (Reliability Index) 3) Predicted ΔΔG 3) Classification of ΔΔG	Predicts both the direction and the estimate of stability. Experimental conditions of pH and Temperature (Celsius) are considered in the stability calculations. Analyses a single mutation at a time only. Output on the web server is better than output requested via email. Use of two different SVM models can lead to discordance between the ΔΔ*G* sign and classification, but is stated to occur only in cases of low RI value.
STRUM: **STR**ucture-based prediction of protein stability changes **U**pon single-point **M**utation REF: [117]	S(St)	ML	https://zhanglab.ccmb.med.umich.edu/STRUM/ Download: Yes	Calculates ΔΔG of mutation using gradient boosting regression algorithm trained on 120 features divided into three groups (sequence, threading and structure). Classification of ΔΔG: ΔΔG < 0: Destabilising ΔΔG > 0: Stabilising	Fasta sequence or PDB file SNP list in two modes: 1) Single or multiple SNPs 2) Systematic: All possible SNPs for user defined amino acid segments.	Results available via e-mail only. Multi-column table containing ΔΔG for SNPs.	Combines sequence profiles and 3D features 3D Structure modelling of query protein sequence by iterative threading assembly refinement simulations Computationally expensive with relatively long runtime.
MAESTRO: **M**ulti **A**g**E**nt **ST**ability p**R**edicti**O**n	St	Multi agent: ML methods and	https://pbwww.che.sbg.ac.at/maestro/web	Multi-agent method where 3 ML methods i.e Artificial NN, SVM and Multiple Linear Regression. are combined to generate a consensus prediction.	PDB code/file Input mode: 1) Specific	Input modes 1 & 2 ΔΔG predictions and confidence intervals.	Ability to analyse mutations independently or in combination ΔΔG predictions are accompanied by

(*continued on next page*)

42

Table1 (*continued*)

Name of tool	Type	Operating Principle	Availability	Summary	User input	Output	User Notes
REF: [150]		statistical scoring functions	Download: Yes	Each agent (ML method) uses 9 input values divided into two categories: SSF functions and protein properties (size, mutational environment, etc.). Classification of ΔΔG: ΔΔG > 0: Destabilising ΔΔG < 0: Stabilising	mutations 2) Sensitivity profile: all possible mutations 3) Scan for destabilising mutations 4) Stability of Disulphide bonds	Graphical display.	confidence intervals. High throughput scanning for all possible point mutations. Specific mode for prediction of stabilising disulphide bonds.
mCSM suite: **m**utational **C**ut-Off **S**canning **M**atrix REF: [122]	St	Graph-based and ML	Protein Stability (PS), Protein-Protein (PP), Protein-DNA (P-NA) http://biosig.unimelb.edu.au/mcsm/ Download: No	Uses graph-based methods to calculate atomic pairwise distance surrounding the wt amino acid. Mutational impact is captured based on a change in the atomic pharmocophore count resulting from the point mutation. Together, this forms the mCSM-signature, and is used to train predictive models for analysing mutational impact on structure stability. Predicted ΔΔG < 0 relates to destabilising, and ΔΔG > 0 relates to stabilising mutational effects. Ternary Classification of Destabilising effect: Mild: −1 < ΔΔG < 0 Moderate: −2 < ΔΔG < -1 High: ΔΔG < -2 Ternary Classification of Stabilising effect: Mild: 0 < ΔΔG < 1 Moderate: 1 < ΔΔG < 2 High: ΔΔG > 2	PDB code/file SNP list Chain ID Input modes: 1) Single mutation 2) Mutation list 3) Systematic: all possible mutation for a single residue	Input mode 1: 1) Predicted ΔΔG 2) Classification of mutational stability change Input modes (2) & (3): Multi-column table with predicted ΔΔG and RSA.	Predicts both the direction and the estimate of stability. Mutant structure is not required. webGL structural visualisation for input mode 1. Works at an atomic level. Demonstrates correlation between atomic-distance pattern of the wild-type residue environment and mutational impact. Calculates overall stability of protein and interactions.
mCSM-lig: **m**utational **C**ut-Off **S**canning **M**atrix on **lig**and affinity REF**:** [88]	St	Graph-based and ML	Protein-ligand affinity (mCSM-lig): http://biosig.unimelb.edu.au/mcsm_lig/prediction Download: No	Based on the mCSM graph-based signatures (as above) with the addition of small-molecule chemical features and ligand physicochemical properties to capture mutational changes. Predictive models trained on a representative set of protein–ligand complexes. Mutational impact on affinity is calculated as the log (ln) affinity fold change as below: $\ln(K_{wt}) - \ln(K_{mt}) = \ln$ (fold-change) Classification of ln (fold-change): ln (fold-change) < 0: Destabilising ln (fold-change) > 0: Stabilising	PDB code/file Single SNP Chain ID 3-letter ligand ID wt-affinity (nano Molar (nM))	Log affinity fold change Distance to ligand (Angstroms) DUET stability change (Kcal/mol) Binary classification of affinity and stability changes.	Predicts both the direction and the estimate of stability. Returns both DUET and ligand affinity changes, along with ligand distance to site. Measures both global and local stability effects. Analyses single mutation at a time. Returns a change in affinity value. Less reliable results for sites > 10 Å from ligand.
Rosetta Flex_ddG REF**:** [151]	St	All-atom energy function	Download only: https://www.rosettacommons.org/software/license-	Based on a mixed physics and knowledge-based approach. Uses all-atom energy function, parameterized from small molecule and X-ray crystal structure.	Customized PDB file Ligand	Each run outputs db3 file containing the changes in the main components of the energy function, ΔG wt, ΔG	For a reliable prediction, at least 35 runs per mutation are required, with each run taking between 2 and 4 h.

43

Table1 (*continued*)

Name of tool	Type	Operating Principle	Availability	Summary	User input	Output	User Notes
			and-download	The *Flex_ddG* protocol models changes in the ΔΔG upon mutation at a protein–protein or protein–ligand interface using the 'backrub' algorithm. This algorithm is used to sample conformational space and produce an ensemble of wt and mt models to estimate the interface ΔΔG values. Ternary Classification of ΔΔG: Destabilising: ΔΔG >=1 Neutral: −1 < ΔΔG < 1 Stabilising: ΔΔG<= −1	parameter file Customized XML protocol file Mutation list Chain ID	mt, and the ΔΔG upon mutation.	Access to HPC may be required for large number of mutations. Protocols are written in XML format. Requires license to download.
INPS-MD **I**mpact of **N**on-synonymous mutations on **P**rotein **S**tability-**M**ulti **D**imension REF: [141]	S/St	SVM regression	https://inpsmd.biocomp.unibo.it/inpsSuite Download: No	Calculates ΔΔG of mutation on sequence and structure. The sequence-based predictions are derived from seven descriptors to account for evolutionary information (INPS), while two additional structural features (RSA and energy difference between wt and mt structures) are included for the structure-based predictions (INPS-3D). SVM regression is used to map the sequence descriptors to the ΔΔG values. Classification of ΔΔG: ΔΔG < 0: Destabilising ΔΔG > 0: Stabilising	Fasta sequence/PDB file SNP list Chain ID (INPS-3D only)	Per SNP in list: Predicted ΔΔG	Predicts both the direction and the estimate of stability. Can operate on both sequence (INPS) and structure (INPS-3D) Accounts for anti-symmetric property of variation i.e ΔΔG (A->B) = - ΔΔG (B->A).
DeepDDG/ iDeepDDG REF: [142]	St	NN/ Ensemble method	http://protein.org.cn/ddg.html Download: No	Calculates ΔΔG of mutation using NN trained on nine categories of sequence and structural features. Operates independently as 'DeepDDG', and in an integrated manner as 'iDeepDDG'. In the latter, predictions from three methods: mCSM, SDM and DUET are fed into the concatenation layer of the NN to generate the consensus prediction. Classification of ΔΔG: ΔΔG < 0: Destabilising ΔΔG > 0: Stabilising	PDB code/file Network model: -DeepDDG -iDeepDDG SNP list in two modes: 1) Single or multiple SNPs 2) All possible mutations	Per SNP/all possible SNPs: Predicted ΔΔG	Predicts both the direction and the estimate of stability. Accounts for anti-symmetric property of variation i.e ΔΔG (A->B) = - ΔΔG (B->A). Runs in independent or integrated modes. 'DeepDDG' allows high throughput scanning for all possible point mutations with relatively fast computation time. For running 'iDeepDDG', user must provide predictions for each mutation from the mCSM DUET server.
DUET REF: [102]	St	Ensemble method: SVM	http://biosig.unimelb.edu.au/duet/ Download: No	Predicts stability effects upon mutation on proteins. Combines predictions from two complementary methods: mCSM and Site Directed Mutator (SDM) in an optimised predictor to generate the DUET prediction.	PDB code/file SNP list Chain ID Input mode1: Single	Input mode 1: 1) Predicted ΔΔG from mCSM, SDM and DUET. Input mode 2: Multi-column table with predicted ΔΔG from mCSM, SDM, DUET and RSA.	Predicts both the direction and the estimate of stability. Mutant structure is not required. webGL structural visualisation for input mode 1.

(*continued on next page*)

Table1 (*continued*)

Name of tool	Type	Operating Principle	Availability	Summary	User input	Output	User Notes
				The optimised predictor is generated using SVM trained with Sequential Minimal Optimisation. Classification of ΔΔG: ΔΔG < 0: Destabilising ΔΔG > 0: Stabilising	mutation Input mode 2: Systematic: all possible mutation for a single residue		
ELASPIC: **E**nsemble **L**earning **A**pproach for **S**tability **P**rediction of **I**nterface and **C**ore mutations REF: [124]	(St)	Ensemble method: ML	http://elaspic.kimlab.org/ Download: No	Predicts stability effects upon mutation in both, domain cores and domain-domain interfaces. Combination of semi-empirical energy terms, sequence conservation, and a wide variety of molecular details with a Stochastic Gradient Boosting of Decision Trees (SGB-DT) algorithm. Uses a combination of sequence, molecular and energy features including prediction scores from other tools.	Uniprot Protein ID or PDB structure SNP list	Multi-column table, with the main output being ΔG of wt and mt structures, and ΔΔG of mutation. Results are downloadable. FoldX generated mutant structures in pdb format Jmol applet showing superimposed wt mt structures.	Can be run as a single or multiple mutations and Protein-protein interactions Option to filter results based on additional criteria. Non-human proteins may take longer to run. An interactive connectivity network showing the affected protein–protein interaction mutations.
DynaMut REF: [118]	St	Ensemble method: NMA	http://biosig.unimelb.edu.au/dynamut/ Download: No	Predicts stability effects based on protein dynamics resulting from vibrational entropy changes. Integrates mCSM signatures and normal model analysis. Combines mutational effect from 3 structure-based prediction tools to generate a consensus prediction. Classification of ΔΔG: ΔΔG < 0 Destabilising ΔΔG > 0: Stabilising	PDB code/file Single SNP/ SNP list Chain ID	NMA based predictions Other structure-based predictions included.	Accounts for protein molecular motion and flexibility. Easy and detailed visualisation of results including interatomic interactions, deformation and fluctuation analysis. Returns a change in stability. Computationally expensive with relatively long runtime.

45

Table1
Sequence and structure-based tools that predict effect of pathogen missense mutations. The table is an up-to date list of currently available tools (as on 3rd August 2020). The type of method for each tool is specified using the following code; **S**: sequence-based, **St:** structure-based, **SA**: sequence alignment, **SS**: sequence and structure, **(St)**: structure if available. Other abbreviations used: MSA (Multiple Sequence alignment), EC (Evolutionary Conservation), NN (Neural Network), SVM (Support Vector Machine), ML (Machine learning), NMA (Normal Mode Analysis), ΔG: Gibbs free energy in Kcal/mol, ΔΔG: Change in Gibbs free energy in Kcal/mol, wt: wild-type, mt: mutant, K_{wt}: affinity of the wild-type protein-ligand complex, K_{mt}: affinity of the mutant complex, RSA: Relative Solvent Accessibility (%).

Name of tool	Type	Operating Principle	Availability	Summary	User input	Output	User Notes
SIFT: **S**orting **I**ntolerant **F**rom **T**olerant REF: [100]	S	EC	http://sift-dna.org Download: Yes	Calculates a normalised probability of substitution score from multiple alignments based on sequence homology using PSI-BLAST. Removes close homologous sequences to prevent over prediction of "tolerated" substitutions. Mutational effect on protein function is classified as damaging (<=0.05) or tolerated (>0.05).	Fasta sequence or aligned sequences SNP list	Per-SNP: 1) SIFT score 2) Binary mutation classification 3) Median sequence conservation	Predictions for submitted SNPs, as well as all possible SNPs (but without a score). Positions are weighted equally within an alignment. Alignments may be user defined. Sequence conservation score provides a useful estimate of whether the alignment contains sufficient variation to support classification.
PROVEAN: **Pro**tein **V**ariation **E**ffect **An**alyzer REF: [104]	S	EC	http://provean.jcvi.org/seq_submit.php Download: Yes	Related sequences are collected with BLAST (using CD-HIT) and clustered based on 75% global sequence identity. The top 30 clusters of closely related sequences form the supporting sequence set, used to generate the prediction. Delta alignment scores are computed for each supporting sequence and averaged within and across clusters to generate the final PROVEAN score. Predicted mutation effects are classified as either deleterious or neutral based on a predefined threshold (-2.5). Available as: PROVEAN Protein PROVEAN Protein Batch* PROVEAN Genome Variants* **Human and Mouse only*	Fasta sequence Mutation list (SNPs and INDELs)	Per-mutation: 1) PROVEAN score 2) Binary mutation classification	Predictions for submitted mutations only. Predict effects for both SNPs and INDELs, but not frameshift mutations. Batch processing of multiple organisms. The classification threshold is fixed in the online version. Stand-alone package only available for PROVEAN Protein.
SNAP2: **S**creening for **N**on-**A**cceptable **P**olymorphisms, v2 REF: [105]	S	NN	https://www.rostlab.org/services/SNAP/ Download: Yes	Combines evolutionary information with an expanded list of original SNAP features (amino acid properties) including features such as AA index, predicted binding residues and disordered regions, residue annotations from Pfam and PROSITE, etc. Mutations are classified as either neutral or effect based on predicted scores, between (-100 to 100) respectively. The prediction algorithm is based on a NN consisting of a feed-forward multi-layer perceptron. 10-fold cross-validation is used to create 10 models, each providing a single score for each output class (neutral/effect). The final score is calculated as the difference between the average scores for each output class.	Fasta sequence	For all possible substitutions: 1) Heatmap representing the predicted effect 2) Multi column table with Predicted Effect, Score and Accuracy.	Predictions for all possible substitutions. Prediction scores are accompanied by an "accuracy metric" to aid interpretation. Uses predicted structural features. Heatmap generated for visualisation of the predictions. Additional method (SNAP2noali) predicts functional effects without alignments. Automatic selection of best method (SNAP2 by default, and SNAP2noali for orphans) with notification to users.
ConSurf REF: [106]	S(St)	EC	https://consurf.tau.ac.il/ Download: No	Estimates evolutionary conservation rate of amino/nucleic acid positions based on the phylogenetic relations between homologous sequences. Homologous sequences are searched using CSI-BLAST, PSI-BLAST or BLAST, with closely related sequences removed using CD-HIT with multiple sequence alignments (MSA) generated by MAFFT by default.	Amino acid/ nucleotide sequence Structure (if available) MSA (if available) *Advanced options:*	Detailed output containing conservation scores, MSA and BLAST results. Estimates mapped onto sequence and structure.	Analysis at amino acid and nucleotide levels. Improved HMMER algorithm to search for homologous proteins. Results are accompanied by confidence intervals. Robust statistical approach to differentiate between apparent conservation (short evolutionary time) and genuine conservation (purifying selection). 'ConSeq' mode used in the absence of a

(*continued on next page*)

46

Table1 (*continued*)

Name of tool	Type	Operating Principle	Availability	Summary	User input	Output	User Notes
				MSA is used to construct phylogenetic relationships using the neighbour joining method. Position specific evolutionary rates are calculated using the empirical Bayesian or Maximum Likelihood methods. Scores graded 1 (variable) to 9 (conserved) for visualisation.	- homologue database - MSA methods - Phylogenetic tree - structural data - calculation method - evolutionary substitution model Optional: user defined MSA and phylogenetic tree.		structure. Site-specific predictions of the buried/exposed status of each position.
MAPP: **M**ultivariate **A**nalysis of **P**rotein **P**olymorphism REF: [110]	SA	EC	Download only: http://mendel.stanford.edu/SidowLab/downloads/MAPP/index.html	Combines MSA with 6 physicochemical properties for amino acids. Calculates a MAPP impact score for each position within the MSA. Sequences in the MSA are weighted to account for phylogenetic correlation. Physicochemical property scores for each column along with their mean and variances are calculated. The deviation of each property is calculated for every possible variant and converted to a single score.	Fasta format MSA Phylogenetic tree	Multicolumn table giving the physico-chemical characteristics of each position, MAPP impact score, and a listing of "good" and "bad" amino acids.	Predictions for every possible substitution, and median MAPP scores calculated for each position. Constructs a physiochemical profile rather than an amino acid profile. Demonstrates value of using only orthologous protein in creating a conservation profile. Scores are continuous and interpreted in a relative manner with higher MAPP scores indicating more conserved areas. Can be optimised for individual genes including MAPP impact score threshold for classification. Requires user defined MSA and phylogenetic tree.
PANTHER-PSEP: **P**rotein **AN**alysis **Th**rough **E**volutionary **R**elationships-**P**osition **S**pecific **E**volutionary **P**reservation REF: [149]	S	EC	http://www.pantherdb.org/tools/csnpScoreForm.jsp Download: Yes	Uses Hidden Markov Model (HMM) to align sequence to protein families and subfamilies in its database to calculate the evolutionary preservation metric. Uses variation over each alignment position to estimate the likelihood of a coding SNP to cause a functional impact on the protein. Score represents the time (in millions of years [my]) a given amino acid has been preserved in the lineage, directly corresponding to the likelihood of a functional impact. Score classified into: Probably damaging, Possibly damaging, Probably benign.	Fasta sequence SNP list *Other parameters:* Organism	Per-SNP: 1) Preservation Time: PANTHER PSEP score 2) Message: Classification of the PSEP score	Positions are weighted equally at all positions within an alignment. Profiles are subfamily specific if they substantially differ from entire family. User defined alignments are not possible since scores are derived from HMMs (PANTHER protein library) together with an ontology of protein function (PANTHER/X – a simplified form of GO) to make predictions.
FoldX suite REF: [113]	St	Empirical force field	Download only: http://foldxsuite.crg.eu/	Empirical force-field used for calculating mutational effects of stability, folding, and dynamics on proteins and nucleic acids ΔG (free energy of unfolding) is calculated using a combination of empirical terms. Empirical data (derived from protein engineering experiments) is used for weighting energy terms for stability calculations.	PDB file SNP list (including chain ID)	Multiple output files where requested. Main output is present in '*Dif_*" files, containing ΔG of wt and mt residues along with ΔΔG of mutation. Output also contains changes in the associated energy terms.	Command line interface. Creates mutant structure models. Can be used to analyse protein–protein and protein-DNA interactions. Calculates actual stabilities of wt and mt structures, as well as change in stability upon mutation (ΔΔG). Easily integrated into custom workflows.

Table1 (*continued*)

Name of tool	Type	Operating Principle	Availability	Summary	User input	Output	User Notes
				Foldx *BuildModel* command calculates stability changes upon mutation based on a full atomic description of the protein structure. Classification of ΔΔG: ΔΔG > 0: Destabilising ΔΔG < 0: Stabilising			Optimised energy function for faster calculations. Requires registration to download.
PoPMuSic (v2.1): **P**rediction **O**f **P**roteins **MU**tations **S**tab**I**lity **C**hanges REF: [115]	St	Physics-based and NN	https://soft.dezyme.com/ Download: No	Stability change upon mutation calculated using a linear combination of statistical energy potentials, accounting for variation in volume of the mutant residue. Predictive models include an optimised set of 52 parameters, whose values are estimated and optimised using a neural network. ΔΔG of point mutation is calculated by a linear combination of 16 terms: 13 statistical potentials, 2 terms for volume of wt and mut residues, and 1 independent term. Classification of ΔΔG: ΔΔG > 0: Destabilising ΔΔG < 0: Stabilising Additional "optimality" score is assigned for each position in the protein sequence. It indicates poorly optimised positions with potential functional consequences.	Only accepts currently available entries in the PDB SNP list in three input modes: 1) Systematic: all possible point mutations 2) Manual: single mutation 3) File: SNP list	Multi-column table containing secondary structure, solvent accessibility (%) and predicted ΔΔG of mutations.	Optimised to rapidly calculate stability changes of all possible mutations in mid-size proteins. Graphical output of sequence optimality scores. No option to upload user-defined PDB files. Requires registration to download.
I-Mutant (2.0) REF: [116]	S(St)	SVM	http://gpcr2.biocomp.unibo.it/cgi/predictors/I-Mutant3.0/I-Mutant3.0.cgi Download: No	Predicts stability effect of a point mutation, as a classification or regression task. The classification task predicts the direction of change, while the regression estimator predicts the ΔΔ*G*. Can be applied to both sequence and structure. RI value (Reliability Index) is computed from the output of the SVM model. Binary classification ΔΔG: ΔΔG < 0: Decrease Stability ΔΔG > 0: Increase Stability Ternary Classification ΔΔG: Large Decrease of Stability: ΔΔG < -0.5 Large Increase of Stability: ΔΔG > 0.5 Neutral Stability: 0.5<=ΔΔG<=0.5	Fasta sequence or PDB code/file Chain ID Single SNP Temperature PH Prediction request: Binary/ Ternary classification	Table containing: 1) RSA (%) of mt residue 2) RI (Reliability Index) 3) Predicted ΔΔG 3) Classification of ΔΔG	Predicts both the direction and the estimate of stability. Experimental conditions of pH and Temperature (Celsius) are considered in the stability calculations. Analyses a single mutation at a time only. Output on the web server is better than output requested via email. Use of two different SVM models can lead to discordance between the ΔΔ*G* sign and classification, but is stated to occur only in cases of low RI value.
STRUM: **STR**ucture-based prediction of protein stability changes **U**pon single-point **M**utation REF: [117]	S(St)	ML	https://zhanglab.ccmb.med.umich.edu/STRUM/ Download: Yes	Calculates ΔΔG of mutation using gradient boosting regression algorithm trained on 120 features divided into three groups (sequence, threading and structure). Classification of ΔΔG: ΔΔG < 0: Destabilising ΔΔG > 0: Stabilising	Fasta sequence or PDB file SNP list in two modes: 1) Single or multiple SNPs 2) Systematic: All possible SNPs for user defined amino acid segments.	Results available via e-mail only. Multi-column table containing ΔΔG for SNPs.	Combines sequence profiles and 3D features 3D Structure modelling of query protein sequence by iterative threading assembly refinement simulations Computationally expensive with relatively long runtime.
MAESTRO: **M**ulti **A**g**E**nt **ST**ability p**R**edicti**O**n	St	Multi agent: ML methods and	https://pbwww.che.sbg.ac.at/maestro/web	Multi-agent method where 3 ML methods i.e Artificial NN, SVM and Multiple Linear Regression. are combined to generate a consensus prediction.	PDB code/file Input mode: 1) Specific	Input modes 1 & 2 ΔΔG predictions and confidence intervals.	Ability to analyse mutations independently or in combination ΔΔG predictions are accompanied by

(*continued on next page*)

48

Table1 (*continued*)

Name of tool	Type	Operating Principle	Availability	Summary	User input	Output	User Notes
REF: [150]		statistical scoring functions	Download: Yes	Each agent (ML method) uses 9 input values divided into two categories: SSF functions and protein properties (size, mutational environment, etc.). Classification of ΔΔG: ΔΔG > 0: Destabilising ΔΔG < 0: Stabilising	mutations 2) Sensitivity profile: all possible mutations 3) Scan for destabilising mutations 4) Stability of Disulphide bonds	Graphical display.	confidence intervals. High throughput scanning for all possible point mutations. Specific mode for prediction of stabilising disulphide bonds.
mCSM suite: **m**utational **C**ut-Off **S**canning **M**atrix REF: [122]	St	Graph-based and ML	Protein Stability (PS), Protein-Protein (PP), Protein-DNA (P-NA) http://biosig.unimelb.edu.au/mcsm/ Download: No	Uses graph-based methods to calculate atomic pairwise distance surrounding the wt amino acid. Mutational impact is captured based on a change in the atomic pharmocophore count resulting from the point mutation. Together, this forms the mCSM-signature, and is used to train predictive models for analysing mutational impact on structure stability. Predicted ΔΔG < 0 relates to destabilising, and ΔΔG > 0 relates to stabilising mutational effects. Ternary Classification of Destabilising effect: Mild: −1 < ΔΔG < 0 Moderate: −2 < ΔΔG < -1 High: ΔΔG < -2 Ternary Classification of Stabilising effect: Mild: 0 < ΔΔG < 1 Moderate: 1 < ΔΔG < 2 High: ΔΔG > 2	PDB code/file SNP list Chain ID Input modes: 1) Single mutation 2) Mutation list 3) Systematic: all possible mutation for a single residue	Input mode 1: 1) Predicted ΔΔG 2) Classification of mutational stability change Input modes (2) & (3): Multi-column table with predicted ΔΔG and RSA.	Predicts both the direction and the estimate of stability. Mutant structure is not required. webGL structural visualisation for input mode 1. Works at an atomic level. Demonstrates correlation between atomic-distance pattern of the wild-type residue environment and mutational impact. Calculates overall stability of protein and interactions.
mCSM-lig: **m**utational **C**ut-Off **S**canning **M**atrix on **lig**and affinity REF**:** [88]	St	Graph-based and ML	Protein-ligand affinity (mCSM-lig): http://biosig.unimelb.edu.au/mcsm_lig/prediction Download: No	Based on the mCSM graph-based signatures (as above) with the addition of small-molecule chemical features and ligand physicochemical properties to capture mutational changes. Predictive models trained on a representative set of protein–ligand complexes. Mutational impact on affinity is calculated as the log (ln) affinity fold change as below: $\ln(K_{wt}) - \ln(K_{mt}) = \ln$ (fold-change) Classification of ln (fold-change): ln (fold-change) < 0: Destabilising ln (fold-change) > 0: Stabilising	PDB code/file Single SNP Chain ID 3-letter ligand ID wt-affinity (nano Molar (nM))	Log affinity fold change Distance to ligand (Angstroms) DUET stability change (Kcal/mol) Binary classification of affinity and stability changes.	Predicts both the direction and the estimate of stability. Returns both DUET and ligand affinity changes, along with ligand distance to site. Measures both global and local stability effects. Analyses single mutation at a time. Returns a change in affinity value. Less reliable results for sites > 10 Å from ligand.
Rosetta Flex_ddG REF**:** [151]	St	All-atom energy function	Download only: https://www.rosettacommons.org/software/license-	Based on a mixed physics and knowledge-based approach. Uses all-atom energy function, parameterized from small molecule and X-ray crystal structure.	Customized PDB file Ligand	Each run outputs db3 file containing the changes in the main components of the energy function, ΔG wt, ΔG	For a reliable prediction, at least 35 runs per mutation are required, with each run taking between 2 and 4 h.

49

Table1 (*continued*)

Name of tool	Type	Operating Principle	Availability	Summary	User input	Output	User Notes
			and-download	The *Flex_ddG* protocol models changes in the ΔΔG upon mutation at a protein–protein or protein–ligand interface using the 'backrub' algorithm. This algorithm is used to sample conformational space and produce an ensemble of wt and mt models to estimate the interface ΔΔG values. Ternary Classification of ΔΔG: Destabilising: ΔΔG >=1 Neutral: −1 < ΔΔG < 1 Stabilising: ΔΔG<= −1	parameter file Customized XML protocol file Mutation list Chain ID	mt, and the ΔΔG upon mutation.	Access to HPC may be required for large number of mutations. Protocols are written in XML format. Requires license to download.
INPS-MD **I**mpact of **N**on-synonymous mutations on **P**rotein **S**tability-**M**ulti **D**imension REF: [141]	S/St	SVM regression	https://inpsmd.biocomp.unibo.it/inpsSuite Download: No	Calculates ΔΔG of mutation on sequence and structure. The sequence-based predictions are derived from seven descriptors to account for evolutionary information (INPS), while two additional structural features (RSA and energy difference between wt and mt structures) are included for the structure-based predictions (INPS-3D). SVM regression is used to map the sequence descriptors to the ΔΔG values. Classification of ΔΔG: ΔΔG < 0: Destabilising ΔΔG > 0: Stabilising	Fasta sequence/PDB file SNP list Chain ID (INPS-3D only)	Per SNP in list: Predicted ΔΔG	Predicts both the direction and the estimate of stability. Can operate on both sequence (INPS) and structure (INPS-3D) Accounts for anti-symmetric property of variation i.e ΔΔG (A->B) = - ΔΔG (B->A).
DeepDDG/ iDeepDDG REF: [142]	St	NN/ Ensemble method	http://protein.org.cn/ddg.html Download: No	Calculates ΔΔG of mutation using NN trained on nine categories of sequence and structural features. Operates independently as 'DeepDDG', and in an integrated manner as 'iDeepDDG'. In the latter, predictions from three methods: mCSM, SDM and DUET are fed into the concatenation layer of the NN to generate the consensus prediction. Classification of ΔΔG: ΔΔG < 0: Destabilising ΔΔG > 0: Stabilising	PDB code/file Network model: -DeepDDG -iDeepDDG SNP list in two modes: 1) Single or multiple SNPs 2) All possible mutations	Per SNP/all possible SNPs: Predicted ΔΔG	Predicts both the direction and the estimate of stability. Accounts for anti-symmetric property of variation i.e ΔΔG (A->B) = - ΔΔG (B->A). Runs in independent or integrated modes. 'DeepDDG' allows high throughput scanning for all possible point mutations with relatively fast computation time. For running 'iDeepDDG', user must provide predictions for each mutation from the mCSM DUET server.
DUET REF: [102]	St	Ensemble method: SVM	http://biosig.unimelb.edu.au/duet/ Download: No	Predicts stability effects upon mutation on proteins. Combines predictions from two complementary methods: mCSM and Site Directed Mutator (SDM) in an optimised predictor to generate the DUET prediction.	PDB code/file SNP list Chain ID Input mode1: Single	Input mode 1: 1) Predicted ΔΔG from mCSM, SDM and DUET. Input mode 2: Multi-column table with predicted ΔΔG from mCSM, SDM, DUET and RSA.	Predicts both the direction and the estimate of stability. Mutant structure is not required. webGL structural visualisation for input mode 1.

(*continued on next page*)

Table1 (*continued*)

Name of tool	Type	Operating Principle	Availability	Summary	User input	Output	User Notes
				The optimised predictor is generated using SVM trained with Sequential Minimal Optimisation. Classification of ΔΔG: ΔΔG < 0: Destabilising ΔΔG > 0: Stabilising	mutation Input mode 2: Systematic: all possible mutation for a single residue		
ELASPIC: **E**nsemble **L**earning **A**pproach for **S**tability **P**rediction of **I**nterface and **C**ore mutations REF: [124]	(St)	Ensemble method: ML	http://elaspic.kimlab.org/ Download: No	Predicts stability effects upon mutation in both, domain cores and domain-domain interfaces. Combination of semi-empirical energy terms, sequence conservation, and a wide variety of molecular details with a Stochastic Gradient Boosting of Decision Trees (SGB-DT) algorithm. Uses a combination of sequence, molecular and energy features including prediction scores from other tools.	Uniprot Protein ID or PDB structure SNP list	Multi-column table, with the main output being ΔG of wt and mt structures, and ΔΔG of mutation. Results are downloadable. FoldX generated mutant structures in pdb format Jmol applet showing superimposed wt mt structures.	Can be run as a single or multiple mutations and Protein-protein interactions Option to filter results based on additional criteria. Non-human proteins may take longer to run. An interactive connectivity network showing the affected protein–protein interaction mutations.
DynaMut REF: [118]	St	Ensemble method: NMA	http://biosig.unimelb.edu.au/dynamut/ Download: No	Predicts stability effects based on protein dynamics resulting from vibrational entropy changes. Integrates mCSM signatures and normal model analysis. Combines mutational effect from 3 structure-based prediction tools to generate a consensus prediction. Classification of ΔΔG: ΔΔG < 0 Destabilising ΔΔG > 0: Stabilising	PDB code/file Single SNP/ SNP list Chain ID	NMA based predictions Other structure-based predictions included.	Accounts for protein molecular motion and flexibility. Easy and detailed visualisation of results including interatomic interactions, deformation and fluctuation analysis. Returns a change in stability. Computationally expensive with relatively long runtime.

51

the S2 subunit in MERS-CoV isolates [109] to aid antiviral strategies.

e. Mapp

MAPP (Multivariate Analysis of Protein Polymorphism) predicts the functional impact of all possible missense mutations. It combines evolutionary conservation and physicochemical information. It uses data from multiple sequence alignments from orthologs to estimate a mean for each of the six physicochemical properties (hydropathy, polarity, charge, volume, and free energy in alpha helices and beta strands) for each position. A single composite value for each physicochemical value is generated based on the deviation from the mean for all twenty amino acids. High MAPP scores indicate highly conserved sites, which in the context of drug resistance can indicate resistance promoting sites [110]. MAPP has been used to develop the ProPhylER [111] tool, used for proteome wide investigation of mutational impact on eukaryotic protein.

2.2. Structure-based methods

When analysing missense mutations, structure-based methods can offer a 3-dimensional explanation of molecular consequences of mutations, which may not be evident from sequence analysis alone [86,89]. These methods include the analysis of the protein structural and functional consequences of mutations, including those on protein folding, stability, dynamics, and alterations to interactions with normal ligands. Protein structure information can be incorporated through rule-based or machine learning based approaches (see Table 1). As acquired resistance can develop through missense mutations, analysing their effects can inform on underlying mechanisms of resistance. In previous analyses, we observed that known resistance mutations arising in the drug-target tend to significantly reduce functional affinities, such as nucleic acid affinity [80–82,93–95]. Resistance mutations in drug activators are associated with large decreases in protein stability or activity [79,80], and those in drug exporters tend to increase protein flexibility to promote drug export [91]. To run these predictors, a crystal structure of the protein or a homology model is required. A summary of the principle methods and applications are described below:

2.2.1. Measures of protein stability

The introduction of resistance-causing missense mutations to a protein structure rarely comes at a negligible cost to protein stability, whether decreasing local stability and affecting protein folding, or increasing local stability and compromising wild-type protein dynamics [112]. Therefore, quantifying the effect of missense mutations on stability presents a good starting point in understanding the basic variant protein changes. Computational tools predicting thermodynamic stability of a protein do so by estimating the Gibbs free energy (ΔG Kcal/mol). The subsequent impact of a point mutation on protein stability is then estimated as a change in the Gibbs free energy ($\Delta\Delta G$ Kcal/mol) between wild-type and mutant proteins, or vice versa. Additionally, these tools provide both the extent (the actual value of $\Delta\Delta G$) as well as the direction (destabilising/stabilising) of the resulting mutational effect. Different *in silico* protein stability predictors are available, of which we highlight a few, based on the methodologies considered in their approximations. Further details for these (and additional) methods can be found in Table 1.

a. FoldX is an empirical-based predictor which provides information on how a single point mutation alters the stability of a protein. It constructs structure models of the protein with the mutation and estimates the stability (ΔG) associated with the mutant protein. Estimation of stability is based on intramolecular interactions such as van der Waals' forces, solvation energies, interactions with water, hydrogen bonds, electrostatic effects and main and side chain entropies. Mutational impact is calculated through a weighted summation of all the intramolecular interactions, and estimated as a change in stability ($\Delta\Delta G$) between mutant and wild-type structures. In this way, ΔG for each mutant protein, $\Delta\Delta G$ upon mutation, and the contribution of each intramolecular interaction, are made available to the user. The extent of the mutational impact (the value of $\Delta\Delta G$) as well as the direction of change ($\Delta\Delta G < 0$: stabilising, $\Delta\Delta G > 0$: destabilising) are captured by the predictions [113,114].
b. PoPMuSic (v2.1) is a statistical method which uses knowledge-based potentials to predict mutational impact on the stability of a protein. It returns the predicted $\Delta\Delta G$ of a single point mutation of a protein and is able to systematically analyse this for all possible point mutations for a given protein. Additionally, an 'optimality' score for each amino acid in the sequence with respect to stability is returned. The optimality score identifies sites of structural weakness i.e. clusters of residues that are considered non-optimal from an evolutionary perspective. Therefore, mutations with desired stability properties ($\Delta\Delta G < 0$: stabilising, $\Delta\Delta G > 0$: destabilising) and poorly optimised positions can be identified. These sites can relate to the protein's function, and be used for rational protein design and other experimental studies. In PoPMuSic, a protein is represented as a statistical potential based on individual residue properties such as sequence position, conformation, solvent accessibility, or a combination of inter-residue distances. The optimality score is computed from the sum of the predicted $\Delta\Delta G$ of all stabilising mutations at a given position in the sequence. Since the majority of the mutations have a destabilising effect, this score is expected to be close to zero for most positions in the sequence, with high negative values indicating sites with strongly stabilising mutations and/or several stabilising mutations with mild effect [115].
c. I-Mutant (v2.0) is an ML based predictor which computes mutational stability changes using support vector machines. It provides an estimate of the $\Delta\Delta G$ upon a single point mutation based on protein structure (or sequence). The resulting $\Delta\Delta G$ highlights the extent as well as the direction of impact ($\Delta\Delta G < 0$: destabilising, $\Delta\Delta G > 0$: stabilising) on the protomer. The predictions consider the mutated residue environment as a 9 Å region (structure) and a 19-residue window (sequence) surrounding the mutation. This environment is combined with experimental pH and temperature conditions, enabling the user to define different pH and temperature conditions on a case-by-case basis to better encompass protein biological conditions [116].
d. STRUM is an ML based predictor and returns an estimate of the $\Delta\Delta G$ of a single point mutation on 3D models based on wild-type sequences. It can be used to analyse single mutations or all possible mutations within a specified region of the protein. Similar to methods above, both the magnitude of change as well as the direction ($\Delta\Delta G < 0$: destabilising, $\Delta\Delta G > 0$: stabilising) are encapsulated in the predictions. The 3D models are generated using iterative threading assembly and combined physics- and knowledge-based energy functions. Predictors are trained based on 3 groups of features: sequence, threading, and I-TASSER structure. A total of 120 features are trained through Gradient Boosted Regression Trees (GBRT) to overcome overfitting effects [117].

2.2.2. Measures of global and local stability within a single framework

The mCSM (mutation Cut-off Scanning Matrix) suite of computational tools accounts for the changes in protein stability dynamics [118], and interactions with other proteins [119], ligands [88] and nucleic acids [120] upon introduction of missense mutation. It estimates change in stability ($\Delta\Delta G$) and change in binding affinity of the ligand. Measuring the impact of missense mutations beyond protein stability, by looking at functional affinities, is crucial to characterise the mechanisms of AMR-associated mutations. This is because affinities to ligands, nucleic acids and other proteins are highly dependent on specific interaction sites, irrespective of protein stability changes. Functionally, protein affinity changes to its ligand is especially important in AMR, as it enables the identification of mutations directly affecting ligand binding. The extent of this importance, however, relates to the drug mode of action, meaning that other functional affinities should also be considered to identify mechanisms beyond direct ligand binding. The mCSM suite of tools quantify these stability and functional measurements using graph-based signatures [121], which summarise the global environment of the protein as a series of nodes for each atom, and represents the local environment at the mutation site as edges on the graph between the nodes at similar distances from the mutation. A pharmacophore count is appended to these signatures to account for any physicochemical changes imparted by the missense mutations [122] **(**Fig. 1**).** Through this graph-based network, the impact of a missense mutation over the whole protein can be calculated. All methods within the mCSM suite are based on ML approaches in quantifying missense mutational changes, and are freely available via their respective web servers.

Ensemble methods like DUET [102] generate a consensus prediction based on two different tools, while the *meta*-predictor tool by Broom, et al. [123] combines predictions from eleven available tools. Similarly, the ELASPIC method [124] combines semi-empirical energy terms, sequence conservation, and several molecular features to predict mutational effect on stability and affinity. Likewise, DynaMut [118] combines graph-based structural predictions with Normal Mode Analysis to account for protein dynamics and molecular motion to assess mutational impact. Consensus approaches have the advantage of improved accuracy over individual tools, but are tightly coupled and sensitive to their availability.

2.2.3. Insights from molecular dynamics simulation experiments

Despite not providing direct thermodynamic measures of mutations, molecular dynamics (MD) remains an invaluable technique for analysing mutational effects on protein conformational movement, especially considering that other techniques run on static protein structures. In the context of AMR, MD simulations enable comparison between wild-type and mutant protein trajectories. Visualising these differences can highlight co-occurring mutations and sites with local protein rigidification. Different MD techniques may be used, depending on computational cost and the level of throughput required.

An all-atom MD method has been adopted to study co-occurring missense mutations V82F/I84V (known to confer resistance to target inhibitors) within HIV-1 protease [125]. This analysis enabled the characterisation of an equilibrium shift imparted by these mutations from a closed to a semi-open conformation as a possible cause of drug resistance [125]. More recently, the effect of G140S mutation on HIV-1C Integrase (IN) protein provided insight into dolutegravir resistance. Decreased stability of IN and higher flexibility around the 140 loop region in the mutant system reduced drug affinity [126]. Similarly, MD simulations also examined artemisinin resistance in malaria. Mutation R539T and C580Y in the *P. falciparum* K13 region revealed local structural destabilisation of the Kelch-repeat propeller (KREP) domain but not the overlapping shallow pocket [78]. In fungal and bacterial enzymes, MD investigation of the interaction of triazole drugs with their target, CYP51, has highlighted the potential to design inhibitors with greater ortholog specificity. While protein-fluconazole interactions were strongly mediated by ligand-HEME interactions in fungal enzymes, the same was mediated by polar interactions in the bacterial counterpart (CYP51 *Mtb*) [127]. Stereochemical changes, rather than electrostatic effects, of ten point mutations in *Mtb katG* led to isoniazid (INH) resistance by restricting access of the drug to its catalytic site [128]. Likewise, conserved motions and unbinding events of 82 point mutations in *Mtb pncA,* linked to PZA resistance, were also discerned through MD simulations. Coupled expansions and contractions of the pncA lid and the side flap were observed in the unbinding of PZA in some mutants, while destabilisation of the "hinge" or nearby residues facilitated lid opening and PZA release from the active site [129].

MD studies have also shed light on AMR mutations in biological pathways. For example, mutations Y59H, M84I and E160D within the RamR homodimerization domain on *ramA* promoter were shown to affect structure stability and binding affinity. These mutations led to dysregulation of the multidrug efflux pump RND, and consequent drug resistance in *Salmonella enterica* [130]. Another example, where extensive modifications modelled by MD simulations of six missense mutations in Thymidylate synthase A (ThyA), a key enzyme in the *Mtb* folate pathway, provided a deeper understanding of Para-aminosalicylic acid resistance [131]. Likewise, investigation of inhA-INH resistance in *Mtb* revealed a ligand "locking" mechanism together with increased vibrational coupling between inhA cofactor binding site residues, responsible for the inhibitory function of the wild-type complex. This insight provided an explanation of how the resistant mutation S94A circumvents these subtle changes in global structural dynamics, with downstream effects in the fatty acid synthase pathway [132]. All-atom MD simulations have also been used to understand the mechanism of anti-microbial peptides within biofilms, which can potentially serve as alternative therapies in the presence of AMR [133].

Although, an all-atom MD approach offers detailed analysis of specific mutations, it is often computationally expensive making it impractical for large mutational datasets. In such cases, an approximated MD technique, known as normal mode analysis (NMA) can be adopted. NMA uses harmonic motion to summarise protein dynamics arising from vibrational entropy changes. This approach is the basis for DynaMut [118] (part of the mCSM-suite of computational tools described above) which predicts missense mutational impact on proteins while accounting for their molecular motions.

3. Applications of the computational tools for characterising drug resistance in TB and other infectious diseases

The tools described above for measuring the effects of mutations within a gene have been used to provide a molecular understanding of how variants can affect pathogen drug resistance in *Mtb* [80,92] and *P. vivax* [134]. In all cases, the different tools have provided complementary information to describe mutational effects under selective pressure as a balance of fitness costs across different protein properties.

To demonstrate the utility of this approach, we explore in more detail *Mtb* variants in two genes *katG* (resistance to isoniazid) and *rpoB* (resistance to rifampicin), which have been associated with drug resistance from GWAS analyses [11,45]. Most *katG* mutations conferred resistance through a disruption of protein stability [80]. Functionally, it is thought that *Mtb* renders the non-essential KatG unstable to impede the activation for prodrug isoniazid, thereby

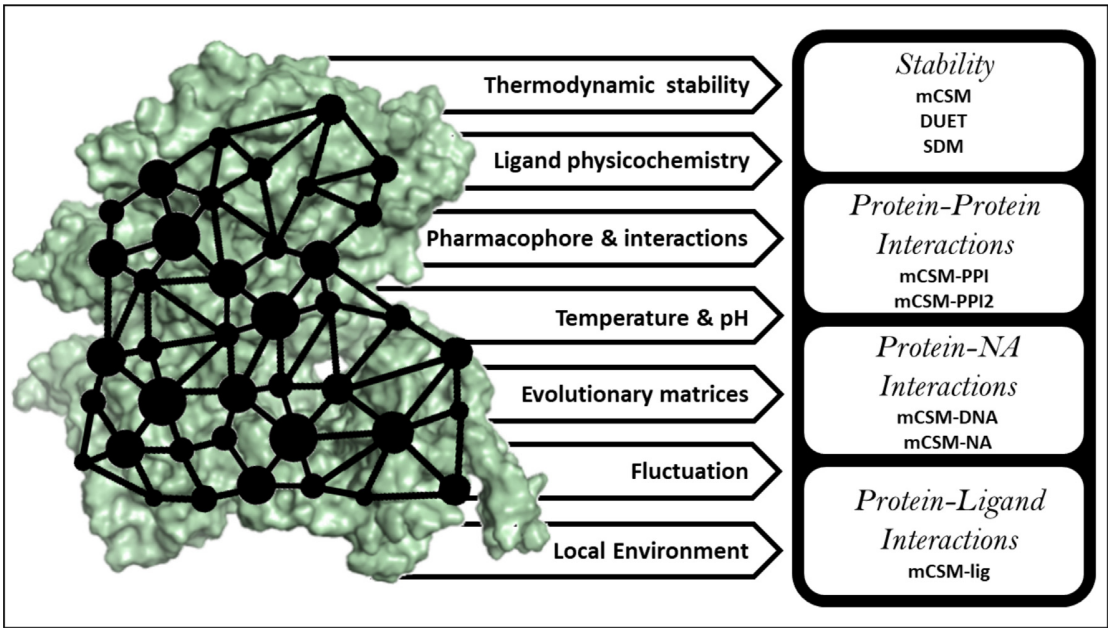


Fig. 1. A summary of mutational Cut-off Scanning Matrix (mCSM) method and its application in measuring mutational effects on protein stability (mCSM DUET), protein–protein interaction (mCSM-PPI, mCSM-PPI2), protein-nucleic acid (mCSM-NA) and protein–ligand affinity (mCSM-lig).

conferring resistance. When considering rifampicin resistant mutations within gene *rpoB*, we found that most mutations disrupt protein–protein interactions, leading to a loss in nucleic acid affinity. Structurally, the effects of these mutations within RpoB, the β-subunit of RNA polymerase, are compensated for by mutations within RpoC, which is the β′ subunit, thereby restoring normal functioning of the RNA polymerase, with an added resistance property [135–137].

Within this analysis, two distinct classes of mutations were observed: (i) those having high allele frequency within GWAS, but which had mild overall effects on protein stability and affinities to ligands, other proteins and nucleic acids, and (ii) those having lower allele frequency but more drastic effects on protein properties. Theoretically, it is thought that a high mutational incidence of class (i) mutations is a result of lower likelihood of evolutionary purging when compared to class (ii) mutations, which is based on the structural and functional effects imparted at the protein level. Mutations from each class were also seen to co-occur as haplotypes, where they are thought to compensate for each other in terms of protein fitness [80].

Using 571 missense SNPs in *katG* across 19265 *Mtb* isolates, we tested for an association between mutation odds ratio and allele frequencies with the biophysical effect on protein stability **(**Fig. 2**).** This analysis suggests a higher proportion of destabilising mutations (~84%, n = 480 vs ~55.5%, n = 105) with only a small proportion of mutations lying within 10 Å of the active site (~10%, n = 57 vs ~15%, n = 28) highlighting the importance of allosteric mutations in INH drug resistance. There is a weak negative correlation between protein stability and odds ratio ($\rho = -0.15$, $P < 0.001$), and between protein stability and allele frequency ($\rho = 0.31$, $P < 0.001$) **(**Fig. 3**a).** Analysis of biophysical effects (destabilising vs stabilising) of *katG* mutations by *Mtb* lineage revealed statistically significant differences **(**Fig. 3**b,** Kolmogorov-Smirnov $P \leq$ 1.3e-08**).**

This type of analysis can be implemented on proteins encoded on plasmids (a common vector of resistance), where this approach has been used to explain the evolution of carbapenem resistance in *Acinetobacter baumannii* [91].

4. Computational structural tools predicting drug resistance

A limitation of current genomic sequencing-based resistance diagnostic approaches is that they require pre-existing knowledge about the phenotypic consequences of a variant. This means we often cannot detect it until it has been established within the population. By contrast, we have shown that using these tools we can pre-emptively identify likely drug resistant mutations in the absence of previous genomic data. These insights are of particular relevance for new drugs without extensive clinical data, and drugs which lack approved diagnostic tests. We have therefore used this approach to explore resistance against the TB drugs BDQ [81] and PZA [82]. The use of our PZA predictive model within the clinic was the first successful translational application of structural guided resistance detection. This revealed the power of combining structural interpretation within existing diagnostic sequencing frameworks [93]. Additionally, other ML based approaches have also been used in predicting drug resistance in *Mtb* [56,138].

5. Designing better antibacterial drugs

It has been suggested that a way to minimise the development of resistance is by making compounds that interact similarly to a natural ligand [139]. The rationale being that this would lead to any resistance hot-spot having a higher fitness cost associated with it. This led to one of the first successful structure-guided drug discovery projects on neuraminidase inhibitors. Computational tools aid molecular characterisation of novel genomic variants, which provide opportunities to pre-empt likely resistant mutations. Anticipating these variants before they arise in a population can inform the drug discovery pipeline, especially in developing compounds less prone to resistance emergence. Such an approach has already been used as part of the drug development efforts against the TB drug target IMPDH [99]. The mutation predicted was the only resistant variant detected in subsequent in vitro resistant assays. Further, compounds designed to avoid this hot-spot were less prone to develop resistance [96–98]. This type of analysis complements the development of new tools that integrate geno-

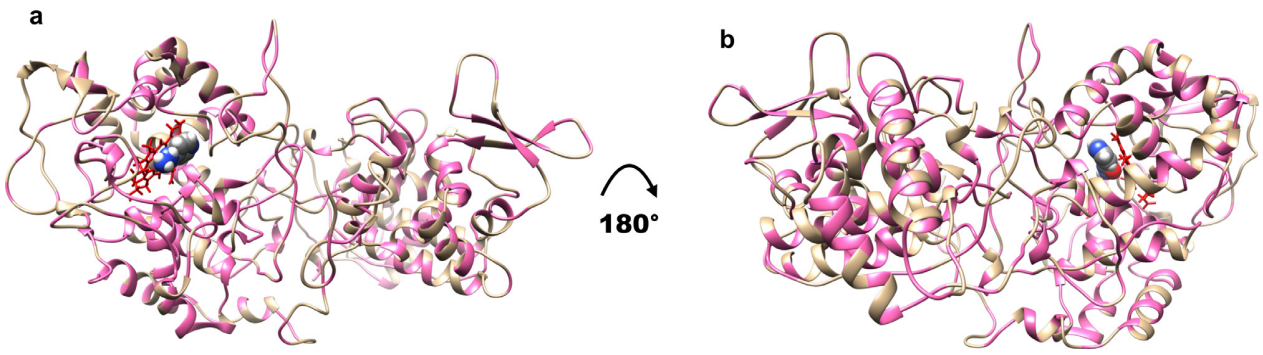


Fig. 2. Structure of katG in complex with the drug isoniazid (INH) coloured by 378 mutational positions linked to 571 SNPs. Areas marked in pink are associated with one or more mutations. HEM is denoted in red, INH is denoted as spheres. Parts a) and b) denote the structure in two different orientations, rotated by 180°. Figure rendered using UCSF Chimera, Version 1.13.1. (For interpretation of the references to colour in this figure legend, the reader is referred to the web version of this article.)

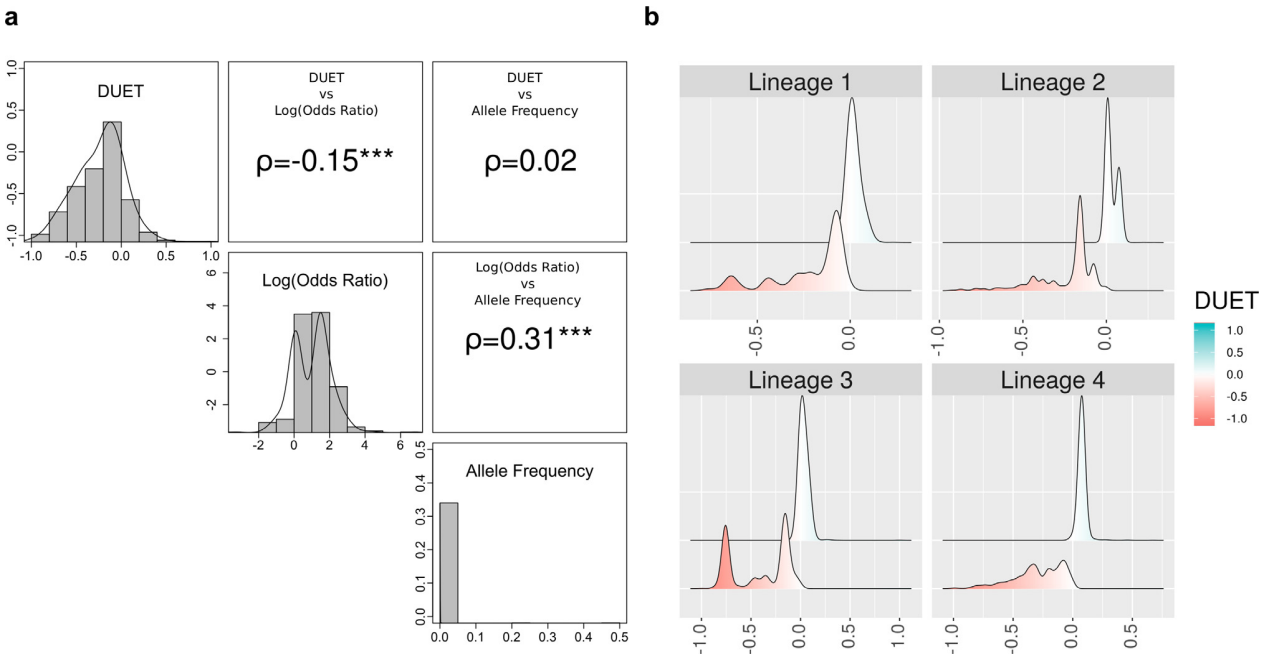


Fig. 3. Relationship between the impact of *katG* mutations on Protein stability (DUET) with Odds Ratio (OR), Allele Frequency (AF) and *Mtb* lineages. **a)** Pairwise correlations between DUET protein stability and GWAS measures of OR and AF of 566 mutations (total number of mutations with associated OR). The upper panel in both plots include the pairwise Spearman correlation values (denoted by ρ) along with their statistical significance (***$P < 0.001$). **b)** Lineage distribution of samples with *katG* mutations showing *Mtb* lineages 1–4 according to DUET protein stability ranging from red (-1, most destabilising) to blue (+1, most stabilising). The number of samples within each lineage are: Lineage 1 (n = 2448), Lineage 2 (n = 6813), Lineage 3 (n = 5020) and Lineage 4 (n = 2739). The number of samples contribute to the 566 *katG* mutations. Figure generated using R statistical software, version 3.6.1. (For interpretation of the references to colour in this figure legend, the reader is referred to the web version of this article.)

mic and structural data such as the Target-Pathogen online resource [140], which prioritises candidate drug targets in ten clinically important and diverse pathogens. This approach underscores the importance of structural data in guiding the drug-discovery process [140].

6. Summary and outlook

Large scale genomic studies have enabled identification of mutational associations with a resistance phenotype, useful for surveying the presence and spread of resistance to a wide range of antimicrobials. However, understanding the functional effects of putative mutations is crucial. Computational tools accounting for anti-symmetric properties of variation i.e. ΔΔG (A->B) = -ΔΔG (B->A) [118,141,142] are able to achieve improved prediction performance complementing experimental studies [85].

Genomic and structural analysis of resistance can infer mutational effects with therapeutic consequences before they become fixed in a pathogen population. This has implications for both infection surveillance and in the development of next generation drugs. The latter is of particular relevance to fragment-based drug discovery (FBDD) [143,144]. For the past 20 years, this has been a powerful route to new therapeutics, for example, in the development of vemurafenib for late-stage melanoma [145], and is increasingly being applied in the search for new antimicrobial drugs [146–148]. FBDD uses a library of low molecular weight, low affinity binding molecules (fragments) to probe a target protein. This approach helps to identify areas that are receptive to binding. Biophysical and structural biology techniques are used to determine which fragment binds, and how. The target can then be used to guide an expansion of the fragment to a higher molecular weight and higher affinity binding molecule. An important

step in this process is elaborating fragments that bind, to generate compounds that can be taken through to clinical testing. This is the stage at which crucial decisions are made about the regions of the drug target to exploit. However, pathogen tolerance is seldom considered, with direct consequences on drug effectiveness or efficacy. Current methods of analysing the effects of mutations either operate at the gene level (identifying known markers of resistance) or focus on a specific effect of the mutation (protein stability) without directly relating it to a resistance phenotype. Combining genomic results with structural analysis permits consideration of mutational impact on a potential drug binding region, providing informed decisions regarding drug efficacy. This has the potential to help the design of better antimicrobial drugs.

CRediT authorship contribution statement

Tanushree Tunstall: Conceptualization, Formal analysis, Visualization, Writing - original draft. **Stephanie Portelli:** Conceptualization, Visualization, Writing - original draft. **Jody Phelan:** Data curation. **Taane G. Clark:** Writing - review & editing. **David B. Ascher:** Conceptualization, Supervision, Writing - review & editing. **Nicholas Furnham:** Conceptualization, Supervision, Writing - review & editing.

Declaration of Competing Interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Acknowledgments

TT is supported by a BBSRC PhD studentship (BBSRC No: BB/S507544/1). SP is supported by an Australian Government Research Training Program Scholarship. JP is funded by a Newton Institutional Links Grant (British Council. 261868591). TGC is funded by the Medical Research Council UK (Grant no. MR/M01360X/1, MR/N010469/1, MR/R025576/1, and MR/R020973/1) and BBSRC (Grant no. BB/R013063/1). DBA is supported by the Jack Brockhoff Foundation [JBF 4186, 2016], a Newton Fund RCUK-CONFAP Grant awarded by The Medical Research Council (MRC) and Fundação de Amparo à Pesquisa do Estado de Minas Gerais (FAPEMIG) [MR/M026302/1] and the National Health and Medical Research Council of Australia [GNT1174405]. We thank Charlotte Ecclestone for her input on the Rosetta tool included in Table 1.

References

[1] WHO AMR-Fact-Sheet. WHO | 10 Facts on Antimicrobial Resistance 2018. https://www.who.int/news-room/facts-in-pictures/detail/antimicrobial-resistance [accessed October 3, 2019].
[2] Walsh CT, Wencewicz TA. Prospects for new antibiotics: a molecule-centered perspective. J Antibiot (Tokyo) 2014;67:7–22. https://doi.org/10.1038/ja.2013.49.
[3] O'Neill Commission. Tackling Drug-Resistant Infections Globally-Final Report and Recommendations. The Review on Antimicrobial Resistance, Chaired by Jim O'Neill. 2016.
[4] Grobusch MP, Kapata N. Global burden of tuberculosis: where we are and what to do. Lancet Infect Dis 2018;18:1291–3. https://doi.org/10.1016/S1473-3099(18)30654-6.
[5] WHO. Global tuberculosis report 2018; 2018.
[6] Dookie N, Rambaran S, Padayatchi N, Mahomed S, Naidoo K. Evolution of drug resistance in Mycobacterium tuberculosis: a review on the molecular determinants of resistance and implications for personalized care. J Antimicrob Chemother 2018;73:1138–51. https://doi.org/10.1093/jac/dkx506.
[7] Zhang Y. The magic bullets and tuberculosis drug targets. Annu Rev Pharmacol Toxicol 2005;45:529–64. https://doi.org/10.1146/annurev.pharmtox.45.120403.100120.
[8] Borgdorff MW, van Soolingen D. The re-emergence of tuberculosis: what have we learnt from molecular epidemiology?. Clin Microbiol Infect 2013;19:889–901. https://doi.org/10.1111/1469-0691.12253.
[9] Gomez JE, McKinney JDM. tuberculosis persistence, latency, and drug tolerance. Tuberculosis 2004;84:29–44. https://doi.org/10.1016/J.TUBE.2003.08.003.
[10] Gengenbacher M, Kaufmann SHE. *Mycobacterium tuberculosis* : success through dormancy. FEMS Microbiol Rev 2012;36:514–32. https://doi.org/10.1111/j.1574-6976.2012.00331.x.
[11] Coll F, Phelan J, Hill-Cawthorne GA, Nair MB, Mallard K, Ali S, et al. Genome-wide analysis of multi- and extensively drug-resistant Mycobacterium tuberculosis. Nat Genet 2018;50:307–16. https://doi.org/10.1038/s41588-017-0029-0.
[12] Phelan JE, O'Sullivan DM, Machado D, Ramos J, Oppong YEA, Campino S, et al. Integrating informatics tools and portable sequencing technology for rapid detection of resistance to anti-tuberculous drugs. Genome Med 2019. https://doi.org/10.1186/s13073-019-0650-x.
[13] Ali J, Rafiq QA, Ratcliffe E. Antimicrobial resistance mechanisms and potential synthetic treatments. Futur Sci OA 2018;4:FSO290. https://doi.org/10.4155/fsoa-2017-0109.
[14] Reygaert WC. An overview of the antimicrobial resistance mechanisms of bacteria. AIMS Microbiol 2018;4:482–501. https://doi.org/10.3934/microbiol.2018.3.482.
[15] Munita JM, Arias CA. Mechanisms of Antibiotic Resistance. Microbiol Spectr 2016;4. https://doi.org/10.1128/microbiolspec.VMBF-0016-2015.
[16] Forsberg KJ, Patel S, Gibson MK, Lauber CL, Knight R, Fierer N, et al. Bacterial phylogeny structures soil resistomes across habitats. Nature 2014;509:612–6. https://doi.org/10.1038/nature13377.
[17] Stegmann E, Frasch H-J, Kilian R, Pozzi R. Self-resistance mechanisms of actinomycetes producing lipid II-targeting antibiotics. Int J Med Microbiol 2015;305:190–5. https://doi.org/10.1016/j.ijmm.2014.12.015.
[18] Fisher RA, Gollan B, Helaine S. Persistent bacterial infections and persister cells. Nat Rev Microbiol 2017;15:453–64. https://doi.org/10.1038/nrmicro.2017.42.
[19] van Hoek AHAM, Mevius D, Guerra B, Mullany P, Roberts AP, Aarts HJM. Acquired antibiotic resistance genes: an overview. Front Microbiol 2011;2:203. https://doi.org/10.3389/fmicb.2011.00203.
[20] Gyles C, Boerlin P. Horizontally transferred genetic elements and their role in pathogenesis of bacterial disease. Vet Pathol 2014;51:328–40. https://doi.org/10.1177/0300985813511131.
[21] Barlow M. What antimicrobial resistance has taught us about horizontal gene transfer. Methods Mol Biol 2009;532:397–411. https://doi.org/10.1007/978-1-60327-853-9_23.
[22] Kanji A, Hasan R, Hasan Z. Efflux pump as alternate mechanism for drug resistance in Mycobacterium tuberculosis. Indian J Tuberc 2019. https://doi.org/10.1016/j.ijtb.2018.07.008.
[23] Kanji A, Hasan R, Ali A, Zaver A, Zhang Y, Imtiaz K, et al. Single nucleotide polymorphisms in efflux pumps genes in extensively drug resistant Mycobacterium tuberculosis isolates from Pakistan. Tuberculosis 2017. https://doi.org/10.1016/j.tube.2017.07.012.
[24] Cowen LE, Sanglard D, Howard SJ, Rogers PD, Perlin DS. Mechanisms of Antifungal Drug Resistance. Cold Spring Harb Perspect Med 2014;5:. https://doi.org/10.1101/cshperspect.a019752a019752.
[25] Wiederhold NP. Antifungal resistance: current trends and future strategies to combat. Infect Drug Resist 2017;10:249–59. https://doi.org/10.2147/IDR.S124918.
[26] Beardsley J, Halliday CL, Chen SCA, Sorrell TC. Responding to the emergence of antifungal drug resistance: perspectives from the bench and the bedside. Future Microbiol 2018;13:1175–91. https://doi.org/10.2217/fmb-2018-0059.
[27] Irwin KK, Renzette N, Kowalik TF, Jensen JD. Antiviral drug resistance as an adaptive process. Virus Evol 2016;2:1–10. https://doi.org/10.1093/ve/vew014.
[28] Strasfeld L, Chou S. Antiviral drug resistance: mechanisms and clinical implications. Infect Dis Clin North Am 2010;24:413–37. https://doi.org/10.1016/j.idc.2010.01.001.
[29] Pramanik PK, Alam MN, Roy Chowdhury D, Chakraborti T. Drug Resistance in Protozoan Parasites: An Incessant Wrestle for Survival. J Glob Antimicrob Resist 2019;18:1–11. https://doi.org/10.1016/j.jgar.2019.01.023.
[30] Lu F, He X-L, Richard C, Cao J. A brief history of artemisinin: Modes of action and mechanisms of resistance. Chin J Nat Med 2019;17:331–6. https://doi.org/10.1016/S1875-5364(19)30038-X.
[31] Vanaerschot M, Huijben S, Van den Broeck F, Dujardin J-C. Drug resistance in vectorborne parasites: multiple actors and scenarios for an evolutionary arms race. FEMS Microbiol Rev 2014;38:41–55. https://doi.org/10.1111/1574-6976.12032.
[32] Woodford N, Ellington MJ. The emergence of antibiotic resistance by mutation. Clin Microbiol Infect 2007;13:5–18. https://doi.org/10.1111/j.1469-0691.2006.01492.x.
[33] Gomez JE, Kaufmann-Malaga BB, Wivagg CN, Kim PB, Silvis MR, Renedo N, et al. Ribosomal mutations promote the evolution of antibiotic resistance in a multidrug environment. Elife 2017;6.. https://doi.org/10.7554/eLife.20420.
[34] Clutter DS, Jordan MR, Bertagnolio S, Shafer RW. HIV-1 drug resistance and resistance testing. Infect Genet Evol 2016;46:292–307. https://doi.org/10.1016/j.meegid.2016.08.031.
[35] He X, Wang F, Huang B, Chen P, Zhong L. Detection and analysis of resistance mutations of hepatitis B virus. Int J Clin Exp Med 2015;8:9630–9.

[36] Zhang Z-H, Wu C-C, Chen X-W, Li X, Li J, Lu M-J. Genetic variation of hepatitis B virus and its significance for pathogenesis. World J Gastroenterol 2016;22:126–44. https://doi.org/10.3748/wjg.v22.i1.126.
[37] Garcia-Rubio R, Alcazar-Fuoli L, Monteiro MC, Monzon S, Cuesta I, Pelaez T, et al. Insight into the Significance of Aspergillus fumigatus cyp51A Polymorphisms. Antimicrob Agents Chemother 2018;62. https://doi.org/10.1128/AAC.00241-18.
[38] Crofts TS, Gasparrini AJ, Dantas G. Next-generation approaches to understand and combat the antibiotic resistome. Nat Rev Microbiol 2017;15:422–34. https://doi.org/10.1038/nrmicro.2017.28.
[39] Adu-Oppong B, Gasparrini AJ, Dantas G. Genomic and functional techniques to mine the microbiome for novel antimicrobials and antimicrobial resistance genes. Ann N Y Acad Sci 2017;1388:42–58. https://doi.org/10.1111/nyas.13257.
[40] Coll F, McNerney R, Preston MD, Guerra-Assunção JA, Warry A, Hill-Cawthorne G, et al. Rapid determination of anti-tuberculosis drug resistance from whole-genome sequences. Genome Med 2015;7:51. https://doi.org/10.1186/s13073-015-0164-0.
[41] Hunt M, Mather AE, Sánchez-Busó L, Page AJ, Parkhill J, Keane JA, et al. ARIBA: rapid antimicrobial resistance genotyping directly from sequencing reads. Microb Genomics 2017;3:. https://doi.org/10.1099/mgen.0.000131e000131.
[42] Shi J, Yan Y, Links MG, Li L, Dillon J-A-R, Horsch M, et al. Antimicrobial resistance genetic factor identification from whole-genome sequence data using deep feature selection. BMC Bioinf 2019;20:535. https://doi.org/10.1186/s12859-019-3054-4.
[43] Desjardins CA, Giamberardino C, Sykes SM, Yu C-H, Tenor JL, Chen Y, et al. Population genomics and the evolution of virulence in the fungal pathogen Cryptococcus neoformans. Genome Res 2017;27:1207–19. https://doi.org/10.1101/gr.218727.116.
[44] Prasanna A, Niranjan V. Classification of Mycobacterium tuberculosis DR, MDR, XDR Isolates and Identification of Signature MutationPattern of Drug Resistance. Bioinformation 2019;15:261–8. https://doi.org/10.6026/97320630015261.
[45] Phelan JE, Lim DR, Mitarai S, de Sessions PF, Tujan MAA, Reyes LT, et al. Mycobacterium tuberculosis whole genome sequencing provides insights into the Manila strain and drug-resistance mutations in the Philippines. Sci Rep 2019;9:9305. https://doi.org/10.1038/s41598-019-45566-5.
[46] Roa MB, Tablizo FA, Morado EKD, Cunanan LF, Uy IDC, Ng KCS, et al. Whole-genome sequencing and single nucleotide polymorphisms in multidrug-resistant clinical isolates of Mycobacterium tuberculosis from the Philippines. J Glob Antimicrob Resist 2018;15:239–45. https://doi.org/10.1016/j.jgar.2018.08.009.
[47] Ramanathan B, Jindal HM, Le CF, Gudimella R, Anwar A, Razali R, et al. Next generation sequencing reveals the antibiotic resistant variants in the genome of Pseudomonas aeruginosa. PLoS ONE 2017;12:. https://doi.org/10.1371/journal.pone.0182524e0182524.
[48] Cannon ME, Mohlke KL. Deciphering the Emerging Complexities of Molecular Mechanisms at GWAS Loci. Am J Hum Genet 2018;103:637–53. https://doi.org/10.1016/j.ajhg.2018.10.001.
[49] Thanabalasingham G, Shah N, Vaxillaire M, Hansen T, Tuomi T, Gašperíková D, et al. A large multi-centre European study validates high-sensitivity C-reactive protein (hsCRP) as a clinical biomarker for the diagnosis of diabetes subtypes. Diabetologia 2011;54:2801–10. https://doi.org/10.1007/s00125-011-2261-y.
[50] Ravenhall M, Campino S, Sepúlveda N, Manjurano A, Nadjm B, Mtove G, et al. Novel genetic polymorphisms associated with severe malaria and under selective pressure in North-eastern Tanzania. PLoS Genet 2018. https://doi.org/10.1371/journal.pgen.1007172.
[51] Diaz Caballero J, Clark ST, Wang PW, Donaldson SL, Coburn B, Tullis DE, et al. A genome-wide association analysis reveals a potential role for recombination in the evolution of antimicrobial resistance in Burkholderia multivorans. PLoS Pathog 2018;14:. https://doi.org/10.1371/journal.ppat.1007453e1007453.
[52] Farhat MR, Freschi L, Calderon R, Ioerger T, Snyder M, Meehan CJ, et al. GWAS for quantitative resistance phenotypes in Mycobacterium tuberculosis reveals resistance genes and regulatory regions. Nat Commun 2019. https://doi.org/10.1038/s41467-019-10110-6.
[53] Oppong YEA, Phelan J, Perdigão J, Machado D, Miranda A, Portugal I, et al. Genome-wide analysis of Mycobacterium tuberculosis polymorphisms reveals lineage-specific associations with drug resistance. BMC Genomics 2019;20:252. https://doi.org/10.1186/s12864-019-5615-3.
[54] Sanglard D. Finding the needle in a haystack: Mapping antifungal drug resistance in fungal pathogen by genomic approaches. PLoS Pathog 2019;15:. https://doi.org/10.1371/journal.ppat.1007478e1007478.
[55] Van Camp P-J, Haslam DB, Porollo A. Bioinformatics Approaches to the Understanding of Molecular Mechanisms in Antimicrobial Resistance. Int J Mol Sci 2020;21:1363. https://doi.org/10.3390/ijms21041363.
[56] Kavvas ES, Catoiu E, Mih N, Yurkovich JT, Seif Y, Dillon N, et al. Machine learning and structural analysis of Mycobacterium tuberculosis pan-genome identifies genetic signatures of antibiotic resistance. Nat Commun 2018;9:4306. https://doi.org/10.1038/s41467-018-06634-y.
[57] Moradigaravand D, Palm M, Farewell A, Mustonen V, Warringer J, Parts L. Prediction of antibiotic resistance in Escherichia coli from large-scale pan-genome data. PLoS Comput Biol 2018;14:. https://doi.org/10.1371/journal.pcbi.1006258e1006258.
[58] Yang JH, Wright SN, Hamblin M, McCloskey D, Alcantar MA, Schrübbers L, et al. A White-Box Machine Learning Approach for Revealing Antibiotic Mechanisms of Action. Cell 2019;177(1649–1661):. https://doi.org/10.1016/j.cell.2019.04.016e9.
[59] Davis JJ, Boisvert S, Brettin T, Kenyon RW, Mao C, Olson R, et al. Antimicrobial Resistance Prediction in PATRIC and RAST. Sci Rep 2016;6:27930. https://doi.org/10.1038/srep27930.
[60] Zankari E, Allesøe R, Joensen KG, Cavaco LM, Lund O, Aarestrup FM. PointFinder: a novel web tool for WGS-based detection of antimicrobial resistance associated with chromosomal point mutations in bacterial pathogens. J Antimicrob Chemother 2017;72:2764–8. https://doi.org/10.1093/jac/dkx217.
[61] Deelder W, Christakoudi S, Phelan J, Benavente ED, Campino S, McNerney R, et al. Machine Learning Predicts Accurately Mycobacterium tuberculosis Drug Resistance From Whole Genome Sequencing Data. Front Genet 2019;10:922. https://doi.org/10.3389/fgene.2019.00922.
[62] Defelipe LA, Do Porto DF, Pereira Ramos PI, Nicolás MF, Sosa E, Radusky L, et al. A whole genome bioinformatic approach to determine potential latent phase specific targets in Mycobacterium tuberculosis. Tuberculosis (Edinb) 2016;97:181–92. https://doi.org/10.1016/j.tube.2015.11.009.
[63] Tam V, Patel N, Turcotte M, Bossé Y, Paré G, Meyre D. Benefits and limitations of genome-wide association studies. Nat Rev Genet 2019;20:467–84. https://doi.org/10.1038/s41576-019-0127-1.
[64] Gallagher MD, Chen-Plotkin AS. The Post-GWAS Era: From Association to Function. Am J Hum Genet 2018;102:717–30. https://doi.org/10.1016/j.ajhg.2018.04.002.
[65] DePristo MA, Weinreich DM, Hartl DL. Missense meanderings in sequence space: a biophysical view of protein evolution. Nat Rev Genet 2005;6:678–87. https://doi.org/10.1038/nrg1672.
[66] Gong LI, Suchard MA, Bloom JD. Stability-mediated epistasis constrains the evolution of an influenza protein. Elife 2013;2:. https://doi.org/10.7554/eLife.00631e00631.
[67] Sikosek T, Chan HS. Biophysics of protein evolution and evolutionary protein biophysics. J R Soc Interface 2014;11:20140419. https://doi.org/10.1098/rsif.2014.0419.
[68] Bloom JD, Lu Z, Chen D, Raval A, Venturelli OS, Arnold FH. Evolution favors protein mutational robustness in sufficiently large populations. BMC Biol 2007;5:29. https://doi.org/10.1186/1741-7007-5-29.
[69] Tokuriki N, Stricher F, Schymkowitz J, Serrano L, Tawfik DS. The stability effects of protein mutations appear to be universally distributed. J Mol Biol 2007;369:1318–32. https://doi.org/10.1016/j.jmb.2007.03.069.
[70] Shoichet BK, Baase WA, Kuroki R, Matthews BW. A relationship between protein stability and protein function. Proc Natl Acad Sci U S A 1995;92:452–6. https://doi.org/10.1073/pnas.92.2.452.
[71] Pandey B, Grover S, Goyal S, Jamal S, Singh A, Kaur J, et al. Novel missense mutations in gidB gene associated with streptomycin resistance in Mycobacterium tuberculosis: insights from molecular dynamics. J Biomol Struct Dyn 2019;37:20–35. https://doi.org/10.1080/07391102.2017.1417913.
[72] Thomas VL, McReynolds AC, Shoichet BK. Structural bases for stability-function tradeoffs in antibiotic resistance. J Mol Biol 2010;396:47–59. https://doi.org/10.1016/j.jmb.2009.11.005.
[73] Sun S, Selmer M, Andersson DI. Resistance to β-lactam antibiotics conferred by point mutations in penicillin-binding proteins PBP3, PBP4 and PBP6 in Salmonella enterica. PLoS ONE 2014;9:. https://doi.org/10.1371/journal.pone.0097202e97202.
[74] Contreras-Martel C, Amoroso A, Woon ECY, Zervosen A, Inglis S, Martins A, et al. Structure-guided design of cell wall biosynthesis inhibitors that overcome β-lactam resistance in Staphylococcus aureus (MRSA). ACS Chem Biol 2011;6:943–51. https://doi.org/10.1021/cb2001846.
[75] Zhang H, Deng J-Y, Bi L-J, Zhou Y-F, Zhang Z-P, Zhang C-G, et al. Characterization of Mycobacterium tuberculosis nicotinamidase/pyrazinamidase. FEBS J 2008;275:753–62. https://doi.org/10.1111/j.1742-4658.2007.06241.x.
[76] Verma JS, Gupta Y, Nair D, Manzoor N, Rautela RS, Rai A, et al. Evaluation of gidB alterations responsible for streptomycin resistance in Mycobacterium tuberculosis. J Antimicrob Chemother 2014;69:2935–41. https://doi.org/10.1093/jac/dku273.
[77] Pokorná J, Pachl P, Karlukova E, Hejdánek J, Řezáčová P, Machara A, et al. Pandemic Influenza Virus. Viruses 2009;2018:10. https://doi.org/10.3390/v10070339.
[78] Coppée R, Jeffares DC, Miteva MA, Sabbagh A, Clain J. Comparative structural and evolutionary analyses predict functional sites in the artemisinin resistance malaria protein K13. Sci Rep 2019;9:10675. https://doi.org/10.1038/s41598-019-47034-6.
[79] Phelan J, Coll F, McNerney R, Ascher DB, Pires DE V, Furnham N, et al. Mycobacterium tuberculosis whole genome sequencing and protein structure modelling provides insights into anti-tuberculosis drug resistance. BMC Med 2016;14:31. https://doi.org/10.1186/s12916-016-0575-9.
[80] Portelli S, Phelan JE, Ascher DB, Clark TG, Furnham N. Understanding molecular consequences of putative drug resistant mutations in Mycobacterium tuberculosis. Sci Rep 2018;8. https://doi.org/10.1038/s41598-018-33370-6.
[81] Karmakar M, Rodrigues CHM, Holt KE, Dunstan SJ, Denholm J, Ascher DB. Empirical ways to identify novel Bedaquiline resistance mutations in AtpE. PLoS ONE 2019;14:. https://doi.org/10.1371/journal.pone.0217169e0217169.

[82] Karmakar M, Rodrigues CHM, Horan K, Denholm JT, Ascher DB. Structure guided prediction of Pyrazinamide resistance mutations in pncA. Sci Rep 2020;10:1875. https://doi.org/10.1038/s41598-020-58635-x.
[83] Gnad F, Baucom A, Mukhyala K, Manning G, Zhang Z. Assessment of computational methods for predicting the effects of missense mutations in human cancers. BMC Genomics 2013;14:S7. https://doi.org/10.1186/1471-2164-14-S3-S7.
[84] Tang H, Thomas PD. Tools for Predicting the Functional Impact of Nonsynonymous Genetic Variation. Genetics 2016;203:635–47. https://doi.org/10.1534/genetics.116.190033.
[85] Sanavia T, Birolo G, Montanucci L, Turina P, Capriotti E, Fariselli P. Limitations and challenges in protein stability prediction upon genome variations: towards future applications in precision medicine. Comput Struct Biotechnol J 2020;18:1968–79. https://doi.org/10.1016/j.csbj.2020.07.011.
[86] Ascher DB, Wielens J, Nero TL, Doughty L, Morton CJ, Parker MW. Potent hepatitis C inhibitors bind directly to NS5A and reduce its affinity for RNA. Sci Rep 2014;4:4765. https://doi.org/10.1038/srep04765.
[87] Kano FS, Souza-Silva FA, Torres LM, Lima BAS, Sousa TN, Alves JRS, et al. The Presence, Persistence and Functional Properties of Plasmodium vivax Duffy Binding Protein II Antibodies Are Influenced by HLA Class II Allelic Variants. PLoS Negl Trop Dis 2016;10:. https://doi.org/10.1371/journal.pntd.0005177e0005177.
[88] Pires DE V, Blundell TL, Ascher DB. mCSM-lig: quantifying the effects of mutations on protein-small molecule affinity in genetic disease and emergence of drug resistance. Sci Rep 2016;6:29575. https://doi.org/10.1038/srep29575.
[89] Pires DEV, Chen J, Blundell TL, Ascher DB. In silico functional dissection of saturation mutagenesis: Interpreting the relationship between phenotypes and changes in protein stability, interactions and activity. Sci Rep 2016. https://doi.org/10.1038/srep19848.
[90] Albanaz ATS, Rodrigues CHM, Pires DE V, Ascher DB. Combating mutations in genetic disease and drug resistance: understanding molecular mechanisms to guide drug design. Expert Opin Drug Discov 2017;12:553–63. https://doi.org/10.1080/17460441.2017.1322579.
[91] Hawkey J, Ascher DB, Judd LM, Wick RR, Kostoulias X, Cleland H, et al. Evolution of carbapenem resistance in Acinetobacter baumannii during a prolonged infection. Microb Genomics 2018;4:. https://doi.org/10.1099/mgen.0.000165e000165.
[92] Holt KE, McAdam P, Thai PVK, Thuong NTT, Ha DTM, Lan NN, et al. Frequent transmission of the Mycobacterium tuberculosis Beijing lineage and positive selection for the EsxW Beijing variant in Vietnam. Nat Genet 2018;50:849–56. https://doi.org/10.1038/s41588-018-0117-9.
[93] Karmakar M, Globan M, Fyfe JAM, Stinear TP, Johnson PDR, Holmes NE, et al. Analysis of a Novel *pncA* Mutation for Susceptibility to Pyrazinamide Therapy. Am J Respir Crit Care Med 2018;198:541–4. https://doi.org/10.1164/rccm.201712-2572LE.
[94] Vedithi SC, Malhotra S, Das M, Daniel S, Kishore N, George A, et al. Structural Implications of Mutations Conferring Rifampin Resistance in Mycobacterium leprae. Sci Rep 2018. https://doi.org/10.1038/s41598-018-23423-1.
[95] Vedithi SC, Rodrigues CHM, Portelli S, Skwark MJ, Das M, Ascher DB, et al. Computational saturation mutagenesis to predict structural consequences of systematic mutations in the beta subunit of RNA polymerase in Mycobacterium leprae. Comput Struct Biotechnol J 2020;18:271–86. https://doi.org/10.1016/j.csbj.2020.01.002.
[96] Park Y, Pacitto A, Bayliss T, Cleghorn LAT, Wang Z, Hartman T, et al. Essential but Not Vulnerable: Indazole Sulfonamides Targeting Inosine Monophosphate Dehydrogenase as Potential Leads against *Mycobacterium tuberculosis*. ACS Infect Dis 2017;3:18–33. https://doi.org/10.1021/acsinfecdis.6b00103.
[97] Singh V, Donini S, Pacitto A, Sala C, Hartkoorn RC, Dhar N, et al. The Inosine Monophosphate Dehydrogenase, Gua B2, Is a Vulnerable New Bactericidal Drug Target for Tuberculosis. ACS Infect Dis 2017;3:5–17. https://doi.org/10.1021/acsinfecdis.6b00102.
[98] Trapero A, Pacitto A, Singh V, Sabbah M, Coyne AG, Mizrahi V, et al. Fragment-Based Approach to Targeting Inosine-5′-monophosphate Dehydrogenase (IMPDH) from *Mycobacterium tuberculosis*. J Med Chem 2018;61:2806–22. https://doi.org/10.1021/acs.jmedchem.7b01622.
[99] Singh V, Pacitto A, Donini S, Ferraris DM, Boros S, Illyés E, et al. Synthesis and Structure-Activity relationship of 1-(5-isoquinolinesulfonyl)piperazine analogues as inhibitors of Mycobacterium tuberculosis IMPDH. Eur J Med Chem 2019;174:309–29. https://doi.org/10.1016/J.EJMECH.2019.04.027.
[100] Kumar P, Henikoff S, Ng PC. Predicting the effects of coding non-synonymous variants on protein function using the SIFT algorithm. Nat Protoc 2009;4:1073–81. https://doi.org/10.1038/nprot.2009.86.
[101] Silk M, Petrovski S, Ascher DB. MTR-Viewer: identifying regions within genes under purifying selection. Nucleic Acids Res 2019;47:W121–6. https://doi.org/10.1093/nar/gkz457.
[102] Pires DE V, Ascher DB, Blundell TL. DUET: a server for predicting effects of mutations on protein stability using an integrated computational approach. Nucleic Acids Res 2014;42:W314–9. https://doi.org/10.1093/nar/gku411.
[103] Zerbino DR, Achuthan P, Akanni W, Amode MR, Barrell D, Bhai J, et al. Ensembl 2018. Nucleic Acids Res 2018;46:D754–61. https://doi.org/10.1093/nar/gkx1098.
[104] Choi Y, Sims GE, Murphy S, Miller JR, Chan AP. Predicting the functional effect of amino acid substitutions and indels. PLoS ONE 2012;7:. https://doi.org/10.1371/journal.pone.0046688e46688.
[105] Hecht M, Bromberg Y, Rost B. Better prediction of functional effects for sequence variants. BMC Genomics 2015;16(Suppl 8):S1. https://doi.org/10.1186/1471-2164-16-S8-S1.
[106] Ashkenazy H, Abadi S, Martz E, Chay O, Mayrose I, Pupko T, et al. ConSurf 2016: an improved methodology to estimate and visualize evolutionary conservation in macromolecules. Nucleic Acids Res 2016;44:W344–50. https://doi.org/10.1093/nar/gkw408.
[107] Portelli S, Olshansky M, Rodrigues CHM, Souza EN, Myung Y, Silk M, et al. COVID-3D: An online resource to explore the structural distribution of genetic variation in SARS-CoV-2 and its implication on therapeutic development. Nat Genet 2020;In Press. https://doi.org/10.1101/2020.05.29.124610.
[108] Kirchdoerfer RN, Ward AB. Structure of the SARS-CoV nsp12 polymerase bound to nsp7 and nsp8 co-factors. Nat Commun 2019;10. https://doi.org/10.1038/s41467-019-10280-3.
[109] Pallesen J, Wang N, Corbett KS, Wrapp D, Kirchdoerfer RN, Turner HL, et al. Immunogenicity and structures of a rationally designed prefusion MERS-CoV spike antigen. Proc Natl Acad Sci U S A 2017;114:E7348–57. https://doi.org/10.1073/pnas.1707304114.
[110] Stone EA, Sidow A. Physicochemical constraint violation by missense substitutions mediates impairment of protein function and disease severity. Genome Res 2005;15:978–86. https://doi.org/10.1101/gr.3804205.
[111] Binkley J, Karra K, Kirby A, Hosobuchi M, Stone EA, Sidow A. ProPhylER: a curated online resource for protein function and structure based on evolutionary constraint analyses. Genome Res 2010;20:142–54. https://doi.org/10.1101/gr.097121.109.
[112] Tokuriki N, Tawfik DS. Stability effects of mutations and protein evolvability. Curr Opin Struct Biol 2009;19:596–604. https://doi.org/10.1016/j.sbi.2009.08.003.
[113] Schymkowitz J, Borg J, Stricher F, Nys R, Rousseau F, Serrano L. The FoldX web server: an online force field. Nucleic Acids Res 2005;33:W382–8. https://doi.org/10.1093/nar/gki387.
[114] Kellogg EH, Leaver-Fay A, Baker D. Role of conformational sampling in computing mutation-induced changes in protein structure and stability. Proteins Struct Funct Bioinforma 2011;79:830–8. https://doi.org/10.1002/prot.22921.
[115] Dehouck Y, Kwasigroch JM, Gilis D, Rooman M. PoPMuSiC 2.1: a web server for the estimation of protein stability changes upon mutation and sequence optimality. BMC Bioinf 2011;12:151. https://doi.org/10.1186/1471-2105-12-151.
[116] Capriotti E, Fariselli P, Casadio R. I-Mutant2.0: predicting stability changes upon mutation from the protein sequence or structure. Nucleic Acids Res 2005;33:W306–10. https://doi.org/10.1093/nar/gki375.
[117] Quan L, Lv Q, Zhang Y. STRUM: structure-based prediction of protein stability changes upon single-point mutation. Bioinformatics 2016;32:2936–46. https://doi.org/10.1093/bioinformatics/btw361.
[118] Rodrigues CHM, Pires DEV, Ascher DB. DynaMut: predicting the impact of mutations on protein conformation, flexibility and stability. Nucleic Acids Res 2018;46:W350–5. https://doi.org/10.1093/nar/gky300.
[119] Rodrigues CHM, Myung Y, Pires DEV, Ascher DB. mCSM-PPI2: predicting the effects of mutations on protein–protein interactions. Nucleic Acids Res 2019;47:W338–44. https://doi.org/10.1093/nar/gkz383.
[120] Pires DEV, Ascher DB. mCSM–NA: predicting the effects of mutations on protein–nucleic acids interactions. Nucleic Acids Res 2017;45:W241–6. https://doi.org/10.1093/nar/gkx236.
[121] Pires DEV, de Melo-Minardi RC, dos Santos MA, da Silveira CH, Santoro MM, Meira W. Cutoff Scanning Matrix (CSM): structural classification and function prediction by protein inter-residue distance patterns. BMC Genomics 2011;12:S12. https://doi.org/10.1186/1471-2164-12-S4-S12.
[122] Pires DEV, Ascher DB, Blundell TL. mCSM: predicting the effects of mutations in proteins using graph-based signatures. Bioinformatics 2014;30:335–42. https://doi.org/10.1093/bioinformatics/btt691.
[123] Broom A, Jacobi Z, Trainor K, Meiering EM. Computational tools help improve protein stability but with a solubility tradeoff. J Biol Chem 2017;292:14349–61. https://doi.org/10.1074/jbc.M117.784165.
[124] Witvliet DK, Strokach A, Giraldo-Forero AFAF, Teyra J, Colak R, Kim PM. ELASPIC web-server: Proteome-wide structure-based prediction of mutation effects on protein stability and binding affinity. Bioinformatics 2016;32:1589–91. https://doi.org/10.1093/bioinformatics/btw031.
[125] Perryman AL, Lin J-H, McCammon JA. HIV-1 protease molecular dynamics of a wild-type and of the V82F/I84V mutant: Possible contributions to drug resistance and a potential new target site for drugs. Protein Sci 2004;13:1108–23. https://doi.org/10.1110/ps.03468904.
[126] Chitongo R, Obasa AE, Mikasi SG, Jacobs GB, Cloete R. Molecular dynamic simulations to investigate the structural impact of known drug resistance mutations on HIV-1C Integrase-Dolutegravir binding. PLoS ONE 2020;15. https://doi.org/10.1371/journal.pone.0223464.
[127] Honorato Siqueira T, Martínez L. Molecular simulations of fluconazole-mediated inhibition of sterol biosynthesis. J Biomol Struct Dyn 2020;38:1659–69. https://doi.org/10.1080/07391102.2019.1614998.
[128] Pimentel AL, de Lima Scodro RB, Caleffi-Ferracioli KR, Siqueira VLD, Campanerut-Sá PAZ, Lopes LDG, et al. Mutations in catalase-peroxidase KatG from isoniazid resistant Mycobacterium tuberculosis clinical isolates: insights from molecular dynamics simulations. J Mol Model 2017;23:121. https://doi.org/10.1007/s00894-017-3290-3.

[129] Sheik Amamuddy O, Musyoka TM, Boateng RA, Zabo S, Tastan BÖ. Determining the unbinding events and conserved motions associated with the pyrazinamide release due to resistance mutations of Mycobacterium tuberculosis pyrazinamidase. Comput Struct Biotechnol J 2020;18:1103–20. https://doi.org/10.1016/j.csbj.2020.05.009.
[130] Liu YY, Chen CC. Computational analysis of the molecular mechanism of RamR mutations contributing to antimicrobial resistance in salmonella enterica. Sci Rep 2017. https://doi.org/10.1038/s41598-017-14008-5.
[131] Pandey B, Grover S, Kaur J, Grover A. Analysis of mutations leading to para-aminosalicylic acid resistance in Mycobacterium tuberculosis. Sci Rep 2019;9:13617–13617. https://doi.org/10.1038/s41598-019-48940-5.
[132] Shaw DJ, Hill RE, Simpson N, Husseini FS, Robb K, Greetham GM, et al. Examining the role of protein structural dynamics in drug resistance in Mycobacterium tuberculosis. Chem Sci 2017;8:8384–99. https://doi.org/10.1039/c7sc03336b.
[133] Koivuniemi A, Fallarero A, Bunker A. Insight into the antimicrobial mechanism of action of β2,2-amino acid derivatives from molecular dynamics simulation: Dancing the can-can at the membrane surface. Biochim Biophys Acta Biomembr 2019;1861:. https://doi.org/10.1016/j.bbamem.2019.07.016183028.
[134] Silvino ACR, Costa GL, de Araújo FCF, Ascher DB, Pires DEV, Fontes CJF, et al. Variation in Human Cytochrome P-450 Drug-Metabolism Genes: A Gateway to the Understanding of Plasmodium vivax Relapses. PLoS ONE 2016;11:. https://doi.org/10.1371/journal.pone.0160172e0160172.
[135] Song T, Park Y, Shamputa IC, Seo S, Lee SY, Jeon H-S, et al. Fitness costs of rifampicin resistance in Mycobacterium tuberculosis are amplified under conditions of nutrient starvation and compensated by mutation in the β' subunit of RNA polymerase. Mol Microbiol 2014;91:1106–19. https://doi.org/10.1111/mmi.12520.
[136] Xu Z, Zhou A, Wu J, Zhou A, Li J, Zhang S, et al. Transcriptional Approach for Decoding the Mechanism of rpoC Compensatory Mutations for the Fitness Cost in Rifampicin-Resistant Mycobacterium tuberculosis. Front Microbiol 2018;9:2895. https://doi.org/10.3389/fmicb.2018.02895.
[137] Comas I, Borrell S, Roetzer A, Rose G, Malla B, Kato-Maeda M, et al. Whole-genome sequencing of rifampicin-resistant Mycobacterium tuberculosis strains identifies compensatory mutations in RNA polymerase genes. Nat Genet 2011;44:106–10. https://doi.org/10.1038/ng.1038.
[138] Carter JJ, Walker TM, Walker AS, Whitfield MG, Morlock GP, Peto TEA, et al. Prediction of pyrazinamide resistance in Mycobacterium tuberculosis using structure-based machine learning approaches. BioRxiv 2019:518142. https://doi.org/10.1101/518142.
[139] Colman PM. Influenza virus neuraminidase: Structure, antibodies, and inhibitors. Protein Sci 1994. https://doi.org/10.1002/pro.5560031007.
[140] Sosa EJ, Burguener G, Lanzarotti E, Defelipe L, Radusky L, Pardo AM, et al. Target-Pathogen: a structural bioinformatic approach to prioritize drug targets in pathogens. Nucleic Acids Res 2018;46:D413–8. https://doi.org/10.1093/nar/gkx1015.
[141] Savojardo C, Fariselli P, Martelli PL, Casadio R. INPS-MD: a web server to predict stability of protein variants from sequence and structure. Bioinformatics 2016;32:2542–4. https://doi.org/10.1093/bioinformatics/btw192.
[142] Cao H, Wang J, He L, Qi Y, Zhang JZ. DeepDDG: Predicting the Stability Change of Protein Point Mutations Using Neural Networks. J Chem Inf Model 2019;59:1508–14. https://doi.org/10.1021/acs.jcim.8b00697.
[143] Erlanson DA, McDowell RS, O'Brien T. Fragment-based drug discovery. J Med Chem 2004;47:3463–82. https://doi.org/10.1021/jm040031v.
[144] de Souza Neto LR, Moreira-Filho JT, Neves BJ, Maidana RLBR, Guimarães ACR, Furnham N, et al. In silico Strategies to Support Fragment-to-Lead Optimization in Drug Discovery. Front Chem 2020;8:93. https://doi.org/10.3389/fchem.2020.00093.
[145] Erlanson DA, Fesik SW, Hubbard RE, Jahnke W, Jhoti H. Twenty years on: the impact of fragments on drug discovery. Nat Rev Drug Discov 2016;15:605–19. https://doi.org/10.1038/nrd.2016.109.
[146] Sabbah M, Mendes V, Vistal RG, Dias DMG, Záhorszká M, Mikušová K, et al. Fragment-based design of mycobacterium tuberculosis inha inhibitors. J Med Chem 2020;63:4749–61. https://doi.org/10.1021/acs.jmedchem.0c00007.
[147] Liu M, Quinn RJ. Fragment-based screening with natural products for novel anti-parasitic disease drug discovery. Expert Opin Drug Discov 2019;14:1283–95. https://doi.org/10.1080/17460441.2019.1653849.
[148] Mello J da FR e., Gomes RA, Vital-Fujii DG, Ferreira GM, Trossini GHG. Fragment-based drug discovery as alternative strategy to the drug development for neglected diseases. Chem Biol Drug Des 2017;90:1067–78. https://doi.org/10.1111/cbdd.13030.
[149] Tang H, Thomas PD. PANTHER-PSEP: predicting disease-causing genetic variants using position-specific evolutionary preservation. Bioinformatics 2016;32:2230–2. https://doi.org/10.1093/bioinformatics/btw222.
[150] Laimer J, Hiebl-Flach J, Lengauer D, Lackner P. MAESTROweb: a web server for structure-based protein stability prediction. Bioinformatics 2016;32:1414–6. https://doi.org/10.1093/bioinformatics/btv769.
[151] Barlow KA, Ó Conchúir S, Thompson S, Suresh P, Lucas JE, Heinonen M, et al. Flex ddG: Rosetta Ensemble-Based Estimation of Changes in Protein-Protein Binding Affinity upon Mutation. J Phys Chem B 2018;122:5389–99. https://doi.org/10.1021/acs.jpcb.7b11367.

References

[1] *WHO Fact Sheet Antimicrobial Resistance*. WHO fact sheet Antimicrobial resistance. 2021. URL: https://www.who.int/news-room/fact-sheets/detail/antimicrobial-resistance (visited on 07/21/2022).

[2] O'Neill Commission. *Tackling Drug-Resistant Infections Globally-Final Report and Recommendations. The Review on Antimicrobial Resistance, Chaired by Jim O'Neill*. 2016.

[3] Antimicrobial Resistance Collaborators. "Global Burden of Bacterial Antimicrobial Resistance in 2019: A Systematic Analysis". In: *Lancet (London, England)* 399.10325 (Feb. 12, 2022), pp. 629–655. ISSN: 1474-547X. DOI: 10.1016/S0140-6736(21)02724-0.

[4] *WHO Fact Sheet HIV Drug Resistance*. WHO fact sheet HIV Drug Resistance. 2021. URL: https://www.who.int/news-room/fact-sheets/detail/hiv-drug-resistance (visited on 09/03/2022).

[5] *WHO Fact Sheet Malaria*. WHO fact sheet malaria. 2021. URL: https://www.who.int/news-room/fact-sheets/detail/malaria (visited on 09/11/2022).

[6] WHO HIV Drug Resistance Report. *HIV Drug Resistance Report*. Geneva: World Health Organization, 2021.

[7] *CDC Fact Sheet Drug-Resistant Candida Species*. CDCs 2019 Antibiotic Resistance Threats Report. 2019.

[8] *Malaria Threat Map*. Malaria Threat Map. 2021. URL: https://apps.who.int/malaria/maps/threats (visited on 07/21/2022).

[9] *World Antimicrobial Awareness Week*. World Antimicrobial Awareness Week 2020 - Handle with care: United to preserve antimicrobials. 2020. URL: https://www.who.int/news-room/events/detail/2020/11/18/default-calendar/world-antimicrobial-awareness-week-2020 (visited on 07/21/2022).

[10] Gabriela Abelenda-Alonso et al. "Antibiotic Prescription during the COVID-19 Pandemic: A Biphasic Pattern". In: *Infection Control & Hospital Epidemiology* 41.11 (Nov. 2020), pp. 1371–1372. ISSN: 0899-823X, 1559-6834. DOI: 10.1017/ice.2020.381.

[11] Bradley J. Langford et al. "Antibiotic Prescribing in Patients with COVID-19: Rapid Review and Meta-Analysis". In: *Clinical Microbiology and Infection: The Official Publication of the European Society of Clinical Microbiology and Infectious Diseases* 27.4 (Apr. 2021), pp. 520–531. ISSN: 1469-0691. DOI: 10.1016/j.cmi.2020.12.018.

[12] V. Jarlier and H. Nikaido. "Mycobacterial Cell Wall: Structure and Role in Natural Resistance to Antibiotics". In: *FEMS microbiology letters* 123.1-2 (Oct. 15, 1994), pp. 11–18. ISSN: 0378-1097. DOI: 10.1111/j.1574-6968.1994.tb07194.x.

[13] Junaid Ali, Qasim A. Rafiq, and Elizabeth Ratcliffe. "Antimicrobial Resistance Mechanisms and Potential Synthetic Treatments." In: *Future science OA* 4.4 (Apr. 2018), FSO290. ISSN: 2056-5623. DOI: 10.4155/fsoa-2017-0109.

[14] Kevin J. Forsberg et al. "Bacterial Phylogeny Structures Soil Resistomes across Habitats". In: *Nature* 509.7502 (2014), pp. 612–616. ISSN: 0028-0836. DOI: 10.1038/nature13377.

[15] Jérôme Lane et al. "A Genome-Wide Association Study of Resistance to HIV Infection in Highly Exposed Uninfected Individuals with Hemophilia A". In: *Human Molecular Genetics* 22.9 (May 1, 2013), pp. 1903–1910. ISSN: 0964-6906. DOI: 10.1093/hmg/ddt033.

[16] Francesc Coll et al. "Rapid Determination of Anti-Tuberculosis Drug Resistance from Whole-Genome Sequences". In: *Genome Medicine* 7.1 (Dec. 2015), p. 51. ISSN: 1756-994X. DOI: 10.1186/s13073-015-0164-0.

[17] Sarah K. Volkman et al. "Genome-Wide Association Studies of Drug-Resistance Determinants". In: *Trends in Parasitology* 33.3 (Mar. 2017), pp. 214–230. ISSN: 1471-5007. DOI: 10.1016/j.pt.2016.10.001.

[18] Genomics England Research Consortium. "The 100, 000 Genomes Project Protocol v4, Genomics England." In: (2017).

[19] Julio Diaz Caballero et al. "A Genome-Wide Association Analysis Reveals a Potential Role for Recombination in the Evolution of Antimicrobial Resistance in Burkholderia Multivorans". In:

PLOS Pathogens 14.12 (Dec. 7, 2018), e1007453. ISSN: 1553-7374. DOI: 10.1371/journal.ppat.1007453.

[20] Delesa Damena et al. "Genome-Wide Association Studies of Malaria Susceptibility and Resistance: Progress, Pitfalls and Prospects". In: *bioRxiv* (Nov. 2018), p. 456707. DOI: 10.1101/456707.

[21] D. J. Smith et al. "Mapping the Antigenic and Genetic Evolution of Influenza Virus". In: *Science* 305.5682 (July 2004), pp. 371–376. ISSN: 0036-8075. DOI: 10.1126/science.1097211.

[22] Jody Phelan et al. "Mycobacterium Tuberculosis Whole Genome Sequencing and Protein Structure Modelling Provides Insights into Anti-Tuberculosis Drug Resistance". In: *BMC Medicine* 14.1 (Dec. 2016), p. 31. ISSN: 1741-7015. DOI: 10.1186/s12916-016-0575-9.

[23] Malancha Karmakar et al. "Analysis of a Novel pncA Mutation for Susceptibility to Pyrazinamide Therapy". In: *American Journal of Respiratory and Critical Care Medicine* 198.4 (Aug. 15, 2018), pp. 541–544. ISSN: 1073-449X. DOI: 10.1164/rccm.201712-2572LE.

[24] James J. Davis et al. "Antimicrobial Resistance Prediction in PATRIC and RAST". In: *Scientific Reports* 6.1 (June 2016), p. 27930. ISSN: 2045-2322. DOI: 10.1038/srep27930.

[25] Erol S. Kavvas et al. "Machine Learning and Structural Analysis of Mycobacterium Tuberculosis Pan-Genome Identifies Genetic Signatures of Antibiotic Resistance". In: *Nature Communications* 9.1 (Dec. 2018), p. 4306. ISSN: 2041-1723. DOI: 10.1038/s41467-018-06634-y.

[26] Malancha Karmakar et al. *SUSPECT-PZA | Home*. 2018.

[27] Malancha Karmakar et al. "Empirical Ways to Identify Novel Bedaquiline Resistance Mutations in AtpE". In: *PloS One* 14.5 (2019), e0217169. ISSN: 1932-6203. DOI: 10.1371/journal.pone.0217169.

[28] Stephanie Portelli et al. "Prediction of Rifampicin Resistance beyond the RRDR Using Structure-Based Machine Learning Approaches". In: *Scientific Reports* 10.1 (1 Oct. 22, 2020), p. 18120. ISSN: 2045-2322. DOI: 10.1038/s41598-020-74648-y.

[29] Ruben Cloete et al. "Molecular Modelling and Simulation Studies of the Mycobacterium Tuberculosis Multidrug Efflux Pump Protein Rv1258c". In: *PLOS ONE* 13.11 (Nov. 2018). Ed. by Claudio M. Soares, e0207605. ISSN: 1932-6203. DOI: 10.1371/journal.pone.0207605.

[30] Joshua J Carter et al. "Prediction of Pyrazinamide Resistance in Mycobacterium Tuberculosis Using Structure-Based Machine Learning Approaches". In: *bioRxiv* (Apr. 2019), p. 518142. DOI: 10.1101/518142.

[31] Brian W. Matthews. "Structural and Genetic Analysis of Protein Stability". In: *Annual Review of Biochemistry* 62.1 (June 1993), pp. 139–160. ISSN: 0066-4154. DOI: 10.1146/annurev.bi.62.070193.001035.

[32] Alan R. Fersht. "Dissection of the Structure and Activity of the Tyrosyl-tRNA Synthetase by Site-Directed Mutagenesis". In: *Biochemistry* 26.25 (Dec. 1987), pp. 8031–8037. ISSN: 0006-2960. DOI: 10.1021/bi00399a001.

[33] Lijun Quan, Qiang Lv, and Yang Zhang. "STRUM: Structure-Based Prediction of Protein Stability Changes upon Single-Point Mutation". In: *Bioinformatics* 32.19 (Oct. 2016), pp. 2936–2946. ISSN: 1367-4803. DOI: 10.1093/bioinformatics/btw361.

[34] Raphael Guerois, Jens Erik Nielsen, and Luis Serrano. "Predicting Changes in the Stability of Proteins and Protein Complexes: A Study of More Than 1000 Mutations". In: *Journal of Molecular Biology* 320.2 (July 2002), pp. 369–387. ISSN: 00222836. DOI: 10.1016/S0022-2836(02)00442-4.

[35] Yves Dehouck et al. "BeAtMuSiC: Prediction of Changes in ProteinProtein Binding Affinity on Mutations". In: *Nucleic Acids Research* 41.W1 (July 2013), W333–W339. ISSN: 1362-4962. DOI: 10.1093/nar/gkt450.

[36] Douglas Pires, Tom L. Blundell, and David B. Ascher. "mCSM-lig: Quantifying the Effects of Mutations on Protein-Small Molecule Affinity in Genetic Disease and Emergence of Drug Resistance". In: *Scientific Reports* 6.1 (Sept. 2016), p. 29575. ISSN: 2045-2322. DOI: 10.1038/srep29575.

[37] Douglas E V Pires and David B Ascher. "mCSM-NA: Predicting the Effects of Mutations on Protein-Nucleic Acids Interactions". In: *Nucleic acids research* 45.W1 (July 2017), W241–W246. ISSN: 0305-1048. DOI: 10.1093/nar/gkx236.

[38] Carlos H M Rodrigues et al. "mCSM-PPI2: Predicting the Effects of Mutations on ProteinProtein Interactions". In: *Nucleic Acids Research* 47.W1 (May 2019), W338–W344. ISSN: 0305-1048. DOI: 10.1093/nar/gkz383.

[39] Shaji M D Kumar et al. "ProTherm and ProNIT: Thermodynamic Databases for Proteins and Protein-Nucleic Acid Interactions and 4 Tsukuba Materials Information". In: *Nucleic Acids Res* 34 (Suppl. 1 2006), pp. D204–D206. DOI: 10.1093/nar/gkj103.

[40] Iain H Moal and Juan Ferná Ndez-Recio. "Structural Bioinformatics SKEMPI: A Structural Kinetic and Energetic Database of Mutant Protein Interactions and Its Use in Empirical Models". In: 28.20 (2012), pp. 2600–2607. DOI: 10.1093/bioinformatics/bts489.

[41] Douglas E.V. Pires, Tom L. Blundell, and David B. Ascher. "Platinum: A Database of Experimentally Measured Effects of Mutations on Structurally Defined ProteinLigand Complexes". In: *Nucleic Acids Research* 43.D1 (Jan. 28, 2015), pp. D387–D391. ISSN: 0305-1048. DOI: 10.1093/nar/gku966.

[42] Emidio Capriotti and Russ B Altman. *Improving the Prediction of Disease-Related Variants Using Protein Three-Dimensional Structure*. S4. BioMed Central, Dec. 2011, S3. DOI: 10.1186/1471-2105-12-S4-S3.

[43] Emidio Capriotti et al. "Predicting Protein Stability Changes from Sequences Using Support Vector Machines". In: *Bioinformatics* 21 (Suppl 2 Sept. 2005), pp. ii54–ii58. ISSN: 1367-4803. DOI: 10.1093/bioinformatics/bti1109.

[44] Emidio Capriotti, Piero Fariselli, and Rita Casadio. "I-Mutant2.0: Predicting Stability Changes upon Mutation from the Protein Sequence or Structure". In: *Nucleic Acids Res* 33 (Suppl.2 2005), W306–W310. DOI: 10.1093/nar/gki375.

[45] Tanushree Tunstall et al. "Combining Structure and Genomics to Understand Antimicrobial Resistance". In: *Computational and Structural Biotechnology Journal* 18 (2020), pp. 3377–3394. ISSN: 2001-0370. DOI: 10.1016/j.csbj.2020.10.017.

[46] WHO Global Tuberculosis Report. *Global Tuberculosis Report 2021*. Geneva: World Health Organization, 2021.

[47] Edine W. Tiemersma et al. "Natural History of Tuberculosis: Duration and Fatality of Untreated Pulmonary Tuberculosis in HIV Negative Patients: A Systematic Review". In: *PloS One* 6.4 (Apr. 4, 2011), e17601. ISSN: 1932-6203. DOI: 10.1371/journal.pone.0017601.

[48] WHO. *Guidelines for the Management of Drug-Resistant Tuberculosis*. 1996.

[49] WHO. *Guidelines for the Programmatic Management of Drug-Resistant Tuberculosis: Emergency Update 2008*. World Health Organization, 2008.

[50] M. WHO Treatment guidelines update. *WHO Guidelines for the Programmatic Management of Drug-Resistant Tuberculosis: 2011 Update*. 2011, pp. 516–528.

[51] WHO Treatment guidelines. *WHO Treatment Guidelines for Drug- Resistant Tuberculosis*. October. 2016.

[52] Simon Tiberi et al. "Classifying New Anti-Tuberculosis Drugs: Rationale and Future Perspectives". In: *International Journal of Infectious Diseases* 56 (2017), pp. 181–184. ISSN: 18783511. DOI: 10.1016/j.ijid.2016.10.026.

[53] DR Silva et al. "New and Repurposed Drugs to Treat Multidrug- and Extensively Drug-Resistant Tuberculosis". In: *J Bras Pneumol* 44 (2018), pp. 153–160.

[54] WHO consolidated guidelines DR-TB. *WHO Consolidated Guidelines on Drug-Resistant Tuberculosis Treatment*. Geneva: World Health Organization, 2019.

[55] WHO. *Global Tuberculosis Report 2018*. 2018.

[56] WHO Global Tuberculosis Report. *Global Tuberculosis Report 2020*. Geneva: World Health Organization, 2020.

[57] *WHO Drug-resistant TB: Treatment and Treatment Coverage*. 3.4 Drug-resistant TB: treatment and treatment coverage. 2021. URL: https://www.who.int/publications/digital/global-

tuberculosis-report-2021/tb-diagnosis-treatment/drug-resistant-treatment (visited on 07/23/2022).

[58] S. T. Cole et al. "Deciphering the Biology of Mycobacterium Tuberculosis from the Complete Genome Sequence". In: *Nature* (1998). ISSN: 00280836. DOI: 10.1038/31159.

[59] Stephen H Gillespie. "Evolution of Drug Resistance in Mycobacterium Tuberculosis: Clinical and Molecular Perspective." In: *Antimicrobial agents and chemotherapy* 46.2 (Feb. 2002), pp. 267–74. ISSN: 0066-4804. DOI: 10.1128/aac.46.2.267-274.2002.

[60] Navisha Dookie et al. "Evolution of Drug Resistance in Mycobacterium Tuberculosis: A Review on the Molecular Determinants of Resistance and Implications for Personalized Care". In: *Journal of Antimicrobial Chemotherapy* 73.5 (2018), pp. 1138–1151. ISSN: 14602091. DOI: 10.1093/jac/dkx506.

[61] Mashael Al-Saeedi and Sahal Al-Hajoj. "Diversity and Evolution of Drug Resistance Mechanisms in Mycobacterium Tuberculosis". In: *Infection and Drug Resistance* Volume 10 (Oct. 2017), pp. 333–342. ISSN: 1178-6973. DOI: 10.2147/IDR.S144446.

[62] Quang Huy Nguyen et al. "Insights into the Processes That Drive the Evolution of Drug Resistance in Mycobacterium Tuberculosis". In: *Evolutionary Applications* 11.9 (2018), pp. 1498–1511. ISSN: 1752-4571. DOI: 10.1111/eva.12654.

[63] Yaa E.A. Oppong et al. "Genome-Wide Analysis of Mycobacterium Tuberculosis Polymorphisms Reveals Lineage-Specific Associations with Drug Resistance". In: *BMC Genomics* (2019). ISSN: 14712164. DOI: 10.1186/s12864-019-5615-3.

[64] Ying Zhang. "THE MAGIC BULLETS AND TUBERCULOSIS DRUG TARGETS". In: *Annual Review of Pharmacology and Toxicology* 45.1 (Feb. 2005), pp. 529–564. ISSN: 0362-1642. DOI: 10.1146/annurev.pharmtox.45.120403.100120.

[65] Jerrold J. Ellner. "The Emergence of Extensively Drug-Resistant Tuberculosis: A Global Health Crisis Requiring New Interventions: Part I: The Origins and Nature of the Problem". In: *Clinical and Translational Science* 1.3 (2008), pp. 249–254. ISSN: 1752-8062. DOI: 10.1111/j.1752-8062.2008.00060.x.

[66] M.W. Borgdorff and D. van Soolingen. "The Re-Emergence of Tuberculosis: What Have We Learnt from Molecular Epidemiology?" In: *Clinical Microbiology and Infection* 19.10 (Oct. 2013), pp. 889–901. ISSN: 1198743X. DOI: 10.1111/1469-0691.12253.

[67] Christoph Lange et al. "Management of Drug-Resistant Tuberculosis". In: *The Lancet* 394.10202 (Sept. 14, 2019), pp. 953–966. ISSN: 0140-6736, 1474-547X. DOI: 10.1016/S0140-6736(19)31882-3.

[68] Veronica Melchionda et al. "Amikacin Treatment for Multidrug Resistant Tuberculosis: How Much Monitoring Is Required?" In: *The European respiratory journal* 42.4 (Oct. 2013), pp. 1148–50. ISSN: 1399-3003. DOI: 10.1183/09031936.00184312.

[69] James E Gomez and John D McKinney. "M. Tuberculosis Persistence, Latency, and Drug Tolerance". In: *Tuberculosis* 84.1-2 (Jan. 2004), pp. 29–44. ISSN: 1472-9792. DOI: 10.1016/J.TUBE.2003.08.003.

[70] Martin Gengenbacher and Stefan H.E. Kaufmann. "Mycobacterium Tuberculosis: Success through Dormancy". In: *FEMS Microbiology Reviews* 36.3 (May 2012), pp. 514–532. ISSN: 1574-6976. DOI: 10.1111/j.1574-6976.2012.00331.x.

[71] M G Reynolds. "Compensatory Evolution in Rifampin-Resistant Escherichia Coli." In: *Genetics* 156.4 (Dec. 2000), pp. 1471–81. ISSN: 0016-6731.

[72] Pedro Eduardo Almeida da Silva et al. "Efflux as a Mechanism for Drug Resistance in \textlessi\textgreaterMycobacterium Tuberculosis\textless/I\textgreater : Table 1". In: *FEMS Immunology & Medical Microbiology* 63.1 (Oct. 2011), pp. 1–9. ISSN: 0928-8244. DOI: 10.1111/j.1574-695X.2011.00831.x.

[73] P.J Brennan. "Structure, Function, and Biogenesis of the Cell Wall of Mycobacterium Tuberculosis". In: *Tuberculosis* 83.1-3 (Feb. 2003), pp. 91–97. ISSN: 1472-9792. DOI: 10.1016/S1472-9792(02)00089-6.

[74] Mary Jackson et al. "Inactivation of the Antigen 85C Gene Profoundly Affects the Mycolate Content and Alters the Permeability of the Mycobacterium Tuberculosis Cell Envelope". In:

Molecular Microbiology 31.5 (Mar. 1999), pp. 1573–1587. ISSN: 0950-382X. DOI: 10.1046/j.1365-2958.1999.01310.x.

[75] Wen-xi Xu et al. "The Wag31 Protein Interacts with AccA3 and Coordinates Cell Wall Lipid Permeability and Lipophilic Drug Resistance in Mycobacterium Smegmatis". In: *Biochemical and Biophysical Research Communications* 448.3 (June 2014), pp. 255–260. ISSN: 0006-291X. DOI: 10.1016/J.BBRC.2014.04.116.

[76] Radhika Gupta et al. "The Mycobacterium Tuberculosis Protein Ldt Mt2 Is a Nonclassical Transpeptidase Required for Virulence and Resistance to Amoxicillin". In: *Nature Medicine* (2010). ISSN: 10788956. DOI: 10.1038/nm.2120.

[77] M. K. Schoonmaker, W. R. Bishai, and G. Lamichhane. "Nonclassical Transpeptidases of Mycobacterium Tuberculosis Alter Cell Size, Morphology, the Cytosolic Matrix, Protein Localization, Virulence, and Resistance to -Lactams". In: *Journal of Bacteriology* 196.7 (Apr. 2014), pp. 1394–1402. ISSN: 0021-9193. DOI: 10.1128/JB.01396-13.

[78] N. Siddiqi et al. "Mycobacterium Tuberculosis Isolate with a Distinct Genomic Identity Overexpresses a Tap-Like Efflux Pump". In: *Infection* 32.2 (Apr. 2004), pp. 109–111. ISSN: 0300-8126. DOI: 10.1007/s15010-004-3097-x.

[79] Meenakshi Balganesh et al. "Efflux Pumps of Mycobacterium Tuberculosis Play a Significant Role in Antituberculosis Activity of Potential Drug Candidates". In: *Antimicrobial Agents and Chemotherapy* 56.5 (May 2012), pp. 2643–2651. ISSN: 0066-4804. DOI: 10.1128/AAC.06003-11.

[80] Francesc Coll et al. "Genome-Wide Analysis of Multi- and Extensively Drug-Resistant Mycobacterium Tuberculosis". In: *Nature Genetics* 50.2 (Feb. 2018), pp. 307–316. ISSN: 1061-4036. DOI: 10.1038/s41588-017-0029-0.

[81] Karolína Buriánková et al. "Molecular Basis of Intrinsic Macrolide Resistance in the Mycobacterium Tuberculosis Complex." In: *Antimicrobial agents and chemotherapy* 48.1 (Jan. 2004), pp. 143–50. ISSN: 0066-4804. DOI: 10.1128/aac.48.1.143-150.2004.

[82] C. E. Maus, B. B. Plikaytis, and T. M. Shinnick. "Mutation of tlyA Confers Capreomycin Resistance in Mycobacterium Tuberculosis". In: *Antimicrobial Agents and Chemotherapy* 49.2 (Feb. 2005), pp. 571–577. ISSN: 0066-4804. DOI: 10.1128/AAC.49.2.571-577.2005.

[83] A Somoskovi, L M Parsons, and M Salfinger. "The Molecular Basis of Resistance to Isoniazid, Rifampin, and Pyrazinamide in Mycobacterium Tuberculosis." In: *Respiratory research* 2.3 (2001), pp. 164–8. ISSN: 1465-9921.

[84] Elena Segala et al. "New Mutations in the Mycobacterial ATP Synthase: New Insights into the Binding of the Diarylquinoline TMC207 to the ATP Synthase C-Ring Structure". In: *Antimicrobial Agents and Chemotherapy* 56.5 (May 2012), pp. 2326–2334. ISSN: 0066-4804. DOI: 10.1128/AAC.06154-11.

[85] S. Boonaiam et al. "Genotypic Analysis of Genes Associated with Isoniazid and Ethionamide Resistance in MDR-TB Isolates from Thailand". In: *Clinical Microbiology and Infection* 16.4 (Apr. 2010), pp. 396–399. ISSN: 1198743X. DOI: 10.1111/j.1469-0691.2009.02838.x.

[86] Sònia Borrell et al. "Epistasis between Antibiotic Resistance Mutations Drives the Evolution of Extensively Drug-Resistant Tuberculosis". In: *Evolution, Medicine, and Public Health* 2013.1 (Jan. 1, 2013), pp. 65–74. ISSN: 2050-6201. DOI: 10.1093/emph/eot003.

[87] M. de Vos et al. "Putative Compensatory Mutations in the rpoC Gene of Rifampin-Resistant Mycobacterium Tuberculosis Are Associated with Ongoing Transmission". In: *Antimicrobial Agents and Chemotherapy* 57.2 (Feb. 2013), pp. 827–832. ISSN: 0066-4804. DOI: 10.1128/AAC.01541-12.

[88] Iñaki Comas et al. "Whole-Genome Sequencing of Rifampicin-Resistant Mycobacterium Tuberculosis Strains Identifies Compensatory Mutations in RNA Polymerase Genes". In: *Nature Genetics* 44.1 (2012), pp. 106–110. ISSN: 10614036. DOI: 10.1038/ng.1038.

[89] Keira A. Cohen et al. "Evolution of Extensively Drug-Resistant Tuberculosis over Four Decades: Whole Genome Sequencing and Dating Analysis of Mycobacterium tuberculosis Isolates from KwaZulu-Natal". In: *PLoS medicine* 12.9 (Sept. 2015), e1001880. ISSN: 1549-1277. DOI: 10.1371/journal.pmed.1001880.

[90] A. Koch, V. Mizrahi, and D. Warner. "The Impact of Drug Resistance on Mycobacterium Tuberculosis Physiology: What Can We Learn from Rifampicin?" In: *Emerging Microbes & Infections* (2014). DOI: 10.1038/emi.2014.17.

[91] Pierre-Alexis Gros, Hervé Le Nagard, and Olivier Tenaillon. "The Evolution of Epistasis and Its Links With Genetic Robustness, Complexity and Drift in a Phenotypic Model of Adaptation". In: *Genetics* 182.1 (May 2009), pp. 277–293. ISSN: 0016-6731. DOI: 10.1534/genetics.108.099127.

[92] Abigail L. Manson et al. "Genomic Analysis of Globally Diverse Mycobacterium Tuberculosis Strains Provides Insights into the Emergence and Spread of Multidrug Resistance". In: *Nature Genetics* 49.3 (Mar. 2017), pp. 395–402. ISSN: 1546-1718. DOI: 10.1038/ng.3767.

[93] Daniel M. Weinreich, Richard A. Watson, and Lin Chao. "Perspective: Sign Epistasis and Genetic Constraint on Evolutionary Trajectories". In: *Evolution; International Journal of Organic Evolution* 59.6 (June 2005), pp. 1165–1174. ISSN: 0014-3820.

[94] Patrick C. Phillips. "Epistasis the Essential Role of Gene Interactions in the Structure and Evolution of Genetic Systems". In: *Nature Reviews Genetics* 9.11 (11 Nov. 2008), pp. 855–867. ISSN: 1471-0064. DOI: 10.1038/nrg2452.

[95] J. Kester and S. Fortune. "Persisters and beyond: Mechanisms of Phenotypic Drug Resistance and Drug Tolerance in Bacteria". In: *Critical reviews in biochemistry and molecular biology* (2014). DOI: 10.3109/10409238.2013.869543.

[96] Travis E. Hartman et al. "Metabolic Perspectives on Persistence". In: *Microbiology Spectrum* 5.1 (Feb. 2017), p. 5.1.16. DOI: 10.1128/microbiolspec.TBTB2-0026-2016.

[97] Jae Jin Lee et al. "Transient Drug-Tolerance and Permanent Drug-Resistance Rely on the Trehalose-Catalytic Shift in Mycobacterium Tuberculosis". In: *Nature Communications* 10.1 (1 July 2, 2019), p. 2928. ISSN: 2041-1723. DOI: 10.1038/s41467-019-10975-7.

Chapter 2

Methods

2.1 Mutation Dataset

The dataset of mutations containing nsSNVs from a genome wide association study (GWAS) of 35,944 *M. tuberculosis* isolates described recently[1] was used to extract the SAVs. These globally diverse, clinical isolates are comprised of the seven main lineages (1, 5, and 6: ancient; 2, 3, and 4: modern; 7: intermediate). Additional metadata relating to these isolates included: drug susceptibility testing (DST) across the eight first-and second line anti-TB drugs, and their link with drug resistance as per the TB-Profiler mutation database.[2] Mutations in this thesis refer to those associated with single amino acid variation (SAV), and as such the terms mutations, SAVs, or SAV mutations will be used interchangeably throughout. Only such mutations which occurred in the protein coding region of the six TB *gene*-drug targets were considered in this project. These were *alr*-cycloserine (DCS), *embB*-ethambutol (EMB), *gidB*-streptomycin (STR), *katG*-isoniazid (INH), *pncA*-pyrazinamide (PZA), *rpoB*-rifampicin (RFP).

These six targets were chosen based on the availability of SAV mutations from the GWAS dataset, 3D protein structure, coverage of both first and second line drugs including diversity of target types (i.e direct and indirect drug targets). Genes *alr*, *embB*, and *rpoB* encode proteins that are considered essential as their respective drugs directly bind to them, while genes *gidB*, *katG*, and *pncA* are indirect targets for their respective drugs. INH and PZA which bind to KatG and PncA respectively are prodrugs which need to be converted into their active form to exert their antibacterial action. GidB is an ancillary protein encoding a S-adenosyl methionine (SAM)-dependent 7-methylguanosine (m7G) methyltransferase enzyme required for the methylation of position G527 in the 16S rRNA required for STR binding.[3]

All SAVs related to these gene-drug targets were extracted using custom Python scripts. The missing values for DST data (designated as 0: Sensitive, 1: Resistant) for each gene-drug target were imputed using a knowledge-based approach: Firstly, for each SAV where DST data was missing, the corresponding TB-Profiler annotation reflecting its link to drug resistance was considered. Thereafter, the mode value of DST for the given SAV across its samples was taken to be the consensus DST value for that SAV. Where a given SAV had a non-unique mode, a DST value corresponding to resistance was prioritised over sensitive. In this manner an aggregate or consensus estimate of DST was obtained to classify each SAV as resistant or sensitive.

Nearly 80% of the clinical isolates (n=28,217/35,944) exhibited SAVs in one or more of the six gene targets attributable to drug resistance. Subsequent downstream analyses performed in the individual

chapters further attempted to determine the proportion of SAVs associated with resistance to the candidate drug.

2.2 Gene-target sequences

Sequences for all six gene-targets as per the *M. tuberculosis* (H37Rv) reference genome were obtained from the Mycobrowser database.[4] The respective gene loci for these targets are: *alr*:Rv3423c, *embB*:Rv3795, *gidB*:Rv3919c, *katG*:Rv1908c, *pncA*:Rv2043c, *rpoB*:Rv0667.

2.3 Structural modelling

The target-drug complex is preferentially sourced from the organism of interest i.e. *M. tuberculosis* in this project, from the RCSB Protein Data Bank database (PDB)[5] of experimentally determined macromolecular structures. In the absence of an available experimentally determined structure, homology modelling was conducted to reconstruct the biological unit of the protein under investigation. For the experimentally determined structures that lacked the interacting ligand/drug and the modelled structures, the small molecules were docked to generate a model of the protein-ligand complex.

2.3.1 Molecular docking

In the absence of a protein-drug complex in *M. tuberculosis* for PncA-PZA and GidB-SRY, molecular docking was performed to facilitate analysis by computational tools estimating mutational impact on ligand affinity. Docking is a computational modelling technique commonly used to predict possible ligand poses/orientation in bound conformations with a target. The bound conformations are associated with their respective predicted binding affinity values. Binding affinity (strength of the ligand interaction with its target) is based on one of several scoring functions, which rank the poses in increasing order of predicted binding affinity.

The software AutoDock Vina version 1.1.2[6] an open source molecular modelling platform, was used to perform the molecular docking of PncA-PZA and GidB-AMP-SAM (described below). The scoring function in AutoDock Vina consists of a conformation-dependent and a conformation-independent component. The conformation-dependent scoring function considers the sum of all the atom pairs which can move relative to each other. Optimisation is performed via the Iterated Local Search global optimiser, which considers position, orientation, and torsion scoring values to generate minima for use during refinement. Binding free energy is calculated using a semi-empirical force field, combining experimental and knowledge-based information. The binding affinity returned by AutoDock Vina is in

Kcal/mol. However in order to use the mCSM-lig tool, binding affinity must be provided in nanomolar (nM). Therefore, the conversion of binding affinity (depicted as K in Equation 1) from Kcal/mol to nM was performed using the equation:

$$\begin{aligned} lnK &= -\Delta G/RT \\ K &= e^{-\Delta G/RT} \end{aligned} \tag{1}$$

Equation 1: Calculation of the dissociation constant (K) associated with binding affinity, where ΔG is the binding free energy, R is the gas constant, 1.987 cal K^{-1} mol^{-1}, and T is the absolute temperature, 298 K. Adapted from Morris, *et. al.* 1998.[7]

The 3D structure of the ligand was sourced from the PDB, and protonation was carried out using UCSF Chimera version 1.14.[8] Identification of rotatable bonds for ligands were carried out in AutoDock Tools version 4.2,[9] where protonation of the ligand is specifically required by AutoDock Vina.[6] Similarly, the removal of explicit solvent, structure minimisation, and other steps per the standard protocol of AutoDock Tools[9] were carried out accordingly. The overall docking workflow is shown in **Figure 1.** The parameters of the configuration file required by AutoDock Vina are explained in **Table 1**.

PatchDock: PatchDock[10,11] was used for docking RNA with *M. tuberculosis* GidB (See section 2.3.2.3). It is available as a web service and was run with default settings. PatchDock is an algorithm based on shape complementarity criteria, and uses the principles of object recognition and image segmentation.

The docking poses were visualised according to the occupation of the search space and diversity of pose conformations in UCSF Chimera version 1.14,[8] PyMol version 2.4.0[12] further analysed using the Protein Ligand Interaction Profiler (PLIP)[13] and Arpeggio.[14]

2.3.2 Target-Drug complexes

2.3.2.1 Existing crystallographic complexes

EmbB-EMB complex: The *M. tuberculosis* EmbB binds to the anti-TB drug ethambutol (EMB). The cryo-EM structure of EmbB-EMB bound complex was obtained from PDB-ID 7BVF.[15]

RpoB RNAP β subunit-RFP complex: The *M. tuberculosis* RpoB RNA polymerase (RNAP) β subunit binds to the anti-TB drug rifampicin, also known as rifampin (RFP). The crystal structure of *M. tuberculosis* RpoB RNAP β subunit bound with RFP was obtained from PDB-ID 5UHC.[16]

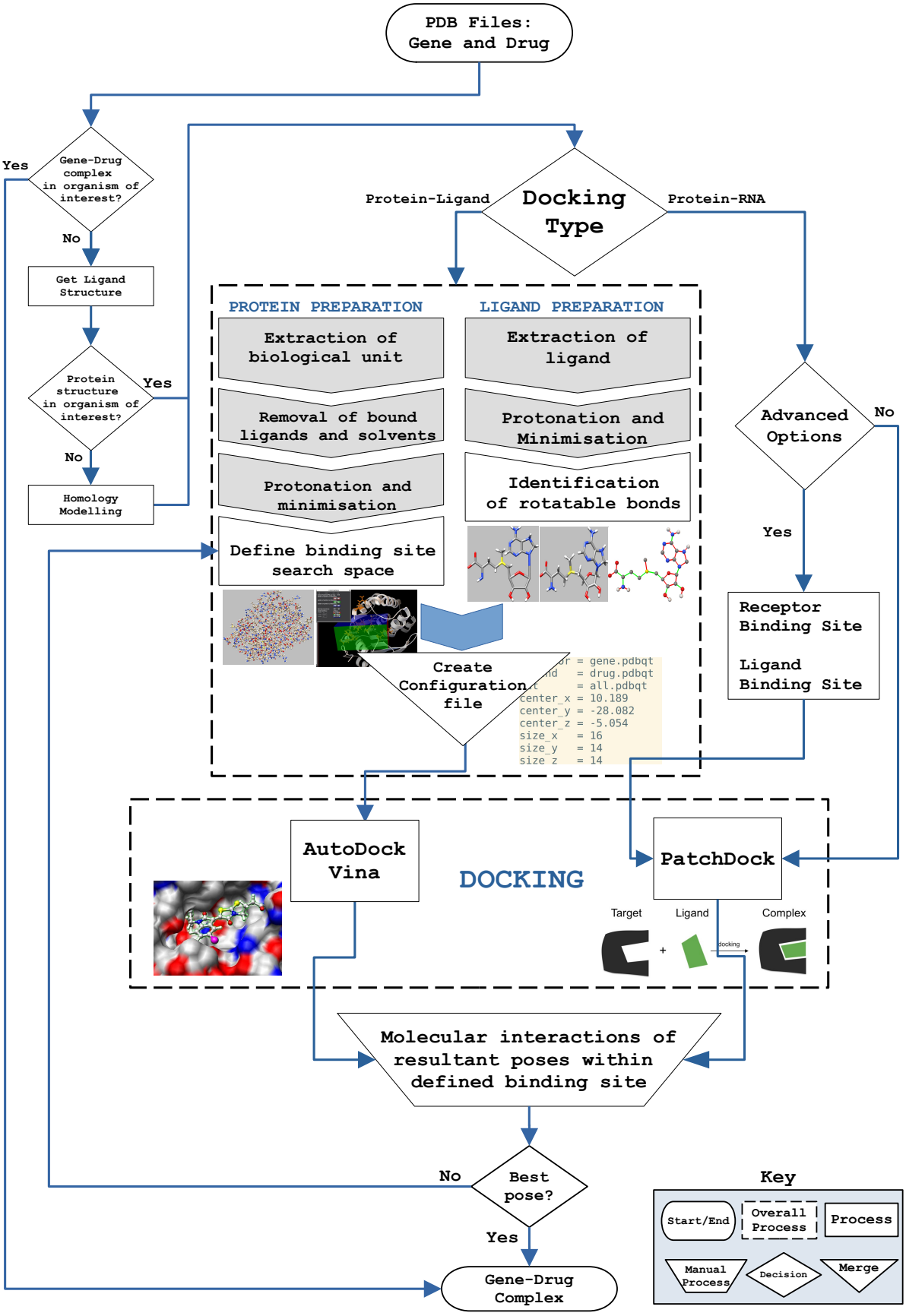

PDB Files:
Gene and Drug
Gene-Drug complex in organism of interest?
Yes
No
Get Ligand Structure
Protein structure in organism of interest?
Yes
No
Homology Modelling
Docking Type
Protein-Ligand
Protein-RNA
PROTEIN PREPARATION
Extraction of biological unit
Removal of bound ligands and solvents
Protonation and minimisation
Define binding site search space
LIGAND PREPARATION
Extraction of ligand
Protonation and Minimisation
Identification of rotatable bonds
Create Configuration file
receptor = gene.pdbqt
ligand = drug.pdbqt
out = all.pdbqt
center_x = 10.189
center_y = -28.082
center_z = -5.054
size_x = 16
size_y = 14
size_z = 14
Advanced Options
No
Yes
Receptor Binding Site
Ligand Binding Site
AutoDock Vina
DOCKING
PatchDock
Target
Ligand
Complex
docking
Molecular interactions of resultant poses within defined binding site
No
Best pose?
Yes
Gene-Drug Complex
Key
Start/End
Overall Process
Process
Manual Process
Decision
Merge

Figure 1: General docking workflow applied in the project
The 3D structure of the protein (gene) and ligand (drug) complex was obtained from the Protein Data Bank (PDB), preferentially one from *M. tuberculosis*. In the absence of a bound complex, the individual structures were also obtained from PDB, again, preferentially from *M. tuberculosis*. Homology modelling was performed where the 3D structure was not available for *M. tuberculosis*. Following this, molecular docking for protein-ligand was performed using AutoDock Vina as per the standard protocol summarised in the workflow. For Protein-RNA docking, PatchDock was used with Advanced options to guide docking. The final pose was selected based on molecular interactions identified using the PLIP web server.

Parameters	Definition	Default value
receptor	Name of file containing protein structure with solvent removed	none
ligand	Name of the file containing the protonated ligand structure	none
out	Output file containing the different poses	none
center_x/y/z & size_x/y/z	Search space coordinates used by AutoDock Vina to guide the docking of the ligand	none
energy_range	The maximum energy range accepted for the resultant poses, i.e. automatically excluding extreme poses	3kcal/mol
exhaustiveness	The number of optimisations run to find the best possible poses in the given search space	8 runs
num_nodes	The maximum number of poses generated	9 poses

Table 1: Description of the parameters required by AutoDock Vina

2.3.2.2 Docked complexes: other sources

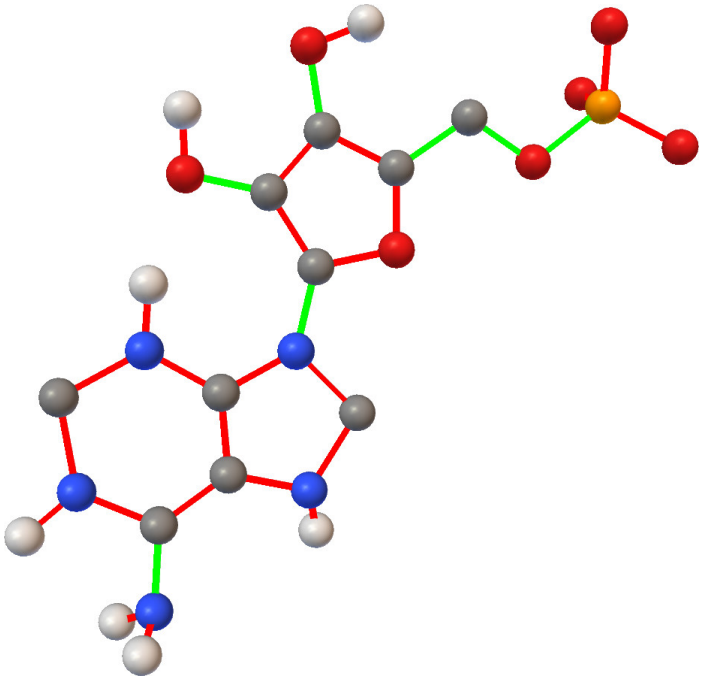
Alr-DCS and KatG-INH: The protein structures of Alr, and KatG were available as crystallographic structures in their *apo* form with PDB-IDs 1XFC[17] and 1SJ2[18] respectively. While Alr binds to the anti-TB drug cycloserine (DCS), KatG binds to the anti-TB drug isoniazid (INH). The docked target-drug complexes were sourced from collaborators for consistency, and are described elsewhere.[19] Briefly, the ligand structure for DCS (covalently bound to pyridoxal 5-phosphate (PLP), DCS-PLP was obtained from holo-homologue structure with PDB-ID 4LUT,[20] while that of INH was obtained similarly from PDB-ID 5SYJ[21] to guide docking of the ligands onto their respective crystal structures. The best binding pose was chosen based on RMSD comparison with the homologue-bound ligands and Arpeggio,[14] and analysis of the active site interactions.[19]

2.3.2.3 Docked complexes: Molecular docking

GidB-STR complex: The target for anti-TB drug streptomycin (STR) is *GidB* or *Gid*, and was formerly known as Ribosomal RNA small subunit methyltransferase G (RsmG). In the absence of a crystal structure of GidB in *M. tuberculosis* until recently (PDB-ID 7CFE, but is yet to be published), the structural modelling was carried out in several stages. Firstly, the structure of *M. tuberculosis* GidB was obtained from the web-accessible *M. tuberculosis* structural and functional proteome database Chopin,[22] and its sequence compared with the one from Mycobrowser's database.[4] A minor discrepancy between the two sequences was noted, where 'F100' in Mycobrowser was reflected as 'S100' in the Chopin database. This was resolved by remodelling the GidB structure using Modeller version 9.25[23] with the sequence from Chopin as the template, and the one from Mycobrowser as the target. The best model was chosen based on the lowest DOPE score and intra-model hydrogen bond interaction. The modelled GidB structure was superimposed on the 7CFE crystal structure with Root Mean Square Deviation (RMSD) of 1.7Å over 1174 atoms using PyMol.[12] Considering the biological importance of all interacting ligands, Adenosine monophosphate (AMP), S-Adenosyl Methionine (SAM), 5nt RNA, and the drug STR on GidB, molecular docking was carried out through the following three stages:

1. AMP docking: Homology modelling identified PDB-ID 3G89[24] as one of the template structures with bound ligands AMP and SAM. 3G89 is a crystal structure of *T. thermophilus* RsmG in complex with AMP and Adomet (also known as SAM) at a 1.5Å resolution. The authors proposed that the AMP binding site could be a potential RNA-binding site. Therefore, AMP was docked first on the *M. tuberculosis* GidB structure created using Modeller following the docking protocol described above. As the biological unit of *M. tuberculosis* GidB is a monomer, AMP was docked on chain A of *M. tuberculosis* GidB. The AMP binding residues for the homologue structure were identified using the PLIP web server, followed by sequence and structure alignment using T-coffee Expresso[25] to identify the equivalent residues: R213 and G214 in *M. tuberculosis* GidB corresponding to R245 and H246 in the 3G89 crystal structure respectively. UCSF Chimera[8] was used to add hydrogens to AMP and to minimise the *M. tuberculosis* GidB structure in preparation for docking using AutoDock Vina. The binding poses were visualised and inspected in UCSF Chimera, and the top pose (referring to the orientation of the docked molecule) was chosen to form the Gid and AMP model, based on H-bond formation between residues R213 and W123 **(Figures 2** and **3)**. The model was then saved as a single complex.

A



B

```
receptor = gidb_m.pdbqt
ligand = amp_H_m.pdbqt

out = gidb_all.pdbqt

center_x = -4.00
center_y = 39.00
center_z = 16.028

size_x = 12
size_y = 14
size_z = 12

exhaustiveness = 8
num_modes = 15
```

Figure 2: Ligand torsion and configuration file for Adenosine monophosphate (AMP) docking
A) Ball and stick representation of AMP with rotatable bonds as identified by AutoDock Tools before docking in AutoDock Vina. The balls are coloured by atom type (grey: carbon, blue: nitrogen, red: oxygen), while the sticks are coloured according to rotatable bonds identified: Green sticks denote rotatable bonds, while red sticks denote un-rotatable bonds. Figure generated using AutoDock Tools version 4.2. **B)** Snapshot of the configuration file used by AutoDock Vina version 1.1.2 for docking AMP on *M. tuberculosis* GidB.

A B

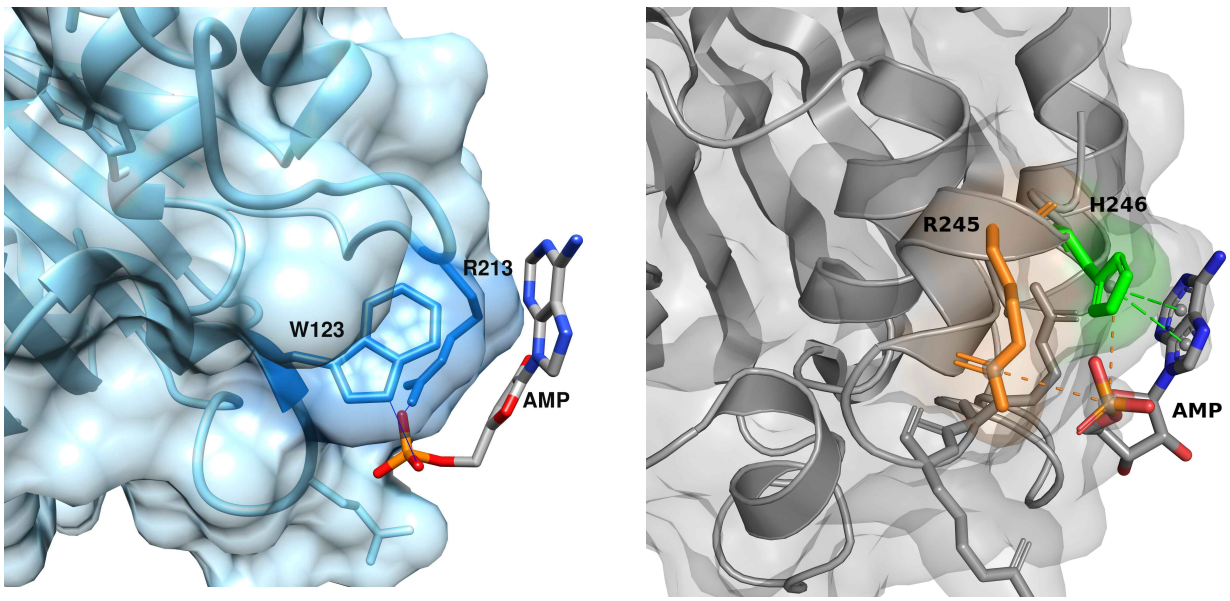
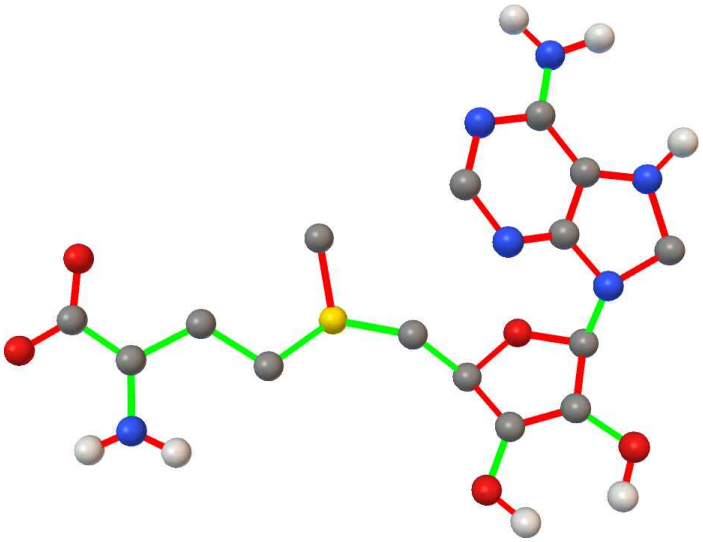


Figure 3: Comparison of Adenosine monophosphate (AMP) poses between *M. tuberculosis* and *T. thermophilus* GidB protein (PDB-ID: 3G89)
A) *M. tuberculosis* GidB with AMP docked, forming hydrogen bond with residues R213 and W123, **B)** Crystal structure of *T. thermophilus* GidB with AMP bound, forming salt bridge with residue R245 shown in orange, and pi-pi stacking with residue H246 appearing in green. The molecular interactions were generated using the PLIP web server, and figures rendered using UCSF Chimera version 1.14 and PyMol version 2.4.0.

2. SAM docking: Docking of SAM on the GidB-AMP complex was similarly guided using PLIP and the T-coffee Expresso alignment to identify equivalent binding/interacting residues for *M. tuberculosis* GidB to 3G89 (the equivalences are: G69:G88, S70:T89, G71:G90, L91:D111, E92:A112, P93:T113, R96:K116, R118:R138, A119:A139, E120:E140, R137:R158), adding hydrogen to the ligand and subsequently minimising the GidB-AMP complex to prepare for docking. The best pose was chosen based on an alignment of the SAM molecule of the homologue structure model along with the molecular interaction of SAM with residues in *M. tuberculosis* GidB. In this manner, the final GidB-AMP-SAM complex was obtained **(Figures 4** and **5)**.

A

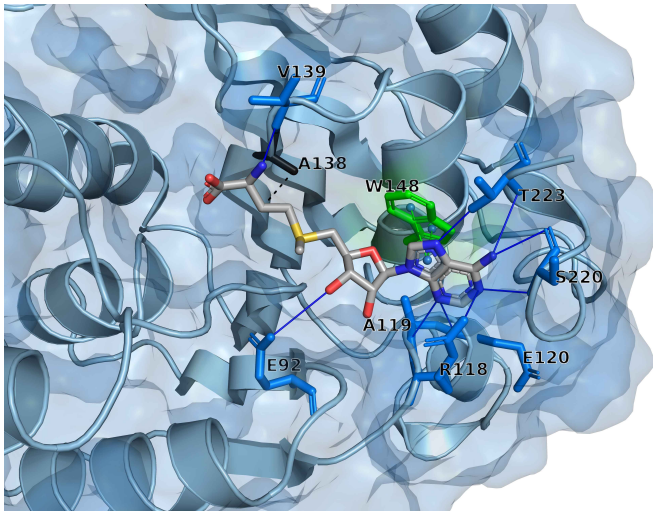


B

```
receptor = gidb_amp_1mdl_m.pdbqt
ligand = sam_H_m.pdbqt

out = gidb_sam_all.pdbqt

size_x = 24
size_y = 18
size_z = 16

center_x = 14
center_y = 41
center_z = 7

exhaustiveness = 15
num_modes = 15
```

Figure 4: Ligand torsion and configuration file for S-Adenosyl Methionine (SAM) docking
A) Ball-and-stick representation of SAM with rotatable bonds as identified by AutoDock Tools before docking in AutoDock Vina. The balls are coloured by atom type (grey: carbon, blue: nitrogen, red: oxygen, yellow: sulphur), while the sticks are coloured according to rotatable bonds identified: Green sticks represent rotatable bonds, while red sticks denote un-rotatable bonds. Figure generated using AutoDock Tools version 4.2. **B)** Snapshot of the configuration file used by AutoDock Vina version 1.1.2 for docking SAM on *M. tuberculosis* GidB.

A



B

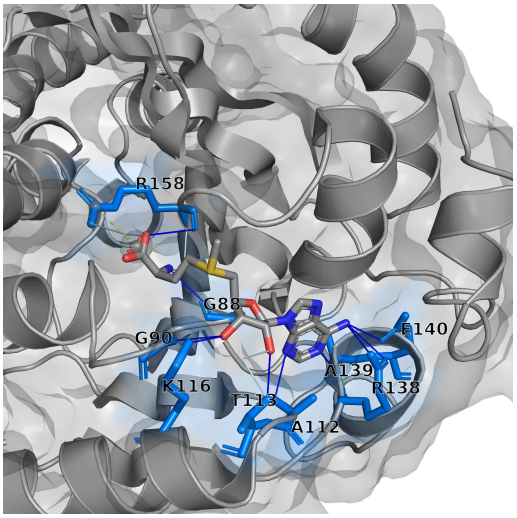


Figure 5: Comparison of S-Adenosyl Methionine (SAM) poses between *M. tuberculosis* and *T. thermophilus* GidB protein (PDB-ID: 3G89)
A) *M. tuberculosis* GidB with SAM docked, forming hydrogen bond with residues E92, R118, A119, E120, V139, S220, and T223 shown in blue. Green indicates pi-pi stacking with residue W148, and hydrophobic interaction with residue A138 indicated in black. **B)** Crystal structure 3G89 of *T. thermophilus* GidB with SAM bound, forming hydrogen bonds with G88, G90, A112, T113, K116, R138, A139, E140 and R158 indicated in blue. The orange dashed line indicates salt bridge interaction with residue R158. The molecular interactions were generated using the PLIP web server, and figures rendered using PyMol version 2.4.0.

3. RNA and STR docking: There is limited information available on the binding site and interactions of RNA and STR in *M. tuberculosis*, with no existing GidB structure in *M. tuberculosis*. However, there exists, a crystal structure of the *T. thermophilus* 30S ribosomal subunit with streptomycin bound (PDB-ID: 4DR3). Docking of RNA is a challenging task due to its size, charged nature and a lack of docking tools capable of docking such large (>10nt) fragments. Two iterations of RNA docking were carried out initially: a 45nt RNA fragment, and a 10nt RNA fragment (containing the methylation site residue G527). Docking was attempted using two different software tools: HDock and PatchDock. The results were inconclusive mainly due to the size of the fragment. Therefore a smaller fragment of 5nt (G526-G530) was extracted for docking. The RNA fragment was left rigid for the docking procedure. PatchDock returned the most promising results on inspection of the docked RNA fragment in relation to SAM, STR, and electrostatic interactions, where the latter was important to ensure that RNA was docked within a positively charged pocket **(Figure 6)**.

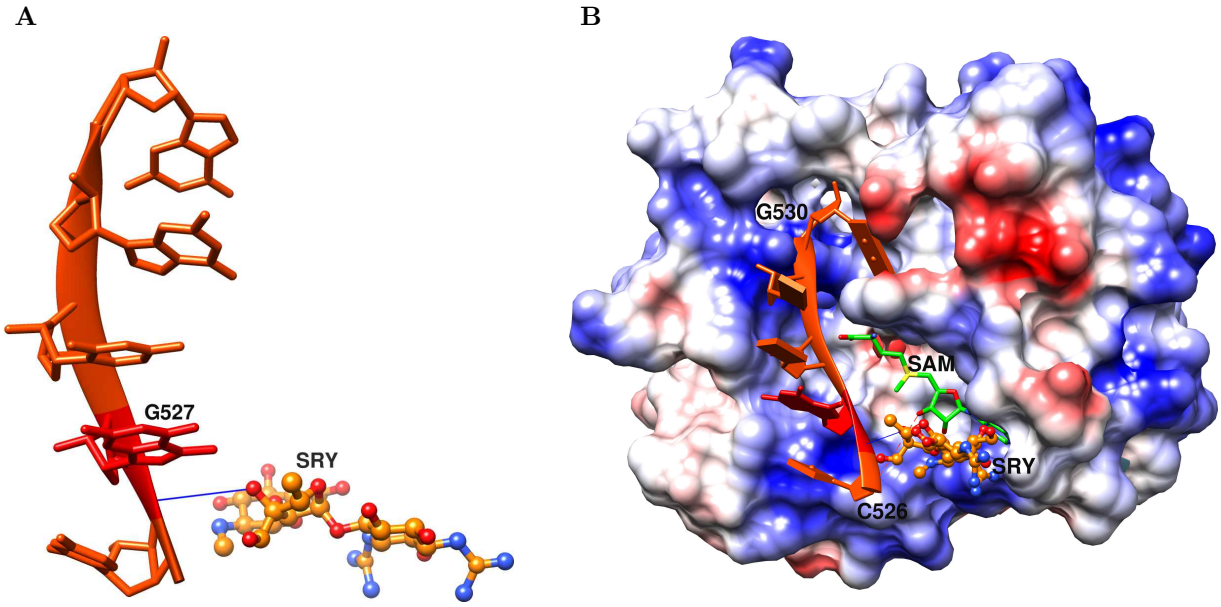
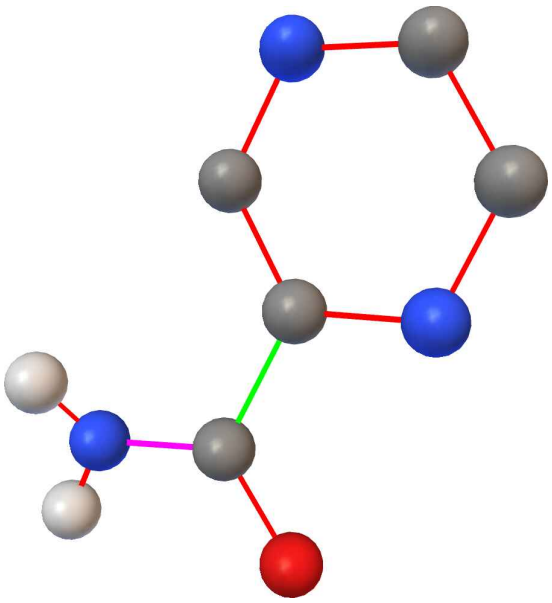


Figure 6: A 5 nucleotide RNA bound to streptomycin docked onto *M. tuberculosis* GidB complex **A)** 5 nucleotide (nt) RNA fragment consisting of residues G526-G530 was extracted from PDB-ID 4DR3, a crystal structure of 30S ribosomal subunit from *T. thermophilus* with streptomycin (STR) bound. The 5nt fragment in shown in red-orange, with G527 residue marked in red forming hydrogen bond with STR indicated in blue. **B)** The 5nt RNA-bound STR docked onto *M. tuberculosis* GidB protein with co-factor S-Adenosyl Methionine (SAM) depicted in green. The structure is coloured according to electrostatics, where blue denotes positively charged surfaces, and red denotes negatively charged surfaces. This highlights that the RNA fragment with STR bound was docked inside a positive pocket on *M. tuberculosis* GidB. The hydrogen bond between G527 and SRY is indicated in blue. RNA docking was performed using PatchDock. Figures were generated using UCSF Chimera version 1.14.

PncA-PZA complex: Generated using the AutoDock Vina.[6] Molecular docking was guided by the active site residues described for PncA (PDB-ID: 3PL1).[26] The general protocol for docking by AutoDock Vina was followed as described in the Molecular Docking section above. The rotatable bonds and configuration file used for PncA-PZA docking are shown in **Figure 7**.

A



B

```
receptor = 3pl1_receptor.pdbqt
ligand = PZA_ligand.pdbqt

out = all.pdbqt

center_x = 10.189
center_y = -28.082
center_z = -5.054

size_x = 16
size_y = 14
size_z = 14

exhaustiveness = 8
num_modes = 15
```

Figure 7: Ligand torsion and configuration file for pyrazinamide (PZA)
A) Ball and stick representation of PZA with rotatable bonds as identified by AutoDock Tools before docking in AutoDock Vina. The balls are coloured by atom type (grey: carbon, blue: nitrogen, red: oxygen), while the sticks are coloured according to rotatable bonds identified. Green sticks represent rotatable bonds, red sticks denote un-rotatable bonds, and magenta sticks indicate non-rotatable bonds. Figure generated using AutoDock Tools version 4.2. **B)** Snapshot of the configuration file used by AutoDock Vina version 1.1.2 for docking PZA on pyrazinamidase (PDB-ID: 3PL1).

A summary of the 3D structural data used in the project appears in **Table 2**.

2.3.3 Active site residue identification for docked complexes

Three software tools, LigPlus version 2.2.4 (downloaded),[27] the PLIP,[13] and Arpeggio[14] web servers were used to curate all possible interacting residues with the ligands and co-factors in the final docked complex. All such residues were considered "active site residues" for the purposes of investigating the molecular consequences of SAV mutations.

2.3.4 Mutational site classification

Mutational sites on the six structural genes were regarded as hotspots when the site presented with multiple (defined as 2 or more) SAVs, while sites with exactly 2 SAVs were considered 'budding' resistant hotspots.

2.4 *In silico* Predictors

Multiple computational tools were used to investigate the structural effects of SAVs on the gene/target and gene/target-drug complex. These structural biophysical effects were related to changes in

Gene-Drug	Structure PDB-ID	Biological unit	Docking	No of chains	Comment
alr-cycloserine (DCS)	1XFC	Homo-dimer	Yes	A,B mutations: Chain A	DCS docked on chain A*
embB-ethambutol (EMB)	7BVF	Hetero-3-mer	No	A,B,P mutations: Chain B	EMB bound in chain B
gidB-streptomycin (STR)	Chopin+Modeller (7CFE, now available)	Monomer	Yes	Chain A mutations: Chain A	AMP+SAM docking: AutoDock Vina RNA docking: PatchDock
katG-isoniazid (INH)	1SJ2	Homo-dimer	Yes	A,B mutations: Chain A	INH docked on chain A*
pncA-pyrazinamide (PZA)	3PL1	Monomer	Yes	Chain A mutations: Chain A	PZA docking: AutoDock Vina
rpoB-rifampicin (RFP)	5UHC	Hetero-6-mer	No	A,B mutations: Chain C	RFP bound in chain C

Table 2: 3D Structural Data
Available 3D structures from the Protein Data Bank and a summary of molecular modelling performed. * 3D structure provided by collaborators.

stability, estimated through a change in Gibbs Free energy ($\Delta\Delta$G Kcal/mol) between the wild-type and mutant-type residues, while changes in molecular binding affinities were estimated as log fold change for ligand affinity. Where applicable, changes in binding affinity for nucleic acid (NA) and protein-protein interactions (estimated as $\Delta\Delta$G Kcal/mol) were also included. Additionally, changes in physicochemical and evolutionary properties between wild-type and mutant-residues were also included.

A detailed description of computational tools, and their application in AMR is described in the review article published as part of this project.[28] The review article is included in full at the end of **Chapter 1** (Introduction) as it forms part of the thesis. A summary description of the computational tools used in this project (adapted from the published paper) appears below:

2.4.1 Sequence based tools

ConSurf,[29] SNAP2,[30] and PROVEAN[31] were used to incorporate evolutionary conservation changes when assessing mutational effects.

ConSurf: The score is based on a multiple sequence alignment which generates probabilistic evolutionary models and phylogenetic links. Through this score, more conserved sites (having slower evolutionary rates), that have important functional and structural consequences, can be identified.[29] Scores are graded 1 (variable) to 9 (conserved) for visualisation.

SNAP2: SNAP2 (Screening for Non-Acceptable Polymorphisms v.2:) characterises the effect of all possible missense mutations as either neutral or deleterious. It accounts for amino acid position probabilities using position-specific independent counts, based on the BLOSUM62 matrix. This predictor

considers other features such as protein fold (Pfam, PROSITE) and functional annotations (SWISS-PROT), and as such is the tool that spans the most comprehensive feature space.[30] Mutations are classified as either neutral or effect based on predicted scores.

PROVEAN: PROVEAN (Protein Variant Effect Analyser) uses the BLOSUM62 substitution matrix as an amino acid probability matrix and combines this with differences in sequence similarity between wild-type and mutant sequences. The sequence context in which variation occurs is also considered, and a numerical score is generated for each variant, which enables the functional classification into deleterious or neutral based on a predefined threshold.[31]

2.4.2 Structure based tools

Protomer stability

These tools measure the mutational impact as a change in stability ($\Delta\Delta$G in Kcal/mol) of the protein structure. Both the extent of the mutational impact (the value of $\Delta\Delta$G) as well as the direction of change ($\Delta\Delta$G classification as stabilising or destabilising) are returned for the predictions.

mCSM-DUET: Combines predictions from two complementary approaches i.e. Site Director Mutator (SDM)[32] and mCSM.[33] The latter refers to the mutation Cut-off Scanning Matrix (mCSM) method which uses graph-based methods to calculate atomic pairwise distance surrounding the wild-type amino acid. Mutational impact is captured based on a change in the atomic pharmacophore count resulting from SAV mutations. Together, this forms the mCSM signature, and is used to train predictive models for analysing mutational impact on structure stability, where $\Delta\Delta G<0$: Destabilising, and $\Delta\Delta G>0$: Stabilising.[34]

DeepDDG: DeepDDG calculates $\Delta\Delta$G of mutation using a neural network trained on nine categories of sequence and structural features. It operates independently as 'DeepDDG', and in an integrated manner as 'iDeepDDG' where predictions from three methods: mCSM, SDM and DUET can be fed into the concatenation layer of the neural network to generate a consensus prediction. Classification of mutational impact is $\Delta\Delta G<0$: Destabilising, and $\Delta\Delta G>0$: Stabilising.[35]

DynaMut2: DynaMut predicts stability effects based on protein dynamics resulting from vibrational entropy changes. It integrates mCSM signatures and normal mode analysis, and thus combines mutational effect from three structure-based prediction tools to generate a consensus prediction. Mutational impact is classified based on $\Delta\Delta G<0$: Destabilising, and $\Delta\Delta G>0$: Stabilising.[36]

FoldX: FoldX is an empirical-based predictor which provides information on how a SAV mutation alters the stability of a protein. Estimation of stability is based on intramolecular interactions such as van der Waals' forces, solvation energies, interactions with water, hydrogen bonds, electrostatic effects and main and side chain entropies. Mutational impact is calculated through a weighted summation of all the intramolecular interactions, and estimated as a change in stability ($\Delta\Delta$G) between mutant and wild-type structures. Mutational impact is classified as $\Delta\Delta G<0$: Stabilising, and $\Delta\Delta G>0$: Destabilising.[37] Of note, the classification score used by FoldX is inverted compared with other tools used in this project, where negative $\Delta\Delta$G denotes a stabilising effect in FoldX, the same is destabilising according to other tools. Similarly, where a positive $\Delta\Delta$G indicates a destabilising effect in FoldX, the same is stabilising according to other tools.

Average protomer stability: The predicted estimates from all four tools described above were averaged to obtain a consensus estimate of changes in protomer stability. The sign associated with FoldX estimates was reversed before calculating the average to account for the different classification criteria mentioned above.

Binding affinity

These tools are based on graph-based methods of the mCSM suite of tools described above, with properties of small molecules, nucleic acid, protein-protein interactions included to account for affinity changes upon mutation. Both the extent of the mutational impact (change in binding affinity) as well as the direction of change (classification as stabilising or destabilising) are returned for the predictions.

mCSM-lig and mmCSM-lig: These tools use the mCSM graph-based structural signature to estimate the ligand affinity change upon mutation. Along with changes in protein stability, small-molecule chemical features and ligand physicochemical properties are considered to capture mutational changes. Mutational impact is given by the log (ln) affinity fold change between wild type and mutant complexes, where ln(fold-change)<0: Destabilising, ln(fold-change)>0: Stabilising.[38] Furthermore, mmCSM-lig (personal communication, unpublished), which is built upon mCSM-lig to include multiple SAVs was also run for all gene-targets.

Average ligand affinity: The predicted estimates from both mCSM- and mmCSM-lig were averaged to obtain a consensus estimate of changes in ligand/drug binding affinity.

mCSM-NA: This estimates a change in the binding affinity of nucleic-acid (NA) upon mutation. The mCSM-graph based structure signature is then extended to include atomic pharmacophore changes for nucleotides (sub-classed into phosphate, sugar and base atom groups) along with residue distance to NA, and the effect of reverse mutation. The predicted affinity change is given by $\Delta\Delta G$ in Kcal/mol, along with the classification corresponding to $\Delta\Delta G<0$: Destabilising, and $\Delta\Delta G>0$: Stabilising.[39] In this project, the NA affinity changes refers to nucleic acid (RNA/DNA) binding affinity changes.

mCSM-PPI2: Mutational effect on protein-protein interaction (PPI) interface is estimated using a combination of structural, evolutionary, PPI network metrics, and energetic terms. Similarly based on the mCSM graph-based structure signature, inter-residue interaction network properties and the effect of the reverse mutation are included. The predicted complex-affinity change is given by $\Delta\Delta G$ in Kcal/mol, along with the classification corresponding to $\Delta\Delta G<0$: Destabilising, and $\Delta\Delta G>0$: Stabilising.[40]

For all affinity estimates, in line with the respective computational tool's threshold criteria, only mutations within 10Å of the ligand, nucleic acid, and protein-protein interface were considered.

2.4.3 Residue level properties

Residue-level properties for the wild-type structure were analysed for all SAVs. Accessible (ASA) and Relative Surface Area (RSA), residue depth (RD), and hydrophobicity values according to the Kyte-Doolittle (KD) scale were obtained. The DSSP program[41,42] was used to extract the ASA and RSA values, while RD values were calculated using the depth server according to the seminal paper.[43] The hydrophobicity KD values were fetched from the expasy server.[44] The computational tools used, and their current availability are listed in Appendix Table 2.A.1.

2.4.4 Prominent mutational effects

When investigating the underlying predominant mutational impact at a given site, effects were prioritised in order of interacting partner size, progressing from mCSM/mmCSM-lig, mCSM-NA, mCSM-PPI2, followed by protomer stability changes. This approach allows molecular interactions to be adequately accounted for, irrespective of interacting partner size and helps identify the most prominent effect at a given site.

2.5 Data Analysis

2.5.0.1 *In silico* framework

A semi-automated framework using a combination of shell, Python and R programming languages was developed for this project. The framework consists of a set of pipelines: data extraction of SAVs, obtaining results from computational predictors, and integrating these for statistical and machine learning analyses. The overall methodology workflow is depicted in **Figure 8**.

2.5.0.2 Minor Allele Frequency, Odds Ratio and Lineage calculations

Across the *M. tuberculosis* isolates tested for drug susceptibility for each gene-drug target, association analysis to estimate the risk of resistance for SAV was performed. For each SAV in each gene, minor allele frequency (MAF) and odds ratio (OR) were calculated in relation to all samples tested for their respective DST. Each drug had a corresponding DST column for each clinical isolate, provided as part of the dataset. The DST values were classified as 0: Sensitive, 1: Resistant. The MAF was calculated based on this as the average occurrence of a given SAV, and OR as the measure of association of a given SAV with its corresponding drug resistance. For each SAV in a given gene, its frequency with respect to the corresponding DST based on the entire dataset was extracted in the form of a contingency table. Since a given mutation can occur in more than one sample (as expected), and displays different DST sensitivity due to belonging to different isolates, the contingency table classifies the mutational frequency with respect to the binary DST sensitivity as rows and columns. Fisher's exact test was then used to calculate the OR and P-values based upon this table. Lineage information was summarised according to the distinct number of lineages, as well as the total number of different lineages per mutation, as well their respective proportional contribution.

2.5.0.3 Normalisation

Results from all computational predictors were normalised between -1 and 1 for comparison and visualisation. For all binding affinity analyses, data was filtered according to distance from interacting site (ligand, nucleic acid, and protein-protein interface) where only residues within 10Å were considered.

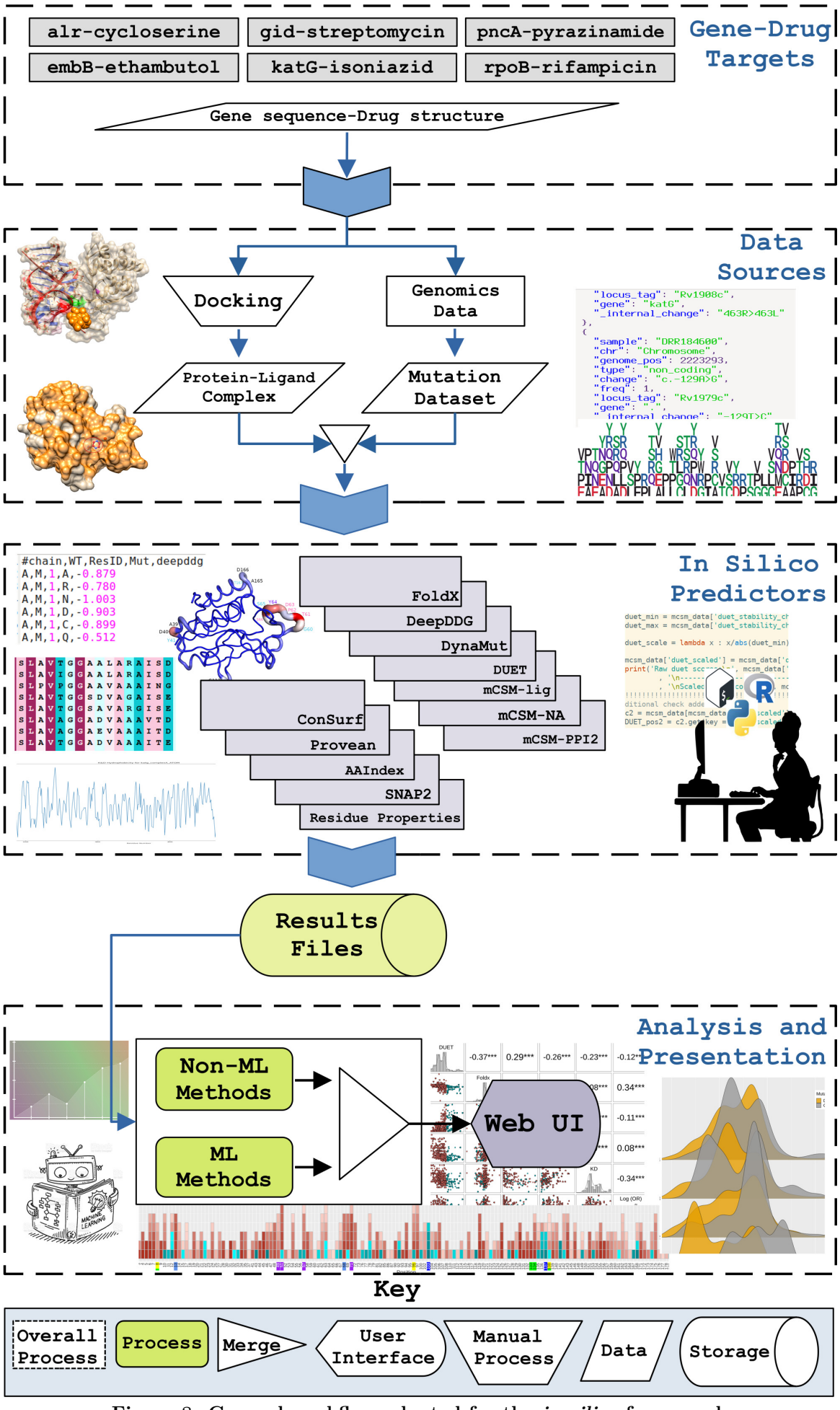


Figure 8: General workflow adopted for the *in silico* framework

2.5.0.4 Statistical analysis

Data was analysed using non-parametric statistical tests. Features distinguishing 'Sensitive' and 'Resistant' mutations were assessed using the unpaired Wilcoxon test. For assessing mutation proportions across lineages, Fisher's exact test was used. Correlations were assessed using the Spearman's rank coefficient (rho, ρ). Correlation thresholds used to assess associations were: $\rho<0.1$: no association, $0.1\leq\rho<0.3$: weak association, $0.3\leq\rho<0.6$: moderate association, $\rho\geq0.6$: strong association. The Kolmogorov-Smirnov (KS) test was used to compare distributions. Statistical significance thresholds used were: .P<0.10, $*$P<0.05, $**$P<0.01, $***$P<0.001, $****$P<0.0001). All statistical analyses were carried out using the R statistical software version 4.0.4.[45]

2.5.0.5 Visualisation

All plots were generated using R statistical software version 4.0.4.[45] Protein and ligand structures were generated using UCSF Chimera version 1.14,[8] PyMol version 2.4.0,[12] and AutoDock Tool version 4.2.[9] An interactive dashboard was also built during this project using Rshiny version 4.0.4[46] as an effective tool to inspect and visualise the interrelationship between structural and genomic data. The dashboard is available at https://thesis.tunstall.in.

2.6 Web-based visualisation tool development

As an offshoot of the plot generation using R, a web-based visualisation tool was developed using Rshiny. The dashboard is hosted on a commercial public cloud server. Initial development on my own desktop made this quite simple due to the fact that I had taken a functional approach when writing my R code. When deployed on the public cloud service, however, performance issues presented themselves. Through a combination of approaches including code refactoring, web server caching, use of the Feather file format rather than flat files such as CSV or JSON, load times were reduced to the point where the tools are usable. While special-purpose hosting solutions for Rshiny apps exist (e.g. https://shinyapps.io), there are sufficient differences that made it impractical to adapt what had been written already. While tools such as Rshiny make initial development very easy, knowledge of modern web technologies is essential to use them effectively. Version control systems such as Git proved invaluable in the overall development process.

References

[1] Gary Napier et al. "Robust Barcoding and Identification of Mycobacterium Tuberculosis Lineages for Epidemiological and Clinical Studies". In: *Genome Medicine* 12.1 (Dec. 1, 2020), pp. 1–10. ISSN: 1756-994X. DOI: 10.1186/s13073-020-00817-3.

[2] Jody E. Phelan et al. "Integrating Informatics Tools and Portable Sequencing Technology for Rapid Detection of Resistance to Anti-Tuberculous Drugs". In: *Genome Medicine* 11.1 (June 24, 2019), p. 41. ISSN: 1756-994X. DOI: 10.1186/s13073-019-0650-x.

[3] Susumu Okamoto et al. "Loss of a Conserved 7-Methylguanosine Modification in 16S rRNA Confers Low-Level Streptomycin Resistance in Bacteria". In: *Molecular Microbiology* 63.4 (Feb. 2007), pp. 1096–1106. ISSN: 0950-382X. DOI: 10.1111/j.1365-2958.2006.05585.x.

[4] Adamandia Kapopoulou, Jocelyne M. Lew, and Stewart T. Cole. "The MycoBrowser Portal: A Comprehensive and Manually Annotated Resource for Mycobacterial Genomes". In: *Tuberculosis (Edinburgh, Scotland)* 91.1 (Jan. 2011), pp. 8–13. ISSN: 1873-281X. DOI: 10.1016/j.tube.2010.09.006.

[5] Helen M. Berman et al. "The Protein Data Bank". In: *Acta Crystallographica Section D: Biological Crystallography* 58 (6 I 2002), pp. 899–907. ISSN: 09074449. DOI: 10.1107/S0907444902003451.

[6] Oleg Trott and Arthur J. Olson. "AutoDock Vina: Improving the Speed and Accuracy of Docking with a New Scoring Function, Efficient Optimization, and Multithreading". In: *Journal of Computational Chemistry* 31.2 (Jan. 2009), NA–NA. ISSN: 01928651. DOI: 10.1002/jcc.21334.

[7] Garrett M Morris et al. "Automated Docking Using a Lamarckian Genetic Algorithm and an Empirical Binding Free Energy Function". In: *Journal of Computational Chemistry* 19.14 (1998). ISSN: 0192-8651.

[8] Eric F. Pettersen et al. "UCSF Chimera?A Visualization System for Exploratory Research and Analysis". In: *Journal of Computational Chemistry* 25.13 (Oct. 2004), pp. 1605–1612. ISSN: 0192-8651. DOI: 10.1002/jcc.20084.

[9] Garrett M. Morris et al. "AutoDock4 and AutoDockTools4: Automated Docking with Selective Receptor Flexibility". In: *Journal of Computational Chemistry* 30.16 (Dec. 2009), pp. 2785–2791. ISSN: 01928651. DOI: 10.1002/jcc.21256.

[10] Dina Duhovny, Ruth Nussinov, and Haim J. Wolfson. "Efficient Unbound Docking of Rigid Molecules". In: *Algorithms in Bioinformatics*. Ed. by Roderic Guigó and Dan Gusfield. Lecture Notes in Computer Science. Berlin, Heidelberg: Springer, 2002, pp. 185–200. ISBN: 978-3-540-45784-8. DOI: 10.1007/3-540-45784-4_14.

[11] Dina Schneidman-Duhovny et al. "PatchDock and SymmDock: Servers for Rigid and Symmetric Docking". In: *Nucleic Acids Research* 33 (Web Server issue July 1, 2005), W363–367. ISSN: 1362-4962. DOI: 10.1093/nar/gki481.

[12] LLC Schrödinger and Warren DeLano. *PyMOL*. Version 2.4.0. May 20, 2020.

[13] Melissa F. Adasme et al. "PLIP 2021: Expanding the Scope of the Protein-Ligand Interaction Profiler to DNA and RNA". In: *Nucleic Acids Research* 49.W1 (July 2, 2021), W530–W534. ISSN: 1362-4962. DOI: 10.1093/nar/gkab294.

[14] Harry C Jubb et al. "Arpeggio: A Web Server for Calculating and Visualising Interatomic Interactions in Protein Structures." In: *Journal of molecular biology* 429.3 (2017), pp. 365–371. ISSN: 1089-8638. DOI: 10.1016/j.jmb.2016.12.004.

[15] Lu Zhang et al. "Structures of Cell Wall Arabinosyltransferases with the Anti-Tuberculosis Drug Ethambutol". In: *Science (New York, N.Y.)* 368.6496 (June 12, 2020), pp. 1211–1219. ISSN: 1095-9203. DOI: 10.1126/science.aba9102.

[16] Wei Lin et al. "Structural Basis of Mycobacterium Tuberculosis Transcription and Transcription Inhibition". In: *Molecular Cell* 66.2 (Apr. 20, 2017), 169–179.e8. ISSN: 1097-2765. DOI: 10.1016/j.molcel.2017.03.001.

[17] Pierre LeMagueres et al. "The 1.9 A Crystal Structure of Alanine Racemase from Mycobacterium Tuberculosis Contains a Conserved Entryway into the Active Site". In: *Biochemistry* 44.5 (Feb. 8, 2005), pp. 1471–1481. ISSN: 0006-2960. DOI: 10.1021/bi0486583.

[18] Thomas Bertrand et al. “Crystal Structure of Mycobacterium Tuberculosis Catalase-Peroxidase”. In: *The Journal of Biological Chemistry* 279.37 (Sept. 10, 2004), pp. 38991–38999. ISSN: 0021-9258. DOI: 10.1074/jbc.M402382200.

[19] Stephanie Portelli et al. “Understanding Molecular Consequences of Putative Drug Resistant Mutations in Mycobacterium Tuberculosis”. In: *Scientific Reports* (2018). ISSN: 20452322. DOI: 10.1038/s41598-018-33370-6.

[20] O. Asojo et al. “Structural and Biochemical Analyses of Alanine Racemase from the Multidrug-Resistant Clostridium Difficile Strain 630”. In: *Acta crystallographica. Section D, Biological crystallography* (2014). DOI: 10.1107/S1399004714009419.

[21] Pietro Vidossich et al. “Binding of the Antitubercular Pro-Drug Isoniazid in the Heme Access Channel of Catalase-Peroxidase (KatG). A Combined Structural and Metadynamics Investigation”. In: *The Journal of Physical Chemistry. B* 118.11 (Mar. 20, 2014), pp. 2924–2931. ISSN: 1520-5207. DOI: 10.1021/jp4123425.

[22] Bernardo Ochoa-Montaño, Nishita Mohan, and Tom L. Blundell. “CHOPIN: A Web Resource for the Structural and Functional Proteome of Mycobacterium Tuberculosis”. In: *Database* 2015 (Jan. 1, 2015). DOI: 10.1093/database/bav026.

[23] Andrej ali and Tom L. Blundell. “Comparative Protein Modelling by Satisfaction of Spatial Restraints”. In: *Journal of Molecular Biology* 234.3 (Dec. 5, 1993), pp. 779–815. ISSN: 0022-2836. DOI: 10.1006/jmbi.1993.1626.

[24] Steven T. Gregory et al. “Structural and Functional Studies of the Thermus Thermophilus 16S rRNA Methyltransferase RsmG”. In: *RNA* 15.9 (Jan. 9, 2009), pp. 1693–1704. ISSN: 1355-8382, 1469-9001. DOI: 10.1261/rna.1652709.

[25] Paolo Di Tommaso et al. “T-Coffee: A Web Server for the Multiple Sequence Alignment of Protein and RNA Sequences Using Structural Information and Homology Extension”. In: *Nucleic Acids Research* 39 (suppl_2 July 1, 2011), W13–W17. ISSN: 0305-1048. DOI: 10.1093/nar/gkr245.

[26] Stéphanie Petrella et al. “3PL1: Crystal Structure of the Pyrazinamidase of Mycobacterium Tuberculosis: Insights into Natural and Acquired Resistance to Pyrazinamide”. In: *PLoS ONE* 6.1 (Jan. 2011), e15785. ISSN: 19326203. DOI: 10.1371/journal.pone.0015785.

[27] Roman A. Laskowski and Mark B. Swindells. “LigPlot+: Multiple Ligand-Protein Interaction Diagrams for Drug Discovery”. In: *Journal of Chemical Information and Modeling* 51.10 (Oct. 24, 2011), pp. 2778–2786. ISSN: 1549-960X. DOI: 10.1021/ci200227u.

[28] Tanushree Tunstall et al. “Combining Structure and Genomics to Understand Antimicrobial Resistance”. In: *Computational and Structural Biotechnology Journal* 18 (2020), pp. 3377–3394. ISSN: 2001-0370. DOI: 10.1016/j.csbj.2020.10.017.

[29] Haim Ashkenazy et al. “ConSurf 2010: Calculating Evolutionary Conservation in Sequence and Structure of Proteins and Nucleic Acids”. In: *Nucleic Acids Research* 38 (suppl_2 July 1, 2010), W529–W533. ISSN: 0305-1048. DOI: 10.1093/nar/gkq399.

[30] Maximilian Hecht, Yana Bromberg, and Burkhard Rost. “Better Prediction of Functional Effects for Sequence Variants”. In: *BMC genomics* 16 Suppl 8 (2015), S1. ISSN: 1471-2164. DOI: 10.1186/1471-2164-16-S8-S1.

[31] Yongwook Choi and Agnes P. Chan. “PROVEAN Web Server: A Tool to Predict the Functional Effect of Amino Acid Substitutions and Indels”. In: *Bioinformatics* 31.16 (Aug. 15, 2015), pp. 2745–2747. ISSN: 1367-4803. DOI: 10.1093/bioinformatics/btv195.

[32] Catherine L Worth, Robert Preissner, and Tom L Blundell. “SDM-a Server for Predicting Effects of Mutations on Protein Stability and Malfunction”. In: *Nucleic Acids Research* 39 (Suppl.2 2011), W215–W222. DOI: 10.1093/nar/gkr363.

[33] Douglas Pires, David B. Ascher, and Tom L. Blundell. “mCSM: Predicting the Effects of Mutations in Proteins Using Graph-Based Signatures”. In: *Bioinformatics* 30.3 (2014), pp. 335–342. ISSN: 13674803. DOI: 10.1093/bioinformatics/btt691.

[34] Douglas Pires, David B. Ascher, and Tom L. Blundell. “DUET: A Server for Predicting Effects of Mutations on Protein Stability Using an Integrated Computational Approach”. In: *Nucleic Acids Research* 42.W1 (2014), pp. 314–319. ISSN: 13624962. DOI: 10.1093/nar/gku411.

[35] Huali Cao et al. “DeepDDG: Predicting the Stability Change of Protein Point Mutations Using Neural Networks”. In: *Journal of Chemical Information and Modeling* 59.4 (Apr. 22, 2019), pp. 1508–1514. ISSN: 1549-960X. DOI: 10.1021/acs.jcim.8b00697.

[36] Carlos H. M. Rodrigues, Douglas E. V. Pires, and David B. Ascher. “DynaMut2: Assessing Changes in Stability and Flexibility upon Single and Multiple Point Missense Mutations”. In: *Protein Science: A Publication of the Protein Society* 30.1 (Jan. 2021), pp. 60–69. ISSN: 1469-896X. DOI: 10.1002/pro.3942.

[37] Joost Schymkowitz et al. “The FoldX Web Server: An Online Force Field”. In: *Nucleic Acids Research* 33 (suppl_2 July 1, 2005), W382–W388. ISSN: 0305-1048. DOI: 10.1093/nar/gki387.

[38] Douglas Pires, Tom L. Blundell, and David B. Ascher. “mCSM-lig: Quantifying the Effects of Mutations on Protein-Small Molecule Affinity in Genetic Disease and Emergence of Drug Resistance”. In: *Scientific Reports* 6.1 (Sept. 2016), p. 29575. ISSN: 2045-2322. DOI: 10.1038/srep29575.

[39] Douglas E V Pires and David B Ascher. “mCSM-NA: Predicting the Effects of Mutations on Protein-Nucleic Acids Interactions”. In: *Nucleic acids research* 45.W1 (July 2017), W241–W246. ISSN: 0305-1048. DOI: 10.1093/nar/gkx236.

[40] Carlos H M Rodrigues et al. “mCSM-PPI2: Predicting the Effects of Mutations on ProteinProtein Interactions”. In: *Nucleic Acids Research* 47.W1 (May 2019), W338–W344. ISSN: 0305-1048. DOI: 10.1093/nar/gkz383.

[41] Wolfgang Kabsch and Christian Sander. “Dictionary of Protein Secondary Structure: Pattern Recognition of Hydrogen-Bonded and Geometrical Features”. In: *Biopolymers* 22.12 (1983), pp. 2577–2637. ISSN: 1097-0282. DOI: 10.1002/bip.360221211.

[42] Wouter G. Touw et al. “A Series of PDB-related Databanks for Everyday Needs”. In: *Nucleic Acids Research* 43 (Database issue Jan. 2015), pp. D364–368. ISSN: 1362-4962. DOI: 10.1093/nar/gku1028.

[43] S. Chakravarty and R. Varadarajan. “Residue Depth: A Novel Parameter for the Analysis of Protein Structure and Stability”. In: *Structure (London, England: 1993)* 7.7 (July 15, 1999), pp. 723–732. ISSN: 0969-2126. DOI: 10.1016/s0969-2126(99)80097-5.

[44] Panu Artimo et al. “ExPASy: SIB Bioinformatics Resource Portal”. In: *Nucleic Acids Research* 40 (Web Server issue July 2012), W597–603. ISSN: 1362-4962. DOI: 10.1093/nar/gks400.

[45] R Core Team. *R: A Language and Environment for Statistical Computing*. Vienna, Austria, 2014.

[46] Winston Chang et al. *Shiny: Web Application Framework for R*. 2021.

Appendix for Chapter 2

2.A Computational tools and web URLs

Name of tool	URL
ConSurf	https://consurf.tau.ac.il/
PROVEAN	http://provean.jcvi.org/seq_submit.php
SNAP2	https://rostlab.org/services/snap/
AAindex	https://www.genome.jp/aaindex/
FoldX	https://foldxsuite.crg.eu/products#foldx
DeepDDG	http://protein.org.cn/ddg.html
Dynamut2	http://biosig.unimelb.edu.au/dynamut2/
mCSM-lig *(Also returns DUET scores)*	http://biosig.unimelb.edu.au/mcsm_lig/
mCSM-NA	http://biosig.unimelb.edu.au/mcsm_na/
mCSM-PPI2	http://biosig.unimelb.edu.au/mcsm_ppi2/
DSSP	https://swift.cmbi.umcn.nl/gv/dssp/
Hydrophobicity	https://web.expasy.org/protscale/
Residue Depth	http://cospi.iiserpune.ac.in/depth/
Arpeggio	http://structure.bioc.cam.ac.uk/arpeggio/
PLIP	https://plip-tool.biotec.tu-dresden.de/plip-web/plip/index
LigPlot	https://www.ebi.ac.uk/thornton-srv/software/LigPlus/
PatchDock	https://bioinfo3d.cs.tau.ac.il/PatchDock/
AutoDock Vina	https://vina.scripps.edu/
AutoDock Tools	https://autodocksuite.scripps.edu/adt/

Table 2.A.1: List of computational tools used and their online availability as of 18 Jul 2022.

Chapter 3

PncA-pyrazinamide results

London School of Hygiene & Tropical Medicine
Keppel Street, London WC1E 7HT

T: +44 (0)20 7299 4646
F: +44 (0)20 7299 4656
www.lshtm.ac.uk

RESEARCH PAPER COVER SHEET

Please note that a cover sheet must be completed for each research paper included within a thesis.

SECTION A – Student Details

Student ID Number	LSH1806129	**Title**	Mrs
First Name(s)	Tanushree		
Surname/Family Name	Tunstall		
Thesis Title	Using Machine Learning to Anticipate Antimicrobial Resistance in Mycobacterium Tuberculosis		
Primary Supervisor	Nicholas Furnham		

If the Research Paper has previously been published please complete Section B, if not please move to Section C.

SECTION B – Paper already published

Where was the work published?	Frontiers in Molecular Biosciences		
When was the work published?	July 2021		
If the work was published prior to registration for your research degree, give a brief rationale for its inclusion			
Have you retained the copyright for the work?*	Yes	Was the work subject to academic peer review?	Yes

*If yes, please attach evidence of retention. If no, or if the work is being included in its published format, please attach evidence of permission from the copyright holder (publisher or other author) to include this work.

SECTION C – Prepared for publication, but not yet published

Where is the work intended to be published?	
Please list the paper's authors in the intended authorship order:	
Stage of publication	Choose an item.

SECTION D – Multi-authored work

For multi-authored work, give full details of your role in the research included in the paper and in the preparation of the paper. (Attach a further sheet if necessary)	I wrote custom Python scripts to extract, format and analyse the GWAS dataset received from Prof Taane Clark's lab (my secondary supervisor). I also wrote custom R scripts for generating all the plots, as well as conducting all the statistical analysis within the manuscript. The structure figures were generated using Chimera using custom scripts. I wrote the initial draft of the manuscript, and circulated and coordinated feedback from all co-authors. Subsequently, I responded to the reviewer comments by incorporating feedback from all co-authors, and handled any follow-on queries from the journal including the final proof.

SECTION E

Student Signature	Tanushree Tunstall
Date	19/07/2022

Supervisor Signature	Nicholas Furnham
Date	19/07/2022

frontiers in Molecular Biosciences

ORIGINAL RESEARCH
published: 23 July 2021
doi: 10.3389/fmolb.2021.619403

Structural and Genomic Insights Into Pyrazinamide Resistance in *Mycobacterium tuberculosis* Underlie Differences Between Ancient and Modern Lineages

*Tanushree Tunstall[1], Jody Phelan[1], Charlotte Eccleston[1], Taane G. Clark[1,2] and Nicholas Furnham[1]**

[1] Department of Infection Biology, London School of Hygiene and Tropical Medicine, London, United Kingdom, [2] Department of Infectious Disease Epidemiology, London School of Hygiene and Tropical Medicine, London, United Kingdom

Resistance to drugs used to treat tuberculosis disease (TB) continues to remain a public health burden, with missense point mutations in the underlying *Mycobacterium tuberculosis* bacteria described for nearly all anti-TB drugs. The post-genomics era along with advances in computational and structural biology provide opportunities to understand the interrelationships between the genetic basis and the structural consequences of *M. tuberculosis* mutations linked to drug resistance. Pyrazinamide (PZA) is a crucial first line antibiotic currently used in TB treatment regimens. The mutational promiscuity exhibited by the *pncA gene* (target for PZA) necessitates computational approaches to investigate the genetic and structural basis for PZA resistance development. We analysed 424 missense point mutations linked to PZA resistance derived from ~35K *M. tuberculosis* clinical isolates sourced globally, which comprised the four main *M. tuberculosis* lineages (Lineage 1–4). Mutations were annotated to reflect their association with PZA resistance. Genomic measures (minor allele frequency and odds ratio), structural features (surface area, residue depth and hydrophobicity) and biophysical effects (change in stability and ligand affinity) of point mutations on pncA protein stability and ligand affinity were assessed. Missense point mutations within *pncA* were distributed throughout the gene, with the majority (>80%) of mutations with a destabilising effect on protomer stability and on ligand affinity. Active site residues involved in PZA binding were associated with multiple point mutations highlighting mutational diversity due to selection pressures at these functionally important sites. There were weak associations between genomic measures and biophysical effect of mutations. However, mutations associated with PZA resistance showed statistically significant differences between structural features (surface area and residue depth), but not hydrophobicity score for mutational sites. Most interestingly *M. tuberculosis* lineage 1 (ancient lineage) exhibited a distinct protein stability profile for mutations associated with PZA resistance, compared to modern lineages.

Keywords: *Mycobacterium tuberculosis*, pncA, nsSNPs, non-synonymous Single Nucleotide Polymorphisms, biophysical effects, thermodynamic stability, mCSM, FoldX

OPEN ACCESS

Edited by:
Arun Prasad Pandurangan, MRC Laboratory of Molecular Biology (LMB), United Kingdom

Reviewed by:
Wim Vranken, Vrije University Brussel, Belgium
Raghavan Varadarajan, Indian Institute of Science (IISc), India

****Correspondence:***
Nicholas Furnham
Nick.Furnham@lshtm.ac.uk

Specialty section:
This article was submitted to Biological Modeling and Simulation, a section of the journal Frontiers in Molecular Biosciences

Received: *20 October 2020*
Accepted: *14 April 2021*
Published: *23 July 2021*

Citation:
Tunstall T, Phelan J, Eccleston C, Clark TG and Furnham N (2021) Structural and Genomic Insights Into Pyrazinamide Resistance in Mycobacterium tuberculosis Underlie Differences Between Ancient and Modern Lineages. Front. Mol. Biosci. 8:619403. doi: 10.3389/fmolb.2021.619403

INTRODUCTION

Tuberculosis (TB), is a highly infectious and contagious air-borne disease caused by the bacterium *Mycobacterium tuberculosis*. Despite its ancient origins and the efforts to develop disease control and prevention measures, the disease continues to cause a global public health burden, with increased drug resistance making control difficult. In 2019, WHO reported around 10 million global cases of TB of which 1.4 million result in death (World Health Organization [WHO], 2020). In 2019, 465,000 cases of rifampicin resistant TB (RR-TB), among which 78% cases of multidrug-resistant TB (MDR-TB, defined as having additional resistance to isoniazid) were reported. Among these RR/MDR cases, ∼6% cases were further resistant to one fluoroquinolone and one injectable second line drug, leading to extensively drug resistant TB (XDR-TB) (World Health Organization [WHO], 2020).

The size of the *M. tuberculosis* genome (reference H37Rv strain) is 4.4 Mb, with a high (65%) GC content. The *M. tuberculosis* genome is clonal, and consists of seven main lineages, which vary by their geographical spread (L1: Indo-Oceanic, L2: East Asian, L3: East-Africa-Indian, and L4: Euro-American) (Phelan et al., 2016). The lineages are further classified into ancient (L1, L5–6), modern (L2–4), and intermediate (L7) strains, with L2 being particularly mobile as evidenced by its recent spread to Europe and Africa from Asia (Phelan et al., 2016). The *M. tuberculosis* lineages appear as distinct clades on phylogenetic trees (Coll et al., 2014) and govern disease transmission and dynamics with phenotypic consequences on clinical severity and drug resistance (Ford et al., 2013; Reiling et al., 2013), including recent reports of lineage-specific associations with the latter (Oppong et al., 2019). Drug resistance in *M. tuberculosis* is almost exclusively due to mutations [including non-synonymous Single Nucleotide Polymorphisms (nsSNPs), insertions and deletions (INDELs)] in genes coding for drug-targets or drug-converting enzymes. Changes in efflux pump regulation may also have an impact on the emergence of resistance (Al-Saeedi and Al-Hajoj, 2017) and putative compensatory mechanisms have been described to overcome fitness impairment that arises during the accumulation of resistance conferring mutations (de Vos et al., 2013). Resistance-associated point mutations have been described for all first-line drugs, including rifampicin, isoniazid and pyrazinamide, as well as for several second-line and newer drugs (fluoroquinolones, bedaquiline) (Somoskovi et al., 2001; Boonaiam et al., 2010; Segala et al., 2012), but knowledge is still incomplete.

Pyrazinamide (PZA) is a crucial antibiotic used in WHO recommended combination therapies in the front-line treatment of TB. It is a pro-drug which is activated by the amidase activity of the enzyme pyrazinamidase/nicotinamidase (PZase; MtPncA) encoded by the *pncA* gene, converting PZA to its active form of pyrazinoic acid (POA). Despite its indispensable status in TB treatment, PZA's exact mode of action remains poorly understood. Other genes (*rpsA* and *panD)* have been implicated in PZA resistance (Dookie et al., 2018) with a recent study suggesting that PZA exerts its antibacterial activity by acting as a target degrader of panD, blocking the synthesis of coenzyme A (targeted by POA) (Gopal et al., 2020). Despite this, mutations in the *pncA* gene remain the most common mechanism of PZA resistance (Khan et al., 2019).

Advances in whole genome sequencing (WGS) is assisting the profiling of *M. tuberculosis* for drug resistance, lineage determination and virulence, and presence in a transmission cluster (Phelan et al., 2019a), thereby informing clinical management and control policies. This is reflected in the WHO recommendation for use of rapid molecular testing for detecting TB and drug resistant TB (World Health Organization [WHO], 2020). The use of WGS can uncover new resistance mutations through genome-wide association studies (GWAS) and convergent evolution analysis (Phelan et al., 2016; Coll et al., 2018).

Furthermore, using protein structure, the biophysical effects of point polymorphisms can be investigated allowing a mechanistic understanding of resistance development (Phelan et al., 2016; Kavvas et al., 2018; Portelli et al., 2018). This approach can highlight important functional resistance mutations before they take hold in a population, corroborate drug susceptibility test results, as well as provide insights in highly polymorphic candidate loci (e.g., *pncA*) where many of the putative mutations have low frequency. It has been observed that sites with multiple mutations (>2) are linked to drug resistance (Comas et al., 2011), but such resistance hotspots may not necessarily lie close to the drug binding site. To this effect, sites with 2 mutations are considered as "emerging" or "budding" resistance hotspots (Portelli et al., 2018).

One assessment of the impact of missense mutations is to measure the change in a protein structure's as well as drug-target complex's physical interactions that contribute to its overall stability. Computational approaches (e.g., the *mCSM* suite; Pires et al., 2014a, 2016; Pires and Ascher, 2016, 2017; Rodrigues et al., 2019) have been developed to predict the effects of missense point mutations on overall protein structure stability, as well as the binding affinity/stability of ligand, protein-protein, and protein-nucleic acid interactions within a single framework, based on either an experimentally resolved structure or derived model. Here we apply such approaches to the effects of missense point mutations in the *pncA* gene. In addition, we also analyse biophysical structural features including surface area, residue depth and hydrophobicity for residues and sites associated with missense point mutations.

A crystal structure for pncA from *M. tuberculosis* has been determined as a monomeric enzyme of 186 amino acids (19.6 kDa) (Petrella et al., 2011). The structure comprises a 6-stranded parallel beta sheets, with helices on either side forming a single α/β domain with a metal cofactor (iron, Fe2+) binding site formed of D49, H51, H57, and H71. The substrate binding cavity in MtPncA is small, approximately 10 Å deep and 7 Å wide. It consists of highly conserved residues F13 and W68 that are essential in substrate binding with Y103 and H137 limiting access to this cavity (Petrella et al., 2011). The catalytic triad consisting of C138, D8, K96 is indicative of a cysteine-based catalytic mechanism (Petrella et al., 2011). Leveraging this crystal structure, we developed an *in silico* framework to assess the biophysical impact of *pncA* mutations and their resistance risk as determined by GWAS. In this study, we attempt to understand PZA resistance by exploring the relationship

between the genomic features and the biophysical consequences of stability and affinity of nsSNPs, and how this is reflected in differences between *M. tuberculosis* lineages.

MATERIALS AND METHODS

SNP Dataset

The dataset consists of 35,944 *M. tuberculosis* isolates, which has been described recently (Napier et al., 2020). In brief, it encompasses all the main lineages (1, 5, and 6, ancient; 2, 3, and 4, modern; 7 intermediate), and drug susceptibility testing across 8 first-and second-line anti-TB drugs. Across these isolates, mutations in the *pncA* coding region with non-synonymous amino acid changes (nsSNPs) were extracted. These nsSNPs were further annotated for their link with drug resistance as defined by their presence in the TB-Profiler mutation database (Phelan et al., 2019b). Initial analysis aimed at understanding the structure and characterising the active site, followed by *in silico* predictions to quantify the enthalpic and entropic effects of GWAS-identified nsSNPs on the pncA protein structure. Subsequently, additional metadata relating to the clinical isolates were studied in relation to the structural effects of mutations. The general methodology workflow followed in this analysis is similar to the one described previously (Portelli et al., 2018).

Drug and Target: Structural Data

In the absence of a drug (PZA) and target (pncA) complex, respective individual structures were obtained from RSCB PDB database (Berman et al., 2000). The crystal structure of *pncA* in *M. tuberculosis* is available as PDB entry 3PL1 (Petrella et al., 2011), while the structure of PZA was extracted from PDB entry 3R55 (Singh et al., 2011). The molecular motion of pncA was analysed by Normal Mode Analysis using the DynaMut tool (Rodrigues et al., 2018) (**Supplementary Figure 1**).

Protein-Ligand Docking: Autodock Vina

The *pncA*-PZA complex was generated using the software AutoDock Vina, version 1.1.2 (Trott and Olson, 2009). Autodock Vina is an open-source, freely available molecular modelling platform to perform protein-ligand docking. Docking was carried out with default settings and guided by the positioning of the ligand within the active site as descried by Petrella et al. (2011). The complex was generated to facilitate downstream analyses by mCSM-lig (Pires et al., 2016) Autodock Vina returns bound conformations with their respective predicted binding affinity values. The prediction of binding affinity (strength of the ligand interaction with its target) is based on one of several scoring functions, which rank the poses in increasing order of predicted binding affinity. Binding free energy is calculated using a semi-empirical force field, combining experimental and knowledge-based information. The docking poses were visualised and inspected in UCSF Chimera 1.13 (Pettersen et al., 2004) according to the occupation of search space and diversity of pose conformations (**Supplementary Figure 2**). The top two binding poses were closely matched with the conformations generated by Karmakar et al. (2018) and Petrella et al. (2011), respectively (**Supplementary Figure 3**). The best pose was chosen considering the ligand orientation generated by molecular docking performed by Karmakar et al. (2018) and comparing interaction of both poses with active site residues through an Arpeggio (Jubb et al., 2017) analysis (**Supplementary Figure 4**).

Ligand extraction and protonation were carried out using UCSF Chimera, version 1.11 (Pettersen et al., 2004) while identification of rotatable bonds was carried out in Autodock tools (available as part of MGL tools, version 1.5.6) (Morris et al., 2009) where protonation of the ligand is specifically required by Autodock Vina (Trott and Olson, 2009). Similarly, protein extraction and explicit removal of solvent were carried out in UCSF Chimera, version 1.11 (Pettersen et al., 2004), and other steps in the overall protein preparation process were carried out in Autodock tools (part of MGL tools, version 1.5.6) (Morris et al., 2009). All the required parameters to perform docking needed to be included in a configuration file.

In silico Predictions: mCSM DUET, FoldX, mCSM-lig

The computational tools based on mutation cut-off scanning matrix, primarily *mCSM DUET* (Pires et al., 2014a) and *mCSM-lig* (Pires et al., 2016) were used to investigate the structural effects of nsSNPs within the pncA target protein. The effects of nsSNPs within *pncA* were analysed with respect to protein stability (DUET and FoldX (Schymkowitz et al., 2005) and ligand affinity (mCSM-lig). The consequences of these effects were to investigate change in protein fold and function, and effect on mechanism of PZA drug activation, respectively. Results from mCSM-lig (Pires et al., 2016) return both ligand affinity and DUET scores, hence only mCSM-lig was run to obtain both the outputs simultaneously.

A semi-automated pipeline was constructed for mCSM and FoldX to submit and extract results for multiple mutations consecutively using python and shell scripts. Both tools require wild type structure, chain ID and a list of nsSNPs in the X <POS> Y format (X: wild type residue; <POS> : position, Y: mutant residue). The residue symbols (X and Y) are specified as one letter amino acid code. DUET and FoldX estimate mutational impact as a change in Gibbs Free energy ($\Delta\Delta G$) in Kcal/mol. The classification of mutational impact based on $\Delta\Delta G$ from these methods are categorised in opposing ways. For example, $\Delta\Delta G < 0$ of a SNP is classified as a "destabilising" according to DUET, while the same is classified as "stabilising" according to FoldX.

The mutational impact on ligand affinity is calculated as a log fold change between wild type and mutant binding affinities. In addition to SNP identifiers, mCSM-lig requires the ligand affinity of the wild-type protein to be specified in nano Molar (nM) for affinity change calculations. Since the binding affinity returned by AutoDock Vina, version 1.1.2 (Trott and Olson, 2009) is in Kcal/mol, these needed to be converted to nM via Eq. 1 (below). The binding affinity for PZA in nM was 0.9911.

$$\Delta G = -RT\ln K. \quad (1)$$

Equation 1: Calculation of binding free energy, ΔG, where R is the gas constant, 1.987 cal K^{-1} mol^{-1} and T is the absolute temperature, 298 K. Adapted from Morris et al. (1998).

The mCSM suite of tools (Pires et al., 2014a, 2016; Pires and Ascher, 2017; Rodrigues et al., 2019) are based on graph-based measures at an atomic level along with machine learning (ML) tools for predicting enthalpic and entropic effects of stability. mCSM achieves this broadly by generating a signature encompassing the wild-type milieu and change in pharmacophore properties upon mutation (Pires et al., 2014b). Owing to the inter-atomic distance pattern within mCSM describing the wild-type residue environment, novel parameters like residue depth and long-range interactions are implicitly considered. In this manner, mCSM is able to characterise both local and global effects of missense point mutations. The mutational change at the atomic level is considered by using a change in the "pharmacophore count" vector, thus obviating the need to have explicit mutant structure. All mCSM tools (Pires et al., 2014a, 2016; Pires and Ascher, 2016, 2017; Rodrigues et al., 2019) use the atomic changes, while DUET (Pires et al., 2014a) is an ensemble method combining methods of mCSM stability (Pires et al., 2014b) and SDM (Worth et al., 2011; Pandurangan et al., 2017). FoldX, however is an empirical-based prediction tool which summarises the change in stability between mutant and wild type protein structures using a combination of energy terms based on fundamental intramolecular interactions (Schymkowitz et al., 2005).

Other Structural Parameters

Additional structural parameters for wild type structure were also included in the analysis. These were: Accessible (ASA) and Relative Surface Area (RSA), residue depth (RD), hydrophobicity values according to the Kyte-Doolittle scale (KD). The DSSP programme (Kabsch and Sander, 1983; Touw et al., 2015) was run to extract the ASA and RSA values, while RD values calculated as described by Chakravarty and Varadarajan (1999) were calculated using the depth server available at http://cospi.iiserpune.ac.in/depth. The KD values were fetched from the expasy server (Artimo et al., 2012) available at https://web.expasy.org/protscale/.

Data Normalisation: DUET, FoldX, and mCSM-lig

The DUET (Pires et al., 2014a), FoldX (Schymkowitz et al., 2005), and mCSM-lig (Pires et al., 2016) scores associated with each SNP were subsequently normalised between the range of −1 and 1. For mCSM-lig analyses, data was filtered according to distance from interacting site and only residues within a distance of 10 Å of the ligand (PZA) were considered for all ligand affinity analyses.

Minor Allele Frequency and Odds Ratio Calculations: SNP Dataset

Across the *M. tuberculosis* isolates tested for PZA drug susceptibility data, we performed association analysis to estimate the risk of resistance for SNP alleles. For each nsSNP, minor allele frequency (MAF) and odds ratio (OR) were calculated in relation to all samples tested for PZA susceptibility. MAF is the average occurrence of a given nsSNP, and OR is the measure of association of a given nsSNP with PZA resistance. In addition to unadjusted odds ratio (OR), and similar to a GWAS approach, adjusted odds ratio (aOR) were estimated using logistic regression models with a kinship matrix adjusting for a random effect representing the SNP-based relationships between samples (e.g., the lineage-based population structure) (Zhou and Stephens, 2012; Coll et al., 2018). *P*-values were estimated using Fisher and Wald test for unadjusted and adjusted ORs, respectively.

Statistical Analyses

Data was analysed using non-parametric statistical tests. For assessing correlations, Spearman correlation values were calculated. For comparing lineage distributions, the Kolmogorov-Smirnov (KS) test was used. Statistical significance thresholds used are $^{*}P < 0.05$, $^{**}P < 0.01$, $^{***}P < 0.001$, $^{****}P < 0.0001$).

Data Visualisation

All plots were generated using R statistical software, version 4.0.2 (R Core Team, 2014). Protein and ligand structures were generated using UCSF Chimera, version 1.11 (Pettersen et al., 2004).

RESULTS

Analysing the pncA Molecular Motion and pncA-PZA Complex

Molecular motion in pncA was analysed by Normal Mode Analysis (NMA). Regions undergoing the greatest movement were limited to residues in loop regions and mainly concentrated to loop 60–66, followed by loop residues 39–41 and 111–113. Residues at site 165–167 within helix 164–178 showed the least flexibility (**Supplementary Figure 1**). The frequency of mutations in these variable regions was most prominent for sites 62–63 (>2 mutations) while the other sites were limited to at most two mutations (**Figure 1**). Mutations within the most flexible region (residues 60–66) of pncA showed mixed effects in relation to their association with PZA resistance with the single mutation at site 64 related to PZA resistance. Sites 39 and 40 within the other highly flexible region 39–41 were not associated with any mutations in our study, while the two mutations at site 41 were not associated with PZA resistance. The region 111–113 is associated with single mutations at sites 111 and 112 which are not linked to PZA resistance, while site 113 was not associated with any mutations in our study. Sites 165–167, which form part of the helix (164–178), are the most stable according to NMA. Two residues (A165 and D166) within this helix were not associated with any mutations in our study, while a single mutation at site T167 was not associated with PZA drug resistance (**Supplementary Figure 1** and **Supplementary Table 1**). Docking with AutoDock vina (Trott and Olson, 2009) generated nine different conformations as per default settings. In six of these poses, the aromatic ring of PZA was oriented towards the substrate binding residue W68 (**Supplementary Figures 2A,B**). The top two poses (1 and 2) returned by Vina were similar to previous molecular docking studies (Petrella et al., 2011; Karmakar et al., 2018)

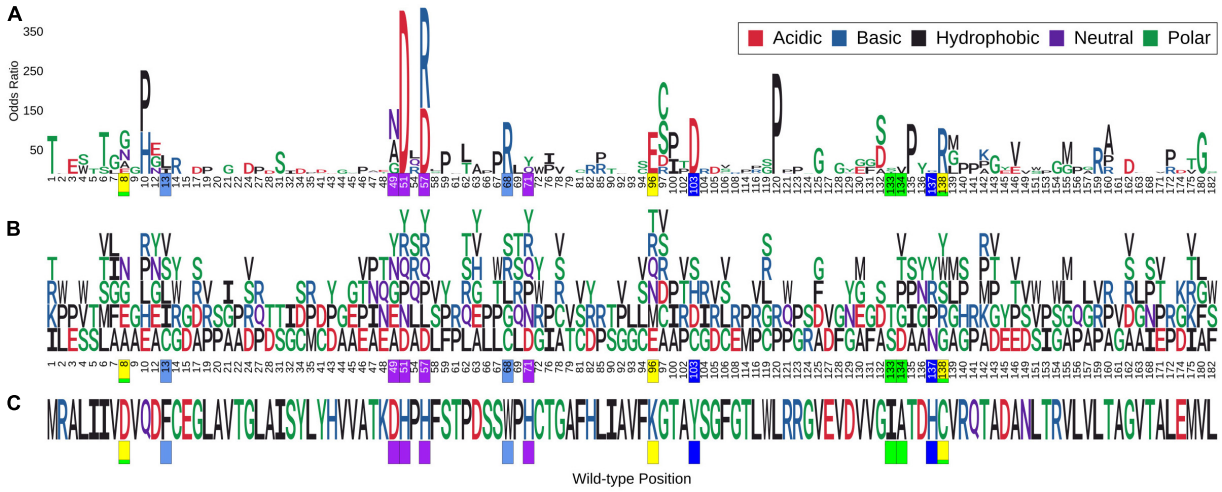


FIGURE 1 | Logo plot showing sites with multiple missense point mutations and association with Odds Ratio. Sites associated with multiple (>2) missense point mutations (i.e., nsSNPs). A total of 386 mutations corresponding to 113 positions on the pncA protein structure were associated with multiple nsSNPs. The horizontal axis in **(A,B)** show the position numbers of sites with multiple nsSNPs, while part **(C)** shows the wild-type residues for each position. The vertical axis in **(A)** represents Odds Ratio (OR) where letters denote mutant residues which are proportional to their corresponding OR highlighting the most resistant mutation at each site and overall. Part **(B)** shows each mutant residue at a given position, highlighting nsSNP diversity by position. The wild-type and mutant residues are coloured according to the amino acid properties as denoted. Positions marked in yellow form the catalytic triad, residues in blue and teal are involved in substrate binding, those in green are involved in hydrogen binding while the ones in purple are involved in the iron centre coordination. The figure is generated using R statistical software (version 4.0.2). nsSNPs, non-synonymous Single Nucleotide Polymorphisms; pncA, pyrazinamidase.

(**Supplementary Figure 3**). A follow-up Arpeggio analysis (Jubb et al., 2017) indicated that pose 1 when compared to pose 2, has more H-bonds (4 vs. 1), fewer aromatic contacts (3 vs. 13), and greater Van der Waals interactions (3 vs. 1) (**Supplementary Figures 4A,B**). Therefore, model with pose 1 was chosen to form the pncA-PZA complex (**Supplementary Figure 5**).

Genomics Data

SNP data from 35,944 *M. tuberculosis* clinical isolates tested for drug susceptibility to a range of first and second line drugs were obtained (Napier et al., 2020). Among these, 39% (n = 13,914) of these isolates were tested for PZA drug susceptibility. The isolates were collected from over 30 different countries and represented the 4 main *M. tuberculosis* lineages (L1, n = 144; L2, n = 1,886; L3, n = 190; L4, n = 2213) (**Supplementary Figure 6**). In order to infer whether the ancestral pncA sequences for each lineage differed, we quantified the number of samples without any mutations in each lineage. The majority of isolates in L1–L4 had an identical *pncA* sequence as the H37Rv reference indicating that the ancestral sequences for these lineages do not differ. The majority were pan susceptible (n = 23,256, 64.7%), with the remainder MDR-TB (n = 6,691, 18.6%), XDR-TB (n = 989, 2.8%), or another type of resistance referred to as DR-TB (n = 5,008, 13.9%) (**Table 1**). From the list, only nsSNPs within the protein coding region of *pncA* (n = 4,731, 13.2%) were considered for our analyses (**Table 1**). The majority of these were MDR-TB (n = 3,290, 69.5%) followed by relatively equal numbers of XDR-TB and DR-TB (n = 625, 13.2% and n = 632, 13.4%, respectively), while only a small percentage were susceptible (n = 184, 3.9%) (**Table 1**). From a total of 13,914 samples tested for PZA drug susceptibility, a minority of those were found to be resistant (n = 2,379, 17.1%) (**Table 1**). However, the burden of PZA resistance among

TABLE 1 | Number of samples analysed.

Item name	Total number (%)
Clinical isolates/samples	35,944
Samples classified Susceptible	23,256 (64.7)
Drug resistant (DR)	5,008 (13.9)
Multi-drug resistant (MDR)	6,691 (18.6)
Extreme drug resistant (XDR)	989 (2.8)
Samples tested for PZA drug susceptibility	13,914
Resistant	2,379 (17.1)
Samples with nsSNPs in the protein coding region of *pncA*	4,731 (13.2)
Susceptible	184 (3.9)
Drug resistant (DR)	632 (13.4)
Multi-drug resistant (MDR)	3,290 (69.5)
Extreme drug resistant (XDR)	625 (13.2)
Samples with *pncA* nsSNPs tested for PZA drug susceptibility	2,289 (48.4)
Samples with *pncA* nsSNPs resistant to PZA	1,677 (73.3)
Unique nsSNPs: No. of sites	424 nsSNPs: 151 sites

Summary of clinical isolates from genome-wide analysis. PZA, pyrazinamide; nsSNPs, non-synonymous Single Nucleotide Polymorphisms.

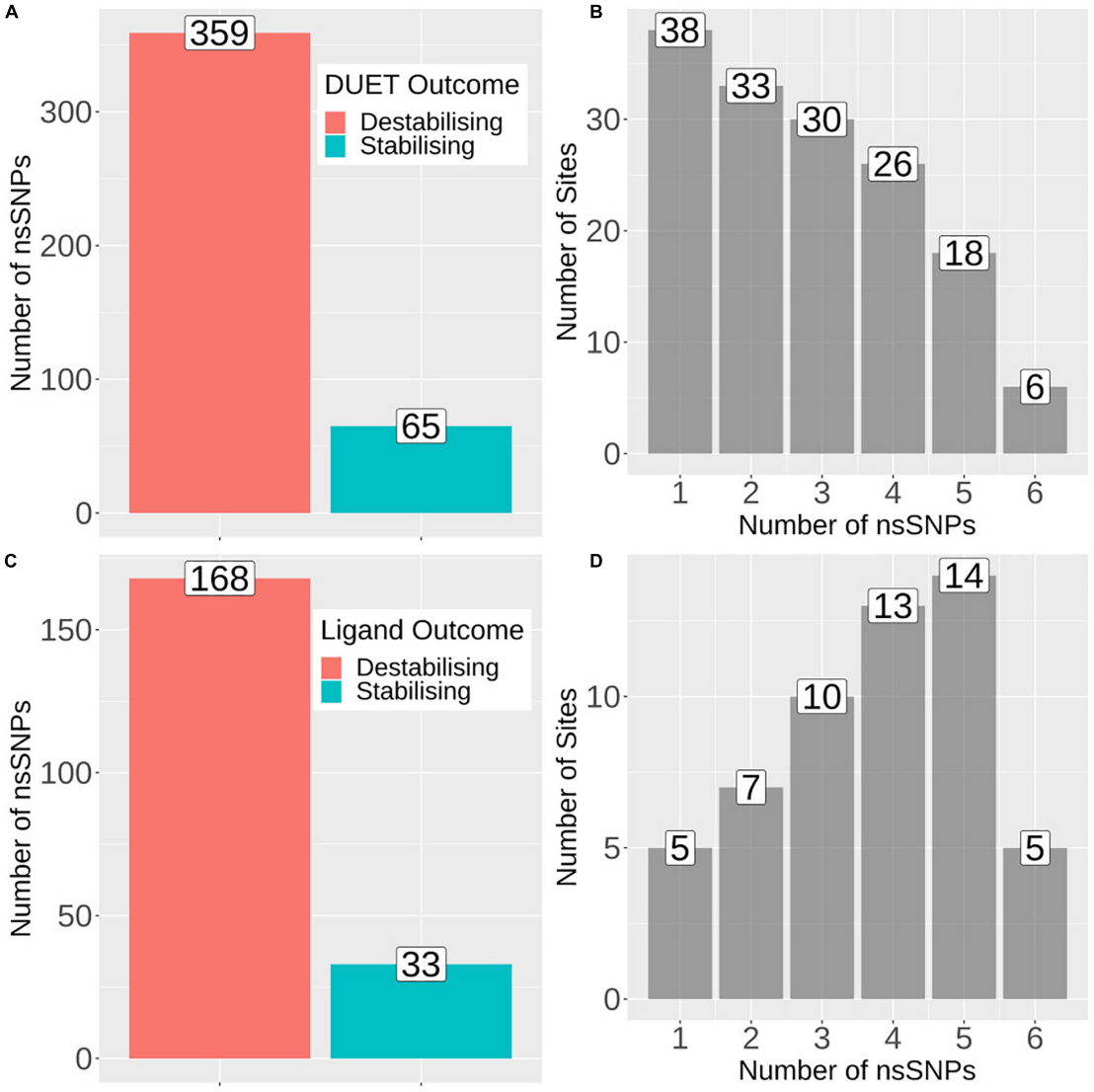


FIGURE 2 | Barplots showing number of mutations and sites associated with protein stability and ligand affinity. **(A)** Number of nsSNPs categorised as destabilising (n = 359) and stabilising (n = 65) according to DUET protein stability. **(B)** Frequency of sites associated with the number of nsSNPs, where horizontal axis denotes the number of nsSNPs and vertical axis denotes the total number of sites/positions corresponding to the number of nsSNPs. **(C)** Barplot showing the number of nsSNPs categorised as destabilising (n = 168) and stabilising (n = 33) according to mCSM ligand affinity where sites lie within 10Å of ligand. **(D)** Frequency of sites associated with the number of nsSNPs, where horizontal axis denotes the number of nsSNPs and vertical axis denotes the total number of sites/positions corresponding to the number of nsSNPs. The figure is generated using R statistical software (version 4.0.2). nsSNPs, non-synonymous Single Nucleotide Polymorphisms.

samples containing nsSNPs in the protein coding region was high (n = 1,677, 73.3%) (**Table 1**).

Across the 4,731 isolates, 424 distinct nsSNPs corresponding to 151 distinct amino acid positions on the pncA structure were identified (**Figures 2A,B**). A total of 201 nsSNPs corresponding to 54 amino acid changes were within 10 Å of the ligand binding site (**Figures 2C,D**). The majority of these nsSNP mutations have been annotated as being linked to PZA resistance within the TBProfiler tool (227/424). The majority of these nsSNP mutations have been annotated as being linked to PZA resistance within the TBProfiler tool (227/424; denoted as DM), while the others (197/424; denoted as OM) were assumed to have weak or no links. Genomic measures like minor allele frequency (MAF) and odds ratio (OR) were obtained for a total of 322 nsSNPs, with adjusted OR (aOR) estimated for a total of 163 nsSNPs. Across the majority of these nsSNPs, the MAFs were low (median: 0.02% range: 0.01–2.11%) (**Supplementary Figure 7A**). Similarly, when considering ORs, the majority of the nsSNPs had high ORs (median: 9.70, range: 0.22–414.61) (**Supplementary Figure 7D**). When looking at the distribution of MAF and OR within mutations associated with PZA resistance (DM) and other mutations (OM) (**Supplementary Figures 7B,E**), DM mutations

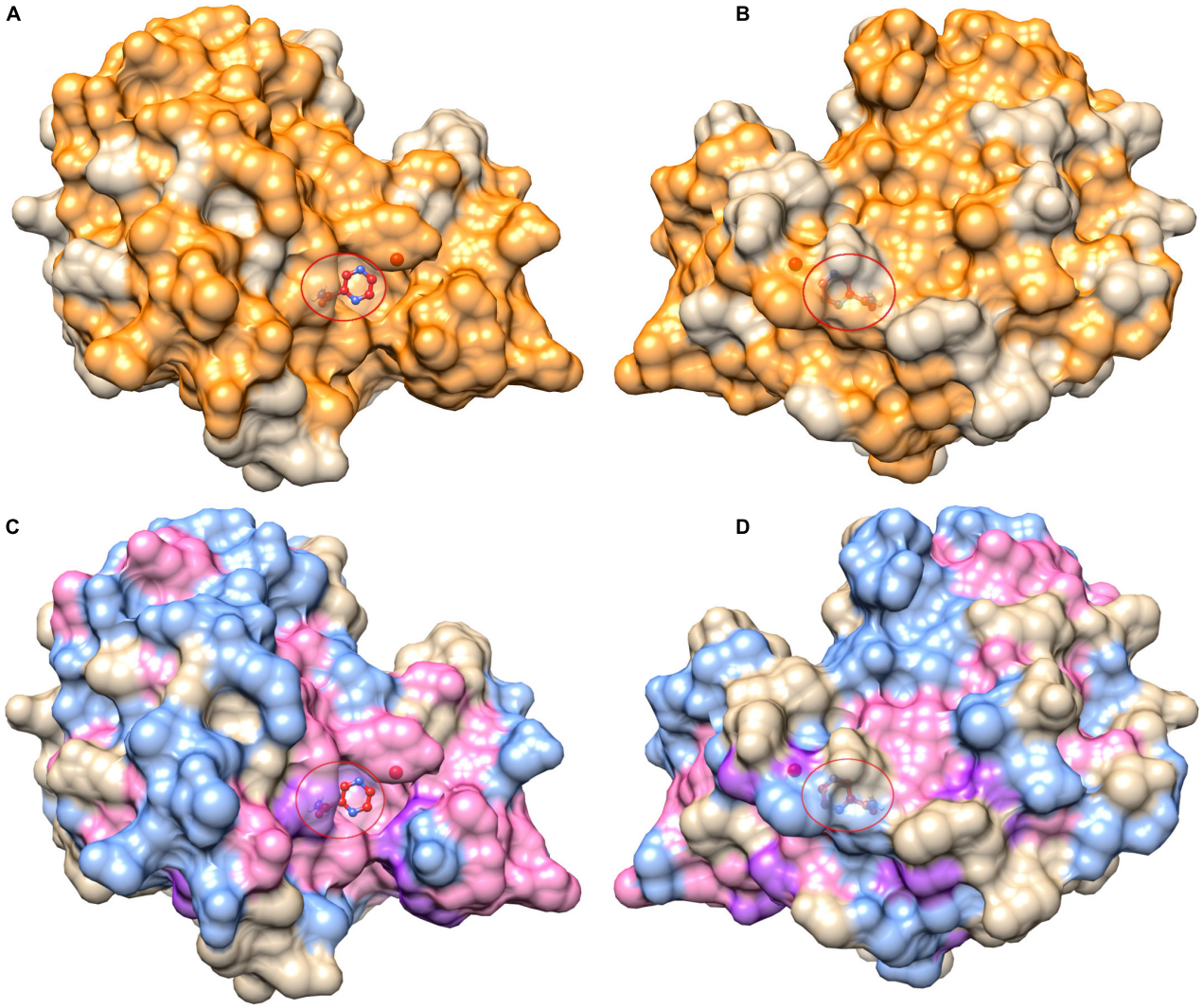


FIGURE 3 | Mutational landscape of pncA structure (3PL1) coloured by positions linked to pyrazinamide drug (PZA) resistance. Panels **(A,B)** show all mutational positions in orange while mutational positions in **(C,D)** are further coloured by mutations classed as either drug resistant mutations (purple) or "other mutations" (blue), while sites linked to mutations belonging to either category are coloured in pink. The right panels **(B,D)** depict the corresponding structure rotated by 180°. The ligand (PZA) is shown as ball and stick within the active site denoted by the red circle. The figure is rendered using UCSF Chimera (version 1.14). pncA, pyrazinamidase.

were associated with significantly higher ($P < 0.0001$) MAF and OR (**Supplementary Figures 7C,F**).

Understanding Mutational Effects on pncA Stability and PZA Binding Affinity

The 424 nsSNPs mapped onto the crystal structure of pncA revealed that mutational landscape of pncA appears distributed (**Figures 3A,B**) throughout the structure. Sites linked to drug resistant mutations were predominant around the PZA binding (active) site, while sites exclusively linked to mutations classed in the "other" category are distal to the active site (**Figures 3C,D**, **4**). Furthermore, active site residues were associated with a multiple point mutation (**Table 2** and **Figures 1B**, **5C**). All active site and hydrogen-bond forming residues with the ligand were associated with multiple mutations (≥2) (**Figure 1B**), thus representing the high diversity of mutations present within pncA. Despite this, there appears to be some degree of clustering around positions 4–14, 46–97, 132–143 involving the active site and metal centre residues (**Figure 5C**).

The biophysical effect of mutations on protomer stability, estimated as $\Delta\Delta G$ (Kcal/mol), was measured using DUET (Pires et al., 2014a) and FoldX (Schymkowitz et al., 2005), while mutational impact on ligand affinity was measured using mCSM-lig (Pires et al., 2016) (see section "Materials and Methods"). Assessing mutational effects on protein stability as measured by

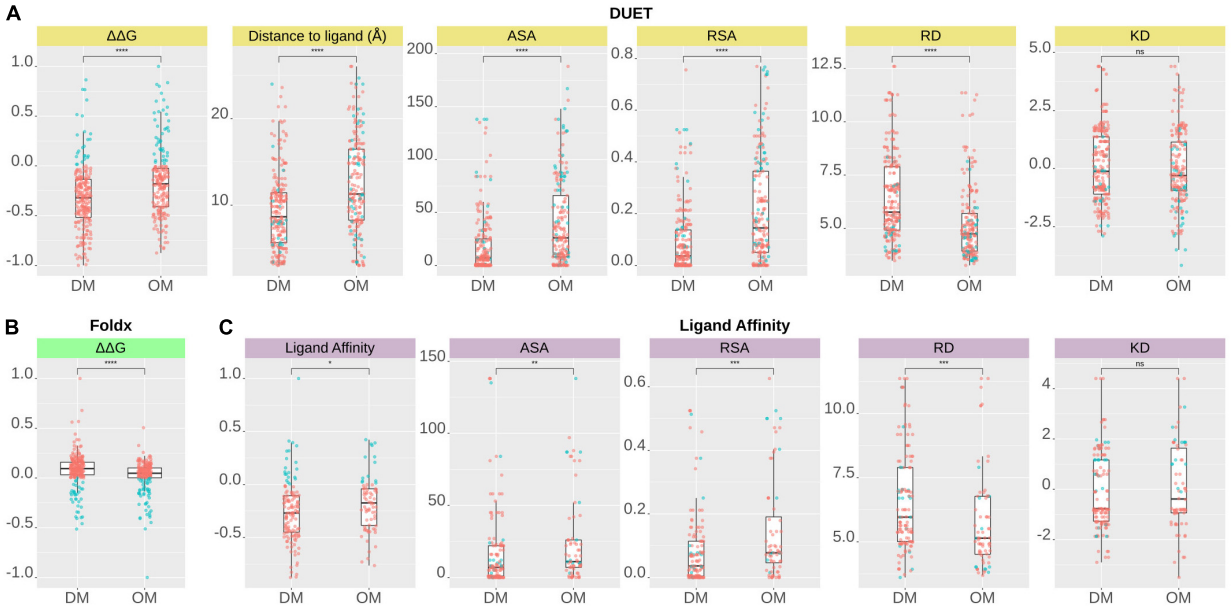


FIGURE 4 | Comparison of structural features between Drug resistance (DM) and other mutations (OM) of pncA gene mutations according to **(A)** DUET protein stability (Δ ΔG), **(B)** FoldX stability (Δ ΔG), and **(C)** Ligand Affinity. A total of 424 nsSNPs for DUET and FoldX (DM, n = 227, OM, n = 197), while a total of 201 nsSNPs (DM, n = 129 OM, n = 72) lying within 10 Å of PZA for ligand affinity were included in the analysis. DM and OM mutations were compared using Wilcoxon rank-sum (unpaired) and statistical significance indicated as: *P < 0.05, **P < 0.01, ***P < 0.001, ****P < 0.0001). The figure is generated using R statistical software (version 4.0.2). ns, non-synonymous Single Nucleotide Polymorphisms; pnca, pyrazinamidase; PZA, pyrazinamide; Å, Angstroms; Δ ΔG, Change in Gibbs free energy in Kcal/mol; ASA, Accessible Surface Area; RSA, Relative surface Area; RD, Residue Depth; KD, Kyte-Doolittle Hydrophobicity values.

DUET, nearly 85% had a destabilising effect (n = 359) compared to nearly 15% mutations with stabilising effects (n = 47) as shown in **Figure 2A**. When assessing ligand affinity, 47.4% (n = 201) SNP mutations were present within 10 Å of the PZA binding site (**Figure 2C**). Similar to DUET stability effects, the majority (84%; n = 168) of nsSNPs were destabilising while 16% (n = 27) were stabilising for ligand binding affinity (**Figure 2C**). More than 50% of the mutational positions were associated with multiple nsSNPs for both protein stability (n = 113) and ligand affinity (n = 49) (**Figures 2B,D**). The average protein stability and ligand affinity effects of all mutations mapped onto the pncA structure (**Figures 5A,B**), highlight mutations with opposing effects for protein stability and ligand affinity. These effects are pronounced for active site residues (I133, A134, H137, C138) (**Figures 5C,D**).

There were 80 sites within *pncA a*ssociated with multiple nsSNPs (>2) (**Figures 1B**, **2B**) which included all active residues except I133 which was associated with 2 mutations (**Figure 1B**). Sites with 2 nsSNPs are considered to be budding resistance hotspots (n = 33 for protein stability, n = 7 for ligand affinity). A total of 57 nsSNPs within 5 Å of PZA were considered to be within the first shell of residues lining the active site (**Table 2**). While majority of the mutational sites associated with more than two mutations comprise of destabilising mutations, positions 1, 2, 10, 12, 43, 46, 51, 57, 63, 67, 69, 78, 82, 92, 96, 100, 104, 105, 129, 135–138, 142, 149, 164, 168, and 174 comprised of both stabilising and destabilising mutations (**Figure 5C**). Similarly, for ligand affinity, most mutational sites had destabilising mutational effects, with positions 7, 8, 13, 27, 49, 72, 78, 96, 102, 103, 105, 134, 137, 138, and 162 associated with mutations resulting in mixed stability impact. Position 163 comprised only of mutations with stabilising effects (**Figure 5D**). The budding resistance hotspot active site residue I133 contained both mutations with destabilising effect for protein stability (**Figure 5C**), while stabilising for ligand affinity (**Figure 5D**). Similarly, for budding resistance hotspots, majority of the nsSNPs were associated with destabilising effects. For protein stability, 9/33 sites had mutations with mixed stability (positions 15, 32, 61, 66, 76, 114, 127, 153, and 161) (**Figure 5C**), while only position 20 showed mixed stability effects for ligand affinity (**Figure 5D**).

Mutations With Extreme Effects

Mutations with extreme effects on protein stability and affinity are summarised in **Table 3**. Overall, the most destabilising mutation according to DUET was L4S, where a change from a hydrophobic to a polar residue may contribute to disruption of local conformation (**Table 3**). The closest most destabilising mutational effect on protein stability was from A134D (wild-type residue involved in hydrogen bonding) (**Table 3**), likely resulting in electrostatic and steric clashes due to a change in charge and volume affecting the overall stability negatively. The most stabilising mutation on protomer stability was from active site residue Y103D, while the closest such mutation was C138R (**Table 3**). The stabilising effect of these mutations on the protein stability and ligand affinity is thought to result from the electrostatic interactions working favourably for sites lying within 5 Å of the ligand. The most destabilising mutation according

TABLE 2 | Mutations close to the active site of PZA.

S. No.	Mutation	Mutation class	MAF (%)	OR	P-value	OR adjusted	P-Wald	DUET ΔΔG	DUET outcome	Distance to ligand (Å)	mCSM-lig (log affinity)	Ligand outcome	Foldx ΔΔG	Foldx outcome	ASA	RSA	Hydro phobicity	Residue depth
1	A134D	Others	0.01	2.42	1.00E+00	NA	NA	−2.98	D	3.05	0.58	S	1.03	D	10	0.08	1.87	6.77
2	A134G	Others	NA	NA	NA	NA	NA	−1.62	D	3.05	−0.38	D	−1.29	S	10	0.08	1.87	6.77
3	A134P	Others	0.01	9.70	1.71E-01	NA	NA	−1.43	D	3.05	0.08	S	−5.20	S	10	0.08	1.87	6.77
4	A134T	Others	NA	NA	NA	NA	NA	−1.93	D	3.05	0.88	S	−0.94	S	10	0.08	1.87	6.77
5	A134V	Drug associated	0.04	19.43	3.68E-03	1.53	3.07E-05	−0.41	D	3.05	0.12	S	−1.46	S	10	0.08	1.87	6.77
6	I133S	Others	0.01	9.70	1.71E-01	NA	NA	−3.22	D	3.05	0.58	S	3.30	D	3	0.02	1.97	7.90
7	I133T	Drug associated	0.32	6.44	2.90E-09	0.86	4.86E-03	−2.79	D	3.05	0.70	S	1.58	D	3	0.02	1.97	7.90
8	D8A	Drug associated	0.01	19.41	2.92E-02	NA	NA	−0.51	D	3.22	−3.27	D	0.54	D	5	0.03	1.63	9.48
9	D8G	Drug associated	0.08	48.69	1.95E-07	1.25	4.42E-02	−0.85	D	3.22	−3.45	D	1.89	D	5	0.03	1.63	9.48
10	D8E	Drug associated	0.03	14.56	1.74E-02	1.19	1.46E-01	−0.79	D	3.22	0.01	S	1.90	D	5	0.03	1.63	9.48
11	D8N	Drug associated	0.05	29.16	1.49E-04	1.24	7.10E-03	−1.18	D	3.22	−1.66	D	−1.26	S	5	0.03	1.63	9.48
12	C138G	Others	NA	NA	NA	NA	NA	−0.02	D	3.28	−0.01	D	1.12	D	12	0.07	1.17	6.70
13	C138S	Drug associated	NA	NA	NA	NA	NA	0.00	D	3.28	0.81	S	−0.23	S	12	0.07	1.17	6.70
14	C138W	Others	NA	NA	NA	NA	NA	−1.05	D	3.28	0.94	S	−1.72	S	12	0.07	1.17	6.70
15	C138Y	Drug associated	NA	NA	NA	NA	NA	−0.52	D	3.28	0.91	S	−0.57	S	12	0.07	1.17	6.70
16	C138R	Drug associated	0.09	116.96	6.10E-10	1.74	4.08E-12	0.10	S	3.28	0.35	S	−2.12	S	12	0.07	1.17	6.70
17	H137N	Others	0.01	2.42	1.00E+00	NA	NA	0.19	S	3.42	−0.12	D	0.40	D	84	0.38	−1.40	4.60
18	H137P	Drug associated	NA	NA	NA	NA	NA	0.37	S	3.42	−0.77	D	2.19	D	84	0.38	−1.40	4.60
19	H137Y	Others	0.01	2.42	1.00E+00	NA	NA	0.86	S	3.42	−0.01	D	0.34	D	84	0.38	−1.40	4.60
20	H137R	Drug associated	0.03	4.85	1.38E-01	0.56	1.21E-04	−0.27	D	3.42	0.47	S	0.49	D	84	0.38	−1.40	4.60

(Continued)

TABLE 2 | Continued

S. No.	Mutation	Mutation class	MAF (%)	OR	P-value	OR adjusted	P-Wald	DUET ΔΔG	DUET outcome	Distance to ligand (Å)	mCSM-lig (log affinity)	Ligand outcome	Foldx ΔΔG	Foldx outcome	ASA	RSA	Hydro phobicity	Residue depth
21	D49G	Drug associated	0.05	29.16	1.49E-04	1.66	4.38E-08	−1.16	D	3.45	−3.46	D	0.46	D	7	0.04	−1.53	7.89
22	D49A	Drug associated	0.04	58.33	2.49E-05	1.67	3.17E-06	−0.45	D	3.45	−3.35	D	−2.07	S	7	0.04	−1.53	7.89
23	D49N	Drug associated	0.06	77.84	7.23E-07	1.51	3.14E-04	−1.68	D	3.45	−1.93	D	−0.33	S	7	0.04	−1.53	7.89
24	D49Y	Drug associated	0.01	9.70	1.71E-01	NA	NA	−0.74	D	3.45	−1.86	D	−2.67	S	7	0.04	−1.53	7.89
25	D49E	Drug associated	0.02	9.70	7.77E-02	NA	NA	−0.47	D	3.45	0.25	S	−0.70	S	7	0.04	−1.53	7.89
26	A102R	Others	0.01	2.42	1.00E+00	NA	NA	−0.70	D	3.50	0.17	S	4.13	D	10	0.08	0.03	5.51
27	A102P	Others	0.06	14.58	5.08E-04	0.66	5.33E-04	−1.25	D	3.50	−0.23	D	−0.62	S	10	0.08	0.03	5.51
28	A102V	Others	0.06	2.43	1.88E-01	0.91	3.00E-01	−0.25	D	3.50	−0.16	D	−1.91	S	10	0.08	0.03	5.51
29	A102T	Drug associated	0.01	19.41	2.92E-02	1.75	4.98E-04	−0.72	D	3.50	0.88	S	−2.03	S	10	0.08	0.03	5.51
30	F13C	Others	0.01	1.21	1.00E+00	0.64	4.31E-03	−2.32	D	3.55	−0.49	D	2.70	D	24	0.10	0.60	6.93
31	F13I	Drug associated	0.03	14.56	1.74E-02	NA	NA	−1.76	D	3.55	−0.45	D	0.89	D	24	0.10	0.60	6.93
32	F13L	Drug associated	0.06	34.04	2.89E-05	1.37	2.29E-03	−2.03	D	3.55	−0.43	D	1.10	D	24	0.10	0.60	6.93
33	F13V	Others	0.01	1.21	1.00E+00	NA	NA	−2.57	D	3.55	−0.56	D	1.40	D	24	0.10	0.60	6.93
34	F13S	Drug associated	0.03	1.62	5.28E-01	0.60	3.07E-04	−3.10	D	3.55	0.22	S	2.59	D	24	0.10	0.60	6.93
35	K96E	Drug associated	0.08	107.17	3.58E-09	1.75	2.79E-06	−2.12	D	3.98	−0.67	D	6.92	D	8	0.03	−1.87	5.96
36	K96Q	Drug associated	0.03	4.85	1.38E-01	0.64	1.17E-01	−1.32	D	3.98	−0.08	D	1.04	D	8	0.03	−1.87	5.96
37	K96T	Drug associated	0.09	58.47	6.68E-09	1.84	2.25E-13	−0.86	D	3.98	−0.57	D	3.54	D	8	0.03	−1.87	5.96
38	K96M	Others	0.01	19.41	2.92E-02	NA	NA	0.41	S	3.98	−1.03	D	0.27	D	8	0.03	−1.87	5.96
39	K96N	Drug associated	0.01	2.42	1.00E+00	NA	NA	−1.16	D	3.98	0.33	S	2.61	D	8	0.03	−1.87	5.96

(Continued)

TABLE 2 | Continued

S. No.	Mutation	Mutation class	MAF (%)	OR	P-value	OR adjusted	P-Wald	DUET ΔΔG	DUET outcome	Distance to ligand (Å)	mCSM-lig (log affinity)	Ligand outcome	Foldx ΔΔG	Foldx outcome	ASA	RSA	Hydro phobicity	Residue depth
40	**K96R**	Drug associated	0.11	19.49	1.66E-07	1.43	2.16E-06	−0.17	D	3.98	0.08	S	−0.74	S	8	0.03	−1.87	5.96
41	**H71D**	Drug associated	0.01	9.70	1.71E-01	NA	NA	−2.69	D	4.18	−2.50	D	5.75	D	5	0.02	−0.77	6.25
42	**H71N**	Drug associated	NA	NA	NA	NA	NA	−2.67	D	4.18	−1.34	D	0.64	D	5	0.02	−0.77	6.25
43	**H71P**	Others	0.01	4.85	3.13E-01	NA	NA	−2.36	D	4.18	−2.89	D	3.26	D	5	0.02	−0.77	6.25
44	**H71Q**	Drug associated	0.01	19.41	2.92E-02	1.75	2.12E-04	−2.29	D	4.18	−1.73	D	1.12	D	5	0.02	−0.77	6.25
45	**H71R**	Drug associated	0.05	1.94	3.42E-01	0.88	2.01E-01	−1.93	D	4.18	−0.83	D	−1.52	S	5	0.02	−0.77	6.25
46	**H71Y**	Drug associated	0.18	25.67	4.52E-13	1.48	5.50E-08	−0.46	D	4.18	−1.96	D	−1.78	S	5	0.02	−0.77	6.25
47	**H57D**	Drug associated	0.73	166.91	2.08E-72	1.24	1.05E-01	−1.85	D	4.56	−1.28	D	1.83	D	16	0.07	−1.30	5.63
48	**H57P**	Drug associated	0.03	38.85	8.53E-04	1.55	1.16E-02	−1.23	D	4.56	−2.12	D	0.15	D	16	0.07	−1.30	5.63
49	**H57Q**	Others	NA	NA	NA	NA	NA	−1.29	D	4.56	−0.95	D	0.85	D	16	0.07	−1.30	5.63
50	**H57R**	Drug associated	0.19	254.92	1.02E-20	1.48	9.69E-09	−1.17	D	4.56	−0.28	D	1.25	D	16	0.07	−1.30	5.63
51	**H57L**	Drug associated	NA	NA	NA	NA	NA	−0.06	D	4.56	−1.92	D	−1.11	S	16	0.07	−1.30	5.63
52	**H57Y**	Drug associated	0.02	29.13	4.99E-03	2.08	7.92E-06	0.41	S	4.56	−1.16	D	−0.15	S	16	0.07	−1.30	5.63
53	**W68C**	Drug associated	0.04	24.29	7.49E-04	1.75	1.67E-04	−1.45	D	4.97	−1.58	D	2.68	D	45	0.16	−1.10	5.49
54	**W68G**	Drug associated	0.14	87.93	2.36E-13	1.58	7.39E-11	−2.57	D	4.97	−2.13	D	4.04	D	45	0.16	−1.10	5.49
55	**W68L**	Drug associated	NA	NA	NA	NA	NA	−1.62	D	4.97	−2.24	D	0.19	D	45	0.16	−1.10	5.49
56	**W68R**	Drug associated	0.20	132.41	4.03E-20	1.50	4.26E-09	−1.61	D	4.97	−0.58	D	0.08	D	45	0.16	−1.10	5.49
57	**W68S**	Drug associated	0.01	9.70	1.71E-01	NA	NA	−2.67	D	4.97	−1.04	D	2.65	D	45	0.16	−1.10	5.49

Fifty-seven mutations (nsSNPs) lying within 5 Å of PZA and the corresponding GWAS measures of minor allele frequency (MAF), Odds Ratio (OR), P-values, adjusted OR (aOR), and P-values from Wald test corresponding to aORs, along with structural measures of distance to ligand, DUET, FoldX, ligand affinity values and effect. Wild type residues for mutations highlighted and marked in green are considered to participate in hydrogen bonding, those in yellow form the catalytic triad, residues in teal (and blue) are involved in substrate binding, while the residues in purple are involved in the iron centre. The columns are coloured to highlight the most significant column attribute with deeper colours denoting the greatest effects. The dark colours in MAF, OR, and aOR columns indicate the highest values, while P-values are coloured with the darkest colour showing the most significant values. Values in the DUET, mCSM-lig, and FoldX columns are coloured according to the extent of their respective effects with red indicating destabilising and blue denoting stabilising effects. nsSNPs, non-synonymous Single Nucleotide Polymorphisms; PZA, pyrazinamide; GWAS, Genome-Wide Association Studies. D, Destabilising; S, Stabilising.

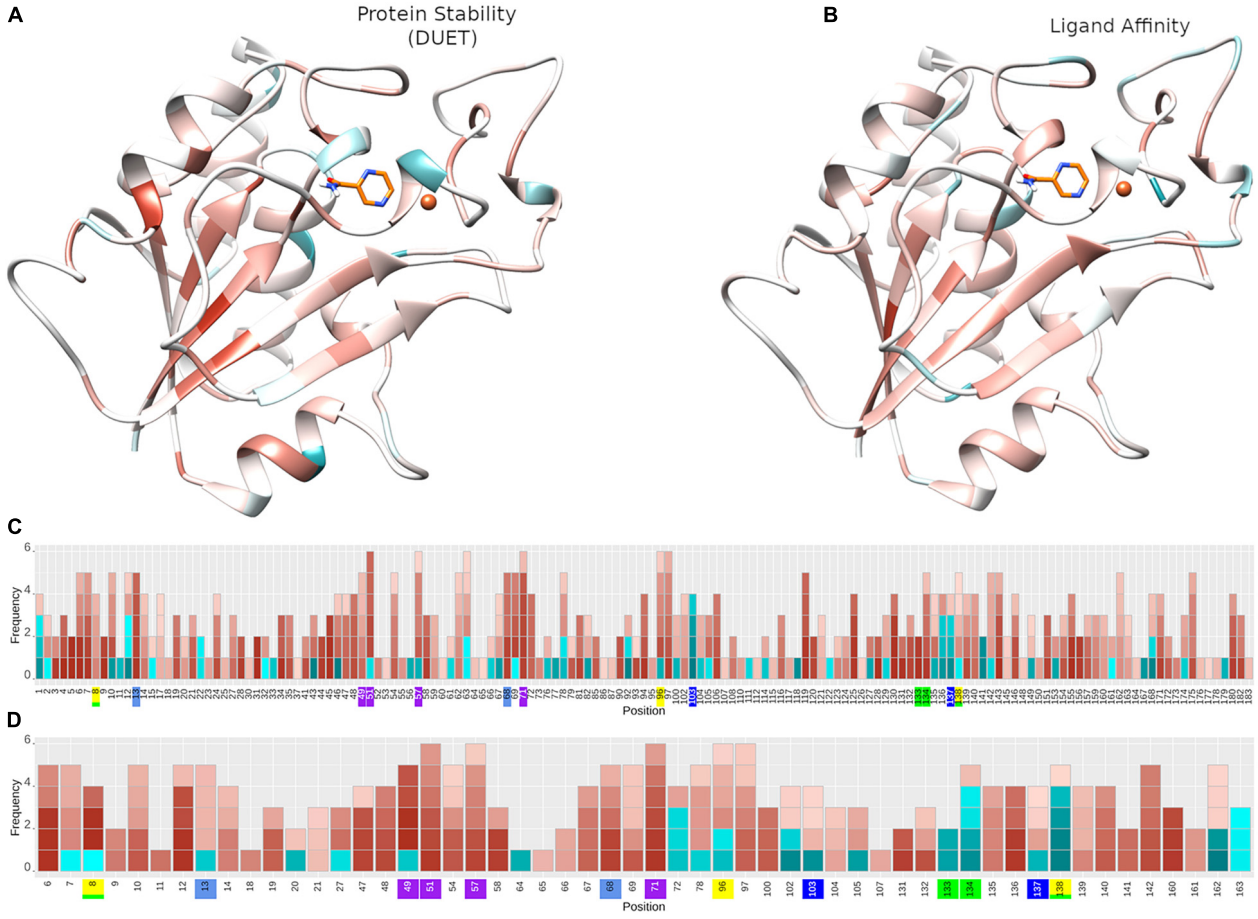


FIGURE 5 | Protein stability and ligand affinity effects of nsSNPs on pncA structure and by position. Mutational impact of nsSNPs on the pncA protein structure coloured by average **(A)** DUET Protein stability (n = 424) and **(B)** ligand affinity (n = 201). Barplots **(C,D)** showing the frequency of mutations within the pncA gene. The horizontal axis shows the mutational positions within pncA and the vertical axis shows the frequency of mutations. Positions on the horizontal axis are coloured to denote the active site residues: green (residues involved in hydrogen bonding with PZA), yellow (catalytic triad), blue and teal (substrate binding), purple (iron centre). For a given position, each corresponding mutation (nsSNP) is coloured by the level of stability according to **(C)** DUET(n = 424) and **(D)** Ligand affinity (n = 201) where the horizontal axis denotes amino acid positions in pnca, and is restricted to positions lying within 10 Å of PZA for ligand affinity. Destabilising mutations are depicted in red and stabilising mutations in blue, where colour intensity reflects the extent of effect, ranging from −1 (most destabilising) to + 1 (most stabilising). The structural figures **(A,B)** are rendered using UCSF Chimera (version 1.14). The barplot figure **(C,D)** is generated using R statistical software (version 4.0.2). nsSNPs, non-synonymous Single Nucleotide Polymorphisms; PZA, pyrazinamide; pncA, pyrazinamidase.

to ligand affinity was D49G located at ∼3.5 Å (**Table 3**). The three subsequent destabilising mutations for ligand affinity were also all within 5 Å of PZA binding site namely D8G (∼3 Å), D49A (∼3.5 Å), and D8A (∼3 Å) (**Supplementary Table 1**), all arising likely due to the loss of charge and volume interfering with ligand interaction. The mutation with the greatest stabilising effect on ligand affinity was G162D, located at ∼8 Å, i.e. outside the first shell of influence (>5 Å) from the ligand. This is possibly due to the resulting electrostatic effects and increase in volume, which may favour hydrogen bond formation with nearby residues and PZA binding, thereby increasing affinity (**Table 3**). The closest most stabilising mutational impact on ligand affinity was due to mutation A134P, though this was a marginal effect (**Table 3**). The most destabilising mutation according to FoldX was C72W, which is located far away from the active site (∼27 Å). Interestingly, mutation A134P was the most stabilising according to FoldX, while the same was estimated to have a destabilising effect according to DUET (**Table 3**). All mutations except A134D and A134P were associated with PZA drug resistance (**Table 3**).

Relating Structural and GWAS Analyses

The minor allele frequencies for the 424 nsSNPs were mapped onto their corresponding amino acid positions of the *pncA* gene (**Supplementary Figure 8**). Position 10 had the highest cumulative minor allele frequency (MAF, ∼2.3%), followed by position 7 (∼1.2%), position 57 (∼1.0%), position 51 (∼0.6%), and position 14 (0.5%). The risk of PZA resistance from the alleles at each SNP was estimated by calculating ORs and P-values using a GWAS approach. Additionally, adjusted OR (aOR) which accounted for the confounding effects of lineage were also

TABLE 3 | Mutations with extreme effects.

Mutational effects	Mutation	Mutation class	MAF (%)	OR	*P*-value	Distance to ligand (Å)	Stability ΔΔG	Ligand affinity
Highest OR	H51D	Drug-associated	0.30	414.61	4.49E-33	5.66	−2.2	−1.82
Most frequent mutation	Q10P	Drug-associated	2.11	156.23	1.28E-207	6.02	−0.63	−1.77
Most deStabilising for protein stability (DUET)	L4S	Drug-associated	0.25	28.46	5.63E-18	15.33	−3.87	−1.08
Closest destabilising for protein stability (DUET)	A134D	Others	0.007	2.43	1.00	3.05	−2.98	0.58
Most stabilising for protein stability (DUET)	Y103D	Others	0.22	142.33	1.24E-21	5.42	1.18	0.85
Closest stabilising for protein stability (DUET)	C138R	Drug-associated	0.09	116.96	6.09E-10	3.28	0.10	0.35
Most destabilising for ligand affinity	D49G	Drug-associated	0.05	29.16	0.0001	3.45	−1.16	−3.46
Closest destabilising for ligand affinity	D8G	Drug-associated	0.08	48.69	1.95E-07	3.22	−0.85	−3.45
Most stabilising for ligand affinity	G162D	Drug-associated	0.03	38.85	0.0008	8.32	−1.04	2.23
Closest stabilising for ligand affinity	A134P	Others	0.007	9.70	1.71E-01	3.05	−1.43	0.08
Most destabilising for protein stability (Foldx)	C72W	Drug-associated	0.01	19.41	0.03	7.05	27.46	–
Most stabilising for protein stability (Foldx)	A134P	Others	0.007	9.70	1.71E-01	3.05	−5.2	–

Mutations (nsSNPs) with extreme effects on odds ratio, frequency, thermodynamic stability, and ligand affinity. For ligand affinity, only mutations lying within 10 Å of PZA (pyrazinamide) were considered. nsSNPs, non-synonymous Single Nucleotide Polymorphisms; Å, Angstroms; MAF, minor allele frequency; OR, Odds Ratio; ΔΔG, Change in Gibbs free energy in Kcal/mol.

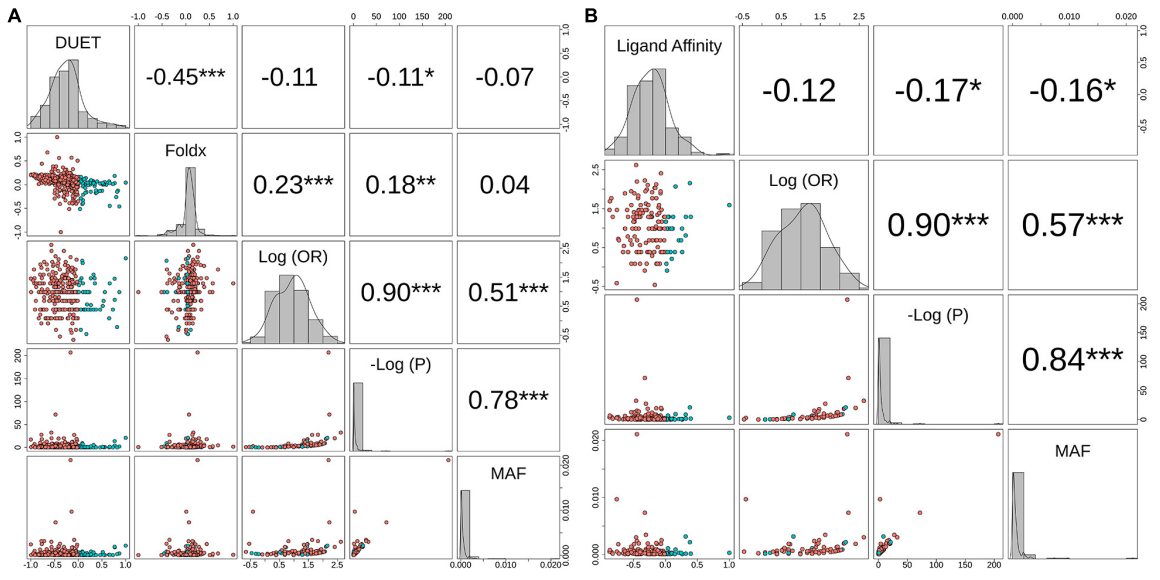


FIGURE 6 | Correlation between biophysical effects and GWAS measures of Odds Ratio (OR), *P*-values (P) and minor allele frequency (MAF). Pairwise correlations between MAF, negative log10 *P*-value [-Log(P)], Log10 (OR) and **(A)** Protein stability (DUET) and FoldX for 424 nsSNPs, **(B)** Ligand affinity of 201 nsSNPs (lying within 10 Å of PZA). The upper panel in both plots include the pairwise Spearman correlation values along with their statistical significance (*$P < 0.05$, **$P < 0.01$, ***$P < 0.001$). The points in the lower panel represent nsSNPs, coloured according to respective stability effects: **(A)** nsSNPs with destabilising effect for DUET and ligand affinity are coloured red, while for FoldX these appear in blue, **(B)** nsSNPs with stabilising effect for DUET and ligand affinity appear in blue, while for FoldX these appear in red. The diagonal plots display the histogram of the corresponding parameter. The figure is generated using R statistical software (version 4.0.2). nsSNPs, non-synonymous Single Nucleotide Polymorphisms; PZA, pyrazinamide; Units for DUET, FoldX and Ligand Affinity (Kcal/mol).

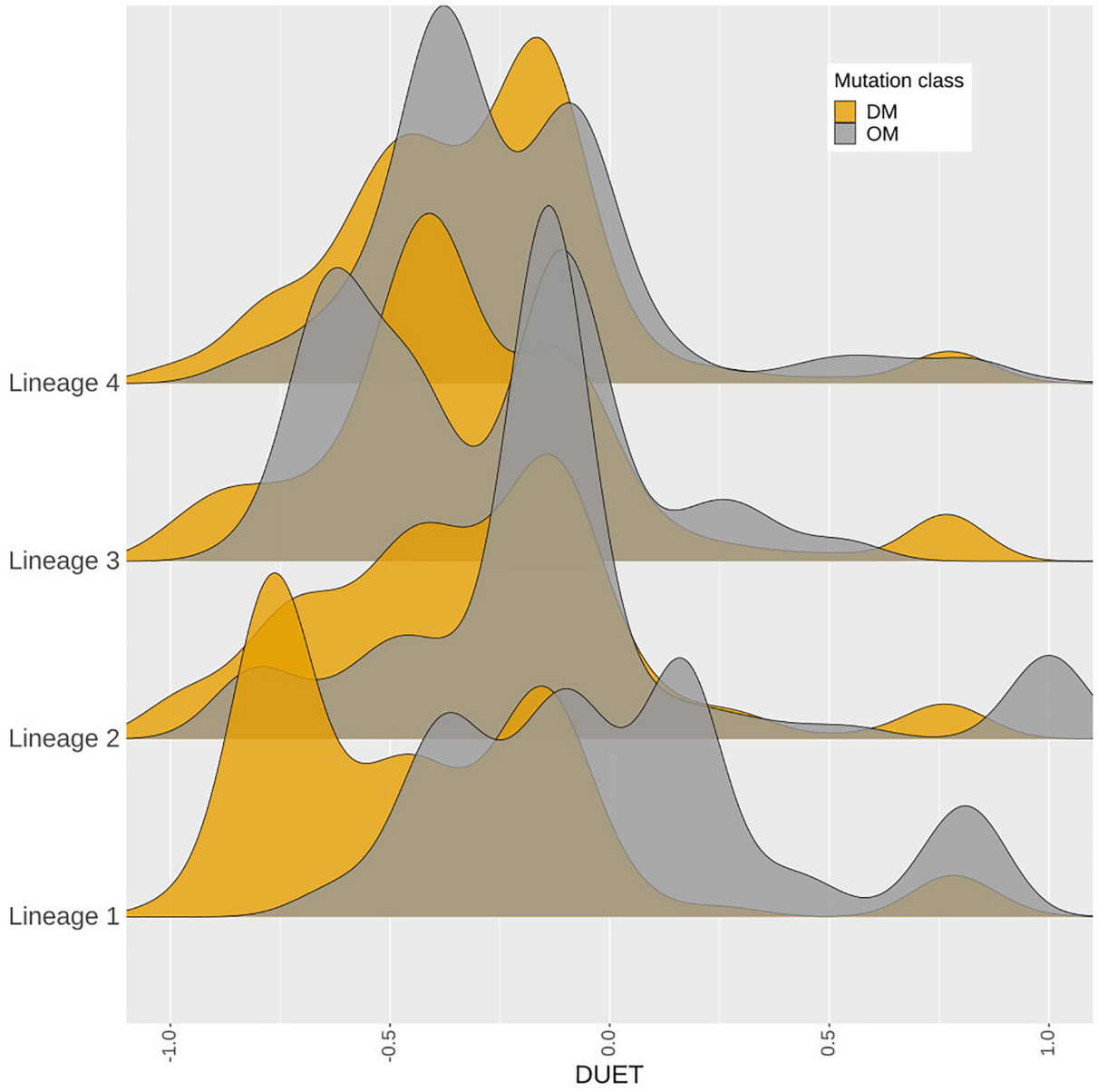


FIGURE 7 | Density distribution of *M. tuberculosis* lineages. A total of 4,433 samples belonging to Lineages 1–4, containing 419'pncA mutations were considered. The horizontal axis shows the DUET stability values (−1, most destabilising) to blue (+1, most stabilising). while the vertical axis shows the density distribution of *M. tuberculosis* lineages coloured by mutation class as either DM (associated with pyrazinamide resistance in orange) or OM (not associated with pyrazinamide drug resistance which appear in grey). DM mutations comprise of a total of 3,565 samples contributing to 226 mutations, while 868 samples contributing to 193 mutations formed part of the OM mutation class. The figure is generated using R statistical software (version 4.0.2). Abbreviations used: nsSNPs: non-synonymous Single Nucleotide Polymorphisms, pncA: pyrazinamidase.

analysed (**Supplementary Figure 9**). The majority of nsSNPs were linked to increased likelihood of being resistant to PZA (OR > 1). For unadjusted ORs, this was 96% (310/322), while for aOR, it was ∼75% (122/163). Wild type position 51 had the highest unadjusted OR (> 350, $P < 10^{-30}$), followed by positions 57, 120 (OR > 250, $P < 10^{-19}$), and subsequently by positions 10, 103, 68, 135, 138, 96, and 180 (OR > 100; $P < 10^{-10}$) (**Figure 1A**, **Supplementary Figure 8**, and **Supplementary Table 1**), with most of these positions being present in the metal binding and active sites.

When assessing sites in relation to mutational diversity, active site residues were among the highest, with residues H51, H57, H71, K96 associated with six distinct mutations, followed by F13, D49, W68, A134, C138 with five mutation

each, while residues D8, Y103, H137 were associated with four distinct mutations and residues I133 associated with two distinct mutations (**Figure 1B**). The dominant effect of a highly frequent mutation (Q10P; MAF = 2.1%, *OR* = 156.23) in the population compared to two other mutations observed at the same position namely Q10R (MAF = 0.13%, *OR* = 83.01) and Q10H (MAF = 0.08%, *OR* = 107.17) (**Supplementary Table 1**), makes position 10 prominent in terms of MAF (**Supplementary Figure 8**) while sites involved in the catalytic activity and iron metal centre are more prominent with respect to SNP diversity (**Supplementary Figure 8**). These results suggest that mutations at these structurally and functionally important sites are likely under selective pressure exerted by the drug resulting in this observed mutational diversity.

The relationship between structural measures of stability and OR was visualised as a bubble plot indicating that mutations associated with greater resistance (high OR) tend not to have extreme effects (**Supplementary Figure 10**). Furthermore, this relationship along with MAF, OR, and *P*-values was assessed through Spearman correlations (**Figures 6A,B**). MAF was strongly correlated with *P*-values for all 424 mutations ($\rho = 0.78$, $P < 0.001$) and 201 mutations lying with 10 Å of PZA ($\rho = 0.84$, $P < 0.001$) (**Figures 6A,B**). As expected, OR and *P*-values were strongly correlated ($\rho = 0.9$, $P < 0.001$) for all 424 nsSNPs and 201 nsSNPs close to PZA binding site (**Figures 6A,B**). FoldX stability and DUET stability values showed moderate correlation ($\rho = 0.45$, $P < 0.001$). The negative sign for the DUET and FoldX associations is expected since stability changes measured by these tools have opposite signs (i.e., $\Delta\Delta G < 0$: destabilising in DUET vs. stabilising in FoldX). FoldX $\Delta\Delta G$ values showed weak but significant correlations with OR ($\rho = 0.23$, $P < 0.001$), and *P*-values ($\rho = 0.18$, $P < 0.01$) (**Figure 6A**), while DUET $\Delta\Delta G$ and ligand affinity showed weak and insignificant association with OR ($\rho = -0.1$, $P > 0.05$) (**Figures 1B**, **6A**), including adjusted OR (**Supplementary Figures 9A**, **8B**).

When considering aOR and its relationship with stability and other structural features [i.e., Accessible (ASA), Relative Surface Area (RSA), residue depth (RD), and hydrophobicity values (KD)], there was high correlation ($\rho > 0.6$, $P < 0.05$) with adjusted and unadjusted ORs (**Supplementary Figure 9A**). DUET $\Delta\Delta G$ showed moderate positive correlation between ASA and RSA ($\rho > 0.6$, $P < 0.05$), while moderately negative correlation with RD ($\rho \sim -0.5$, $P < 0.05$), and weak negative correlation with KD values ($\rho \sim -0.2$, $P < 0.05$) (**Supplementary Figure 9A**). The same structural features, however, did not demonstrate correlation with either FoldX $\Delta\Delta G$ (**Supplementary Figure 9A**) or ligand affinity (**Supplementary Figure 9B**).

Structural Differences in Drug Associated Mutations

Comparing stability effect (DUET and FoldX), ligand affinity, ligand distance, and other structural features (ASA, RSA, RD, KD) between mutations associated with PZA drug resistance (DM) and other mutations (OM), revealed statistically significant differences ($P < 0.05$) between all features except hydrophobicity values. The difference in structural features were most prominent when all 424 SNP mutations were considered ($P < 0.0001$) (**Figures 4A,B**) with lesser significance for ligand affinity ($P < 0.05$), ASA ($P < 0.01$), and RSA and RD ($P < 0.001$) values when 201 nsSNPs lying within 10 Å were considered (**Figure 4C**). Mutations associated with PZA resistance have lower DUET (**Figure 4A**, top left) but higher FoldX stability changes (**Figure 4B**, bottom left), and lower binding affinity (**Figure 4C**, second from bottom left) compared to OM. Additionally, it also appears that that while drug mutations need not necessarily occur at the hydrophobic sites (KD values, $P > 0.05$), they tend to lie buried indicated by higher RD values, and consequently lower surface area (ASA and RSA) compared to OM (**Figures 4A,B**).

Distinct Stability Profile for Drug Mutations and Lineage 1

A total of 419 nsSNPs are lineage specific (L1: 74; L2: 277; L3: 104; L4: 311). The greatest diversity of nsSNPs was observed in L3 (54.7%), followed by L1 (51.4%) and Lineage 2 (14.7%) with L4 showing the lowest diversity (14.1%) despite containing the highest number of samples (**Supplementary Figure 6**). Statistical analysis of the DUET $\Delta\Delta G$ distributions revealed significant differences between all lineages except between L3 and L4. Lineage differences for DUET $\Delta\Delta G$ were most prominent between L2 and L4 ($P < 0.0001$), followed by L1 and L4 ($P < 0.001$) (**Supplementary Table 2A**). Within each lineage, mutational distributions were significantly different between DM and OM mutation classes ($P < 0.0001$) except L3 (**Supplementary Table 2B**). Interestingly, a distinct stability profile was observed for DM mutations within L1. Mutations associated with drug resistance showed a marked peak around the extreme end (-0.75 DUET $\Delta\Delta G$) of the destabilising spectrum (**Figure 7**) within L1.

DISCUSSION

Genetic mutations including nsSNPs present within drug-targets and their activating genes are the main drivers of resistance development in TB (Schön et al., 2017). The motivation for investigating the missense mutations within the protein coding region only of the *pncA* gene was to enable understanding of the phenotypic mutational effects in relation to PZA resistance development. While the exact molecular mechanisms of PZA resistance are yet to be fully elucidated, the binding pocket of PZA and its key interactions are well known and characterised (Petrella et al., 2011; Ali et al., 2020; Sheik Amamuddy et al., 2020; Khan et al., 2021). This knowledge was used to guide the molecular docking of PZA to generate the pncA-PZA complex in the absence of an experimentally solved structure of the bound complex in Mtb. While docking generates a variety of ligand conformations (poses), choosing the "best" pose is based on considerations around key molecular interactions formed by the ligand, interaction energy of the docked complex and subject expertise. Using these guides, docking pose 1 was chosen due to its molecular interactions

with known key residues and close alignment with previously published studies (Karmakar et al., 2018; Ali et al., 2020; Khan et al., 2021). In addition, we analysed the top two docking poses using the mCSM pipeline (**Supplementary Figure 3**). The resulting mutational effects on pncA stability and ligand affinity did not differ between poses indicating the small differences in pose did not affect downstream analysis. It also suggests that due to the small size of the PZA molecule, the orientation of the aromatic ring within the cavity may have more flexibility in its orientation and interaction with the neighbouring residues, but without drastically impacting the molecular interactions for global protomer stability and ligand binding affinity.

The molecular motion of pncA assessed by NMA was visualised to understand the mutational effects with regard to flexibility (**Supplementary Figure 1**). Sites displaying high mutational frequency or association with drug resistance mutations were not located in regions with high flexibility, with large molecular motions mainly restricted to the loop region 60–66. This suggests the molecular motion in pncA does not interfere with PZA binding as active site residues were not associated with high fluctuations.

Normal mode analysis shows large scale molecular motions. Molecular dynamics (MD) studies offer insights into the finer grained atomic motions and are an excellent way to investigate molecular mechanisms. However, these studies are computationally intensive and are difficult to scale for studying hundreds of mutations. A recent MD study on a subset of mutations found within our dataset analysed seven pncA nsSNPs (F94L, F94S, K96N, K96R, G97C, G97D, and G97S) showed that these destabilising mutations altered the binding pocket, allowing increased PZA flexibility (Khan et al., 2021). All seven mutations were associated with PZA resistance and also showed destabilising effects in our study. A similar study of destabilising mutations R123P, T76P, H7R associated with PZA resistance showed that the mechanism of resistance could be through increasing the flexibility of the region they are located in, thereby changing the binding pocket volume (Ali et al., 2020). Another MD study of mutations P54L and H57P showed that they decrease overall stability along with reduced ligand affinity leading to PZA resistance (Mehmood et al., 2019). All of these observations are concordant with our analysis.

Destabilising effects of nsSNPs are thought to be the main reason for impeding protein function through directly effecting protomer stability or ligand affinity. However, large stabilising effects can have an equally deleterious impact on protein function through rigidification, impeding flexibility and dynamic molecular motions. This has been implicated more generally within a disease context (Gerasimavicius et al., 2020) and more specifically in PZA resistance (Rajendran and Sethumadhavan, 2014). It offers an explanation for the observance of the stabilising mutation site 103. Drug associated mutations at this site (Y103C, Y103H, and Y103S) could result from the rigidification of the binding pocket leading to reduced binding affinity measured as destabilising PZA affinity.

Mutations within *pnca* are scattered along the entire gene length observed in studies (Stoffels et al., 2012; Miotto et al., 2014; Whitfield et al., 2015). While two other genes, *rpsA* and *panD* have also been linked to PZA resistance, a clear link between *rpsA* and PZA resistance is lacking (Shi et al., 2011; Alexander et al., 2012; Simons et al., 2013; Tan et al., 2014) although there is increasing evidence to support *panDs* association with PZA resistance (Pandey et al., 2016; Werngren et al., 2017; Gopal et al., 2020). In our analysis, there were only a few samples with *rpsA* and *panD* mutations, therefore limiting attempts at assessing their synergistic relationship with PZA resistance. Mutations within the *pncA* gene and its promoter remain the most common route to PZA resistance (Dookie et al., 2018) (Khan et al., 2019). Nearly 70% of the MDR isolates and 13% XDR isolates had nsSNPs in the *pncA* coding region. The burden of pncA mutations in the MDR and XDR isolates was lower in our analysis compared to 88.0% and ∼20% observed by Pang et al. (2017). In another study, 70% of the MDR isolates, and significantly higher i.e., 96% of XDR isolates harboured pncA mutations including nsSNPs (Allana et al., 2017). An alternative route to resistance for pncA as a non-essential gene encoding an enzyme that transforms a prodrug to drug would be by INDELs or mutations leading to premature stop codons resulting in the protein being degraded on translation. A recent report analysing the *pncAc.85_86insG* frameshift mutation using structural and biophysical analysis showed the mutation resulted in a truncated and incomplete protein lacking the active site pocket (Karmakar et al., 2018). Despite this obvious route to resistance, only 1% samples in our dataset showed INDELs and stop codons, compared to 13% of samples that showed missense point mutations in pncA. This is consistent with the knowledge that nsSNPs in pncA remain the major route to resistance for PZA (Khan et al., 2019).

Destabilising effects are considered detrimental to the downstream protein function (via disruption of drug affinity, nucleic acid affinity or overall complex stability) and are thus given higher consideration in classifying mutations (Wylie and Shakhnovich, 2011). In our analysis, around 85% of mutations were destabilising for overall protein stability as well as complex affinity. It is thought that the resistant phenotype is imparted either through affecting protein folding, instability of the PZase protein, prevention of coenzyme complex (Gopal et al., 2016) or loss of virulence factor synthesis (Gopal et al., 2016). Further, this is thought to come without a high bacterial fitness cost since pncA is primarily an activator of the PZA drug. This is similar to a recent observation reported in the *katG* gene (target for the anti-TB pro-drug, isoniazid) with a high proportion of destabilising mutations (Portelli et al., 2018). Also, a higher proportion 60% (n = 253) of SNP mutations showed electrostatic changes compared to ∼35% reported by Portelli et al. (2018). This likely due to the larger sample size of our dataset.

All active site residues appear to be under drug selection pressures due to multiple mutations (>2) associated with these with the exception of I133, considered to be an emerging or budding-resistance hotspot. In our analyses, there were 22 such sites while 83 sites within *pncA a*ssociated with > 2 nsSNPs linked to PZA drug resistance (categorised as DM). However mutations

were not restricted to the active site, with less than 50% resistant variants lying within 10 Å of the active site of PZA, indicating the possible role of distal residues in resistance development (Portelli et al., 2018). Mutations associated with drug resistance tend to have lower stability, lie buried within the structure with lesser surface area as shown by Karmakar et al. (2020).

Our study compares results from two different computational stability predictors: mCSM and FoldX (Schymkowitz et al., 2005). Unsurprisingly, most mutations were found to have a destabilising effect (**Supplementary Figure 11**). FoldX reported ∼85% (vs. ∼80% estimated by DUET) nsSNPs with destabilising effect. The range for absolute ΔΔG values was greater for FoldX (median: 2.0; range: −5.2, 27.46) compared to DUET (median: −0.1; range: −3.9, 1.2). There was however, 77% agreement between FoldX and DUET outcomes (data not shown). Interestingly, drug associated mutations displayed higher FoldX ΔΔG predictions compared to mCSM-DUET ΔΔG predictions. A possible explanation for this is the differences in the underlying parameters the different methods use. FoldX constructs mutant structures by mutating the target residue and searching for the optimal conformation by iteratively altering the position of the neighbouring side chains. The stability of the mutant structure is estimated using an empirical force field made of several energy terms. This compares to DUET where estimates of the structural effects are based on differences between the wild-type environment and pharmacophore atomic changes resulting from the mutation, without the need to generate mutant structures. With this in mind, it appears that the DM mutations have larger local perturbations in the mutated region considered by FoldX, resulting in higher ΔΔG predictions compared to the lesser effects of surface area considered by DUET. Drug resistance mutations displaying smaller surface area compared to their susceptible counterparts were also observed in recent studies investigating nsSNPs in Mtb genes (Portelli et al., 2018; Karmakar et al., 2020) indicating the role of compensatory mutations, alleviating any fitness penalty in the development of the drug resistance phenotype. The extent of the contribution of surface area in these methods is reflected in the observation of moderate correlations between DUET and structural features, and the weaker associations between FoldX and structural features (**Supplementary Figure 9A**). Structural associations for ligand affinity were also observed to be weak (**Supplementary Figure 9B**) most likely due to the role of factors involved in short-range interactions (like Van der Waal's forces) not considered in our analysis. A similar view emerged in the recent study by Karmakar et al. (2020) where no significant differences were observed for PZA binding affinity.

It has been suggested that frequently occurring mutations may not confer extreme changes in biophysical stability measures, with mild stability effects offering local fitness advantages (Portelli et al., 2018). Our data presented us with the opportunity to test this theory empirically by assessing relationships of stability with GWAS measures of MAF, OR, and *P*-values. At a glance, it appears that mutations with high OR tend be less extreme in their impact on protein stability and ligand affinity (**Supplementary Figure 10**). However, we did not find any significant association with high frequency mutations and extreme changes in stability or affinity parameters (**Figure 6**). One possible explanation is that the fitness landscape is gene and function specific, optimised differently for genes directly coding for drug targets and for non-essential genes like *pncA*. Another major consideration is that resistance is often acquired through a stepwise ordinal accumulation of mutations (Woodford and Ellington, 2007; Ismail et al., 2019). The genetic background can dramatically influence fitness effects associated with mutations (Wong, 2017). Consequently, the mutational impact differs when occurring against a sequence background of extant resistant mutations, a phenomenon known as epistasis (Wong, 2017). Since resistance development is a balanced interplay between fitness effects and cost of resistance, epistasis warrants due consideration in efforts to understand and limit the evolution of multi-drug resistance.

The use of mCSM suite of tools has the advantage of studying global (protein stability) as well as local effects (ligand affinity, protein-protein interaction, and protein nucleic-acid interaction). Additionally, it also provides the methodological consistency for comparing molecular effects and benefits application of machine learning methods (ML) to explore greater mechanistic details. While computationally intensive, ML methods would benefit from using tools such as DynaMut (Rodrigues et al., 2018) which account for protein molecular motions when estimating mutational effect on protein stability. Additionally methods which consider anti-symmetric properties of mutational impact i.e., $\Delta\Delta G\ (A \rightarrow B) = -\Delta\Delta G\ (B \rightarrow A)$ like DeepDDG (Cao et al., 2019) and INPS-MD (Savojardo et al., 2016) have the potential to build robust predictive models and improve the "learning" capability of ML methods in the context of machine learning.

Mtb lineages have been associated with virulence, disease transmission, drug resistance, and clinical outcome (Ford et al., 2013; Reiling et al., 2013; Novais et al., 2017; Correa-Macedo et al., 2019; Oppong et al., 2019; McHenry et al., 2020). Lineage specific differences between lineages 2 and 4 have recently been noted in the development of TB drug resistance, especially related to MDR and XDR strains (Oppong et al., 2019). Our study highlighted the most significant differences between L2 and L4 with respect to protomer stability demonstrating the biophysical phenotypic manifestation of these underlying genotypic changes. The observance of a distinct peak for destabilising mutations related to drug resistance within L1 suggests that the extreme mutational consequences of such mutations in the "ancient" lineage 1 may be rapidly giving way to other "modern" *M. tuberculosis* lineages linked to MDR and XDR-TB and virulence.

Our study is based on a well-characterised clinical dataset sourced globally from over 35 K clinical isolates, and leverages the availability of robust metadata (lineage, geography, DST, etc.) for each isolate. We show that the framework used in our work allows us to investigate the interrelationships between genomic features from GWAS analysis and the biophysical measures of nsSNPs, helping to contextualise the underlying bacterial fitness and mutational landscape. The need to consider multiple stability predictors with different underlying principles to validate these associations has also been highlighted. Lineage associations of drug resistance, and their biophysical consequences, require

further investigation and the functional characteristics of mutations should be validated in future experiments. We hope such a framework can be used to understand and inform therapeutic and stewardship efforts.

DATA AVAILABILITY STATEMENT

All pre-generated TB-profiler results were downloaded for all isolates from tbdr.lshtm.ac.uk/sra.

AUTHOR CONTRIBUTIONS

TT was responsible for the molecular docking, integrating genomics and structural data, data analysis, and writing the initial draft. JP made available and generated the genomics data results. CE provided the FoldX pipeline. TC and NF provided the overall supervision and contributed to revising and refining the manuscript. All authors contributed to the manuscript and approved the submitted version.

FUNDING

TT was supported by a BBSRC Ph.D. studentship (Grant No: BB/S507544/1). JP was funded by a Newton Institutional Links Grant (British Council. 261868591). CE was funded by the Medical Research Council UK (Grant No. MR/T000171/1). TC was funded by the Medical Research Council UK (Grant No. MR/M01360X/1, MR/N010469/1, MR/R025576/1, and MR/R020973/1) and BBSRC (Grant No. BB/R013063/1).

SUPPLEMENTARY MATERIAL

The Supplementary Material for this article can be found online at: https://www.frontiersin.org/articles/10.3389/fmolb.2021.619403/full#supplementary-material

REFERENCES

Alexander, D. C., Ma, J. H., Guthrie, J. L., Blair, J., Chedore, P., and Jamieson, F. B. (2012). Gene sequencing for routine verification of pyrazinamide resistance in *Mycobacterium tuberculosis*: a role for pncA but Not rpsA. *J. Clin. Microbiol.* 50, 3726–3728. doi: 10.1128/jcm.00620-12

Ali, A., Khan, M. T., Khan, A., Ali, S., Chinnasamy, S., Akhtar, K., et al. (2020). Pyrazinamide resistance of novel mutations in pncA and their dynamic behavior. *RSC Adv.* 10, 35565–35573. doi: 10.1039/d0ra06072k

Allana, S., Shashkina, E., Mathema, B., Bablishvili, N., Tukvadze, N., Shah, N. S., et al. (2017). PncA gene mutations associated with pyrazinamide resistance in drug-resistant Tuberculosis, South Africa and Georgia. *Emerg. Infect. Dis.* 23, 491–495. doi: 10.3201/eid2303.161034

Al-Saeedi, M., and Al-Hajoj, S. (2017). Diversity and evolution of drug resistance mechanisms in *Mycobacterium tuberculosis*. *Infect. Drug Resist.* 10, 333–342. doi: 10.2147/idr.s144446

Artimo, P., Jonnalagedda, M., Arnold, K., Baratin, D., Csardi, G., de Castro, E., et al. (2012). ExPASy: SIB bioinformatics resource portal. *Nucleic Acids Res.* 40, W597–W603.

Berman, H. M., Westbrook, J., Feng, Z., Gilliland, G., Bhat, T. N., Weissig, H., et al. (2000). The protein data bank. *Nucleic Acids Res.* 28, 235–242.

Boonaiam, S., Chaiprasert, A., Prammananan, T., and Leechawengwongs, M. (2010). Genotypic analysis of genes associated with isoniazid and ethionamide resistance in MDR-TB isolates from Thailand. *Clin. Microbiol. Infect.* 16, 396–399. doi: 10.1111/j.1469-0691.2009.02838.x

Cao, H., Wang, J., He, L., Qi, Y., and Zhang, J. Z. (2019). DeepDDG: predicting the stability change of protein point mutations using neural networks. *J. Chem. Inf. Model.* 59, 1508–1514. doi: 10.1021/acs.jcim.8b00697

Chakravarty, S., and Varadarajan, R. (1999). Residue depth: a novel parameter for the analysis of protein structure and stability. *Structure* 7, 723–732. doi: 10.1016/s0969-2126(99)80097-5

Coll, F., McNerney, R., Guerra-Assunção, J. A., Glynn, J. R., Perdigão, J., Viveiros, M., et al. (2014). A robust SNP barcode for typing *Mycobacterium tuberculosis* complex strains. *Nat. Commun.* 5:4812.

Coll, F., Phelan, J., Hill-Cawthorne, G. A., Nair, M. B., Mallard, K., Ali, S., et al. (2018). Genome-wide analysis of multi- and extensively drug-resistant *Mycobacterium tuberculosis*. *Nat. Genet.* 50, 307–316.

Comas, I., Borrell, S., Roetzer, A., Rose, G., Malla, B., Kato-Maeda, M., et al. (2011). Whole-genome sequencing of rifampicin-resistant Mycobacterium tuberculosis strains identifies compensatory mutations in RNA polymerase genes. *Nat. Genet.* 44, 106–110. doi: 10.1038/ng.1038

Correa-Macedo, W., Cambri, G., and Schurr, E. (2019). The interplay of human and *Mycobacterium Tuberculosis* genomic variability. *Front. Genet.* 10:865.

de Vos, M., Müller, B., Borrell, S., Black, P. A., van Helden, P. D., Warren, R. M., et al. (2013). Putative compensatory mutations in the rpoC gene of rifampin-resistant *Mycobacterium tuberculosis* are associated with ongoing transmission. *Antimicrob. Agents Chemother.* 57, 827–832. doi: 10.1128/aac.01541-12

Dookie, N., Rambaran, S., Padayatchi, N., Mahomed, S., and Naidoo, K. (2018). Evolution of drug resistance in *Mycobacterium tuberculosis*: a review on the molecular determinants of resistance and implications for personalized care. *J. Antimicrob. Chemother.* 73, 1138–1151. doi: 10.1093/jac/dkx506

Ford, C. B., Shah, R. R., Maeda, M. K., Gagneux, S., Murray, M. B., Cohen, T., et al. (2013). Mycobacterium tuberculosis mutation rate estimates from different lineages predict substantial differences in the emergence of drug-resistant tuberculosis. *Nat. Genet.* 45, 784–790. doi: 10.1038/ng.2656

Gerasimavicius, L., Liu, X., and Marsh, J. A. (2020). Identification of pathogenic missense mutations using protein stability predictors. *Sci. Rep.* 10:15387.

Gopal, P., Sarathy, J. P., Yee, M., Ragunathan, P., Shin, J., Bhushan, S., et al. (2020). Pyrazinamide triggers degradation of its target aspartate decarboxylase. *Nat. Commun.* 11:1661.

Gopal, P., Yee, M., Sarathy, J., Liang Low, J., Sarathy, J. P., Kaya, F., et al. (2016). Pyrazinamide resistance is caused by two distinct mechanisms: prevention of coenzyme a depletion and loss of virulence factor synthesis. *ACS Infect. Dis.* 2, 616–626. doi: 10.1021/acsinfecdis.6b00070

Ismail, N., Ismail, N. A., Omar, S. V., and Peters, R. P. H. (2019). In vitro study of stepwise acquisition of rv0678 and atpE mutations conferring bedaquiline resistance. *Antimicrob. Agents Chemother.* 63:e00292-19.

Jubb, H. C., Higueruelo, A. P., Ochoa-Montaño, B., Pitt, W. R., Ascher, D. B., and Blundell, T. L. (2017). Arpeggio: a web server for calculating and visualising interatomic interactions in protein structures. *J. Mol. Biol.* 429, 365–371. doi: 10.1016/j.jmb.2016.12.004

Kabsch, W., and Sander, C. (1983). Dictionary of protein secondary structure: pattern recognition of hydrogen-bonded and geometrical features. *Biopolymers* 22, 2577–2637. doi: 10.1002/bip.360221211

Karmakar, M., Globan, M., Fyfe, J. A. M., Stinear, T. P., Johnson, P. D. R., Holmes, N. E., et al. (2018). Analysis of a Novel pncA mutation for susceptibility to pyrazinamide therapy. *Am. J. Respir. Crit. Care Med.* 198, 541–544. doi: 10.1164/rccm.201712-2572le

Karmakar, M., Rodrigues, C. H. M., Horan, K., Denholm, J. T., and Ascher, D. B. (2020). Structure guided prediction of Pyrazinamide resistance mutations in pncA. *Sci. Rep*. 10:1875.

Kavvas, E. S., Catoiu, E., Mih, N., Yurkovich, J. T., Seif, Y., Dillon, N., et al. (2018). Machine learning and structural analysis of *Mycobacterium tuberculosis* pan-genome identifies genetic signatures of antibiotic resistance. *Nat. Commun.* 9:4306.

Khan, M. T., Malik, S. I., Ali, S., Masood, N., Nadeem, T., Khan, A. S., et al. (2019). Pyrazinamide resistance and mutations in pncA among isolates of *Mycobacterium tuberculosis* from Khyber Pakhtunkhwa, Pakistan. *BMC Infect. Dis*. 19:116.

Khan, T., Khan, A., Ali, S. S., Ali, S., and Wei, D. Q. (2021). A computational perspective on the dynamic behaviour of recurrent drug resistance mutations in the pncA gene from: *Mycobacterium tuberculosis*. *RSC Adv*. 11, 2476–2486. doi: 10.1039/d0ra09326b

McHenry, M. L., Bartlett, J., Igo, R. P., Wampande, E. M., Benchek, P., Mayanja-Kizza, H., et al. (2020). Interaction between host genes and *Mycobacterium tuberculosis* lineage can affect tuberculosis severity: evidence for coevolution? *PLoS Genet.* 16:e1008728. doi: 10.1371/journal.pgen.1008728

Mehmood, A., Khan, M. T., Kaushik, A. C., Khan, A. S., Irfan, M., and Wei, D.-Q. (2019). Structural dynamics behind clinical mutants of PncA-Asp12Ala, Pro54Leu, and His57Pro of *Mycobacterium tuberculosis* associated with Pyrazinamide resistance. *Front. Bioeng. Biotechnol.* 7:404.

Miotto, P., Cabibbe, A. M., Feuerriegel, S., Casali, N., Drobniewski, F., Rodionova, Y., et al. (2014). Mycobacterium tuberculosis pyrazinamide resistance determinants: a multicenter study. *MBio* 5:e001819-14.

Morris, G. M., Goodsell, D. S., Halliday, R. S., Huey, R., Hart, W. E., Belew, R. K., et al. (1998). Automated docking using a lamarckian genetic algorithm and an empirical binding free energy function. *J. Comput. Chem*. 19, 1639–1662.

Morris, G. M., Huey, R., Lindstrom, W., Sanner, M. F., Belew, R. K., Goodsell, D. S., et al. (2009). AutoDock4 and AutoDockTools4: automated docking with selective receptor flexibility. *J. Comput. Chem.* 30, 2785–2791. doi: 10.1002/jcc.21256

Napier, G., Campino, S., Merid, Y., Abebe, M., Woldeamanuel, Y., Aseffa, A., et al. (2020). Robust barcoding and identification of *Mycobacterium tuberculosis* lineages for epidemiological and clinical studies. *Genome Med.* 12:114.

Novais, E., Bastos, H., Machado, H., Sousa, J., Veiga, M. I., Ramos, A., et al. (2017). Tuberculosis severity and its association with pathogen phylogeny and properties. *Eur. Respir. J.* 50:A3046.

Oppong, Y. E. A., Phelan, J., Perdigão, J., Machado, D., Miranda, A., Portugal, I., et al. (2019). Genome-wide analysis of *Mycobacterium tuberculosis* polymorphisms reveals lineage-specific associations with drug resistance. *BMC Genomics* 20:252.

Pandey, B., Grover, S., Tyagi, C., Goyal, S., Jamal, S., Singh, A., et al. (2016). Molecular principles behind pyrazinamide resistance due to mutations in panD gene in *Mycobacterium tuberculosis*. *Gene* 581, 31–42. doi: 10.1016/j.gene.2016.01.024

Pandurangan, A. P., Ochoa-Montaño, B., Ascher, D. B., and Blundell, T. L. (2017). SDM: a server for predicting effects of mutations on protein stability. *Nucleic Acids Res.* 45, W229–W235.

Pang, Y., Zhu, D., Zheng, H., Shen, J., Hu, Y., Liu, J., et al. (2017). Prevalence and molecular characterization of pyrazinamide resistance among multidrug-resistant *Mycobacterium tuberculosis* isolates from Southern China. *BMC Infect. Dis.* 17:711.

Petrella, S., Gelus-Zietal, N., Maudry, A., Laurans, C., Boudjelloul, R., and Sougakoff, W. (2011). 3PL1: crystal structure of the pyrazinamidase of mycobacterium tuberculosis: insights into natural and acquired resistance to pyrazinamide. *PLoS One* 6:e15785. doi: 10.1371/journal.pone.0015785

Pettersen, E. F., Goddard, T. D., Huang, C. C., Couch, G. S., Greenblatt, D. M., Meng, E. C., et al. (2004). UCSF chimera?a visualization system for exploratory research and analysis. *J. Comput. Chem.* 25, 1605–1612. doi: 10.1002/jcc.20084

Phelan, J., Coll, F., McNerney, R., Ascher, D. B., Pires, D. E. V., Furnham, N., et al. (2016). Mycobacterium tuberculosis whole genome sequencing and protein structure modelling provides insights into anti-tuberculosis drug resistance. *BMC Med.* 14:31.

Phelan, J., Lim, D. R., Mitarai, S., de Sessions, P. F., Tujan, M. A. A., Reyes, L. T., et al. (2019a). Mycobacterium tuberculosis whole genome sequencing provides insights into the Manila strain and drug-resistance mutations in the Philippines. *Sci. Rep*. 9:9305.

Phelan, J., O'Sullivan, D. M., Machado, D., Ramos, J., Oppong, Y. E. A., Campino, S., et al. (2019b). Integrating informatics tools and portable sequencing technology for rapid detection of resistance to anti-tuberculous drugs. *Genome Med.* 11:41.

Pires, D., and Ascher, D. B. (2016). mCSM-AB: a web server for predicting antibody-antigen affinity changes upon mutation with graph-based signatures. *Nucleic Acids Res.* 44, W469–W473.

Pires, D., and Ascher, D. B. (2017). mCSM–NA: predicting the effects of mutations on protein–nucleic acids interactions. *Nucleic Acids Res.* 45, W241–W246.

Pires, D., Ascher, D. B., and Blundell, T. L. (2014a). DUET: a server for predicting effects of mutations on protein stability using an integrated computational approach. *Nucleic Acids Res.* 42, W314–W319.

Pires, D., Ascher, D. B., and Blundell, T. L. (2014b). mCSM: predicting the effects of mutations in proteins using graph-based signatures. *Bioinformatics* 30, 335–342. doi: 10.1093/bioinformatics/btt691

Pires, D., Blundell, T. L., and Ascher, D. B. (2016). mCSM-lig: quantifying the effects of mutations on protein-small molecule affinity in genetic disease and emergence of drug resistance. *Sci. Rep*. 6:29575.

Portelli, S., Phelan, J. E., Ascher, D. B., Clark, T. G., and Furnham, N. (2018). Understanding molecular consequences of putative drug resistant mutations in *Mycobacterium tuberculosis*. *Sci. Rep*. 8:15356.

R Core Team (2014). *R: a Language and Environment for Statistical Computing*. Vienna: R Development Core Team.

Rajendran, V., and Sethumadhavan, R. (2014). Drug resistance mechanism of PncA in *Mycobacterium tuberculosis*. *J. Biomol. Struct. Dyn.* 32, 209–221. doi: 10.1080/07391102.2012.759885

Reiling, N., Homolka, S., Walter, K., Brandenburg, J., Niwinski, L., Ernst, M., et al. (2013). Clade-specific virulence patterns of Mycobacterium tuberculosis complex strains in human primary macrophages and aerogenically infected mice. *MBio* 4:e00250-13.

Rodrigues, C. H. M., Myung, Y., Pires, D. E. V., and Ascher, D. B. (2019). mCSM-PPI2: predicting the effects of mutations on protein–protein interactions. *Nucleic Acids Res.* 47, W338–W344.

Rodrigues, C. H. M., Pires, D. E. V., and Ascher, D. B. (2018). DynaMut: predicting the impact of mutations on protein conformation, flexibility and stability. *Nucleic Acids Res.* 46, W350–W355.

Savojardo, C., Fariselli, P., Martelli, P. L., and Casadio, R. (2016). INPS-MD: a web server to predict stability of protein variants from sequence and structure. *Bioinformatics* 32, 2542–2544. doi: 10.1093/bioinformatics/btw192

Schön, T., Miotto, P., Köser, C. U., Viveiros, M., Böttger, E., and Cambau, E. (2017). *Mycobacterium tuberculosis* drug-resistance testing: challenges, recent developments and perspectives. *Clin. Microbiol. Infect.* 23, 154–160. doi: 10.1016/j.cmi.2016.10.022

Schymkowitz, J., Borg, J., Stricher, F., Nys, R., Rousseau, F., and Serrano, L. (2005). The FoldX web server: an online force field. *Nucleic Acids Res.* 33, W382–W388.

Segala, E., Sougakoff, W., Nevejans-Chauffour, A., Jarlier, V., and Petrella, S. (2012). New mutations in the Mycobacterial ATP synthase: new insights into the binding of the diarylquinoline TMC207 to the ATP $ynthase C-Ring $tructure. *Antimicrob. Agents Chemother.* 56, 2326–2334. doi: 10.1128/aac.06154-11

Sheik Amamuddy, O., Musyoka, T. M., Boateng, R. A., Zabo, S., and Tastan Bishop, Ö (2020). Determining the unbinding events and conserved motions associated with the pyrazinamide release due to resistance mutations of *Mycobacterium tuberculosis* pyrazinamidase. *Comput. Struct. Biotechnol. J.* 18, 1103–1120. doi: 10.1016/j.csbj.2020.05.009

Shi, W., Zhang, X., Jiang, X., Yuan, H., Lee, J. S., Barry, C. E., et al. (2011). Pyrazinamide inhibits trans-translation in *Mycobacterium tuberculosis*. *Science* 333, 1630–1632. doi: 10.1126/science.1208813

Simons, S. O., Mulder, A., van Ingen, J., Boeree, M. J., and van Soolingen, D. (2013). Role of rpsA gene sequencing in diagnosis of pyrazinamide resistance: table 1. *J. Clin. Microbiol.* 51, 382–382. doi: 10.1128/jcm.02739-12

Singh, R. P., Pandey, N., Singh, A. K., Sinha, M., Kaur, P., Sharma, S., et al. (2011). Crystal structure of the complex of goat lactoperoxidase with Pyrazinamide at 2.1 A resolution. doi: 10.2210/pdb3R55/pdb

Somoskovi, A., Parsons, L. M., and Salfinger, M. (2001). The molecular basis of resistance to isoniazid, rifampin, and pyrazinamide in *Mycobacterium tuberculosis*. *Respir. Res.* 2, 164–168.

Stoffels, K., Mathys, V., Fauville-Dufaux, M., Wintjens, R., and Bifani, P. (2012). Systematic analysis of pyrazinamide-resistant spontaneous mutants and clinical isolates of *Mycobacterium tuberculosis*. *Antimicrob. Agents Chemother.* 56, 5186–5193. doi: 10.1128/aac.05385-11

Tan, Y., Hu, Z., Zhang, T., Cai, X., Kuang, H., Liu, Y., et al. (2014). Role of pncA and rpsA gene sequencing in detection of pyrazinamide resistance in *Mycobacterium tuberculosis* Isolates from Southern China. *J. Clin. Microbiol.* 52, 291–297. doi: 10.1128/jcm.01903-13

Touw, W. G., Baakman, C., Black, J., te Beek, T. A. H., Krieger, E., Joosten, R. P., et al. (2015). A series of PDB-related databanks for everyday needs. *Nucleic Acids Res.* 43, D364–D368.

Trott, O., and Olson, A. J. (2009). AutoDock vina: improving the speed and accuracy of docking with a new scoring function, efficient optimization, and multithreading. *J. Comput. Chem.* 31, 455–461.

Werngren, J., Alm, E., and Mansjö, M. (2017). Non- pncA gene-mutated but pyrazinamide-resistant *Mycobacterium tuberculosis*: why is that? *J. Clin. Microbiol.* 55, 1920–1927. doi: 10.1128/jcm.02532-16

Whitfield, M. G., Soeters, H. M., Warren, R. M., York, T., Sampson, S. L., Streicher, E. M., et al. (2015). A global perspective on pyrazinamide resistance: systematic review and meta-analysis. *PLoS One* 10:e0133869. doi: 10.1371/journal.pone.0133869

Wong, A. (2017). Epistasis and the evolution of antimicrobial resistance. *Front. Microbiol.* 8:246.

Woodford, N., and Ellington, M. J. J. (2007). The emergence of antibiotic resistance by mutation. *Clin. Microbiol. Infect.* 13, 5–18. doi: 10.1111/j.1469-0691.2006.01492.x

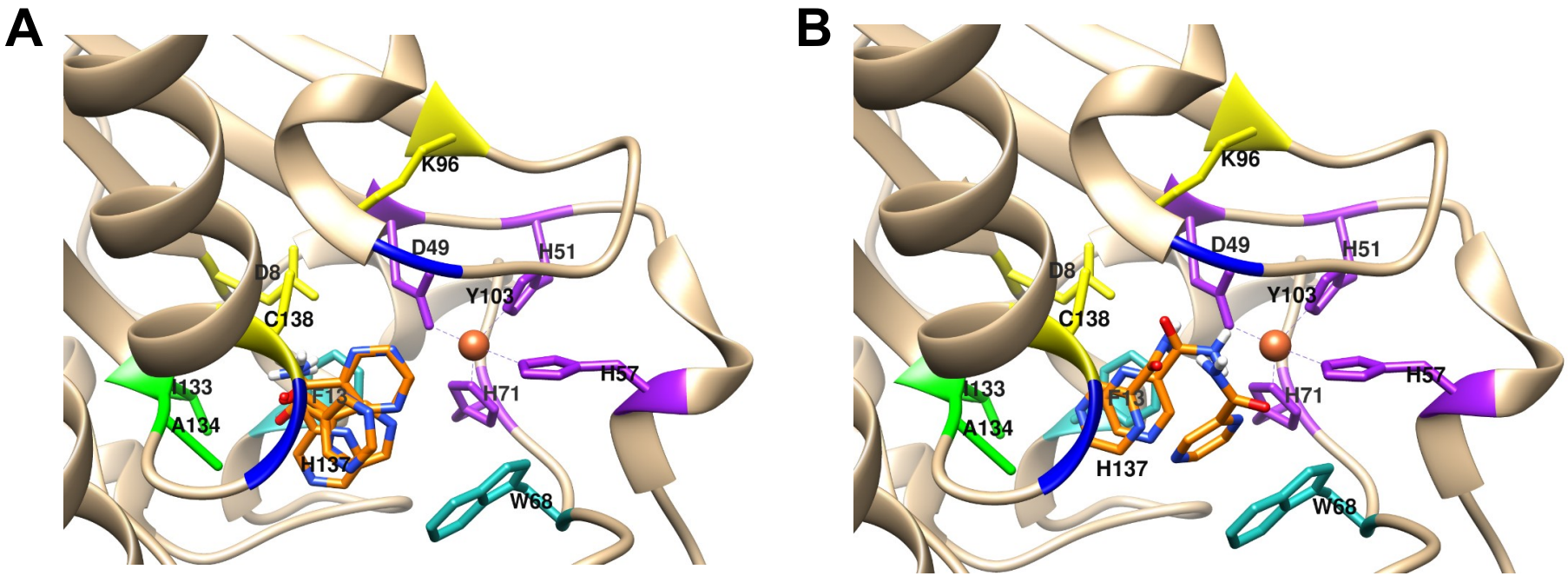
World Health Organization [WHO] (2020). *WHO Report on TB 2020*. Geneva: WHO.

Worth, C. L., Preissner, R., and Blundell, T. L. (2011). SDM-a server for predicting effects of mutations on protein stability and malfunction. *Nucleic Acids Res.* 39, W215–W222.

Wylie, C. S., and Shakhnovich, E. I. (2011). A biophysical protein folding model accounts for most mutational fitness effects in viruses (in press). *Proc. Natl. Acad. Sci. U. S. A.* 108, 9916–9921. doi: 10.1073/pnas.1017572108

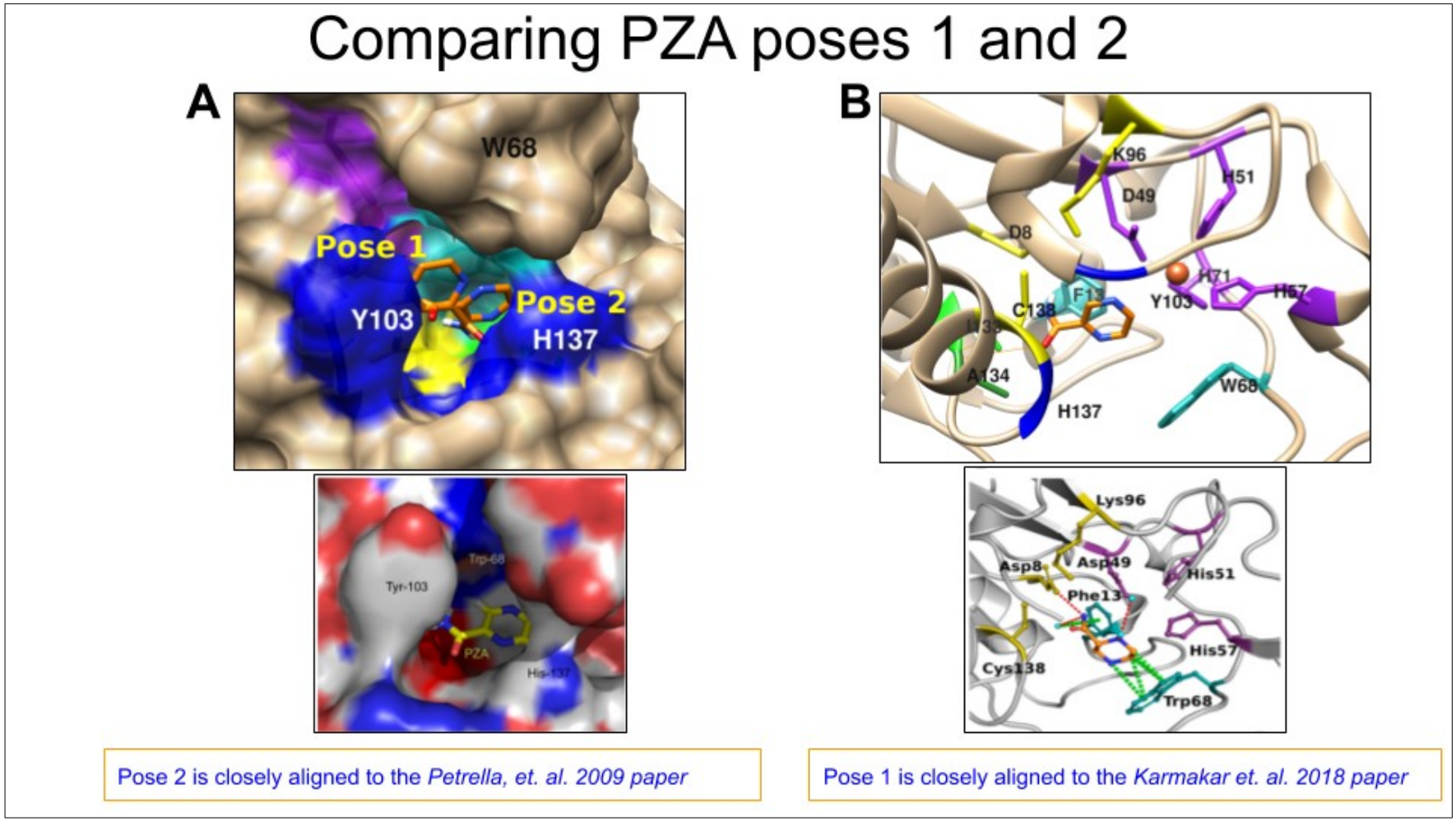
Zhou, X., and Stephens, M. (2012). Genome-wide efficient mixed-model analysis for association studies. *Nat. Genet.* 44, 821–824. doi: 10.1038/ng.2310

Conflict of Interest: The authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.



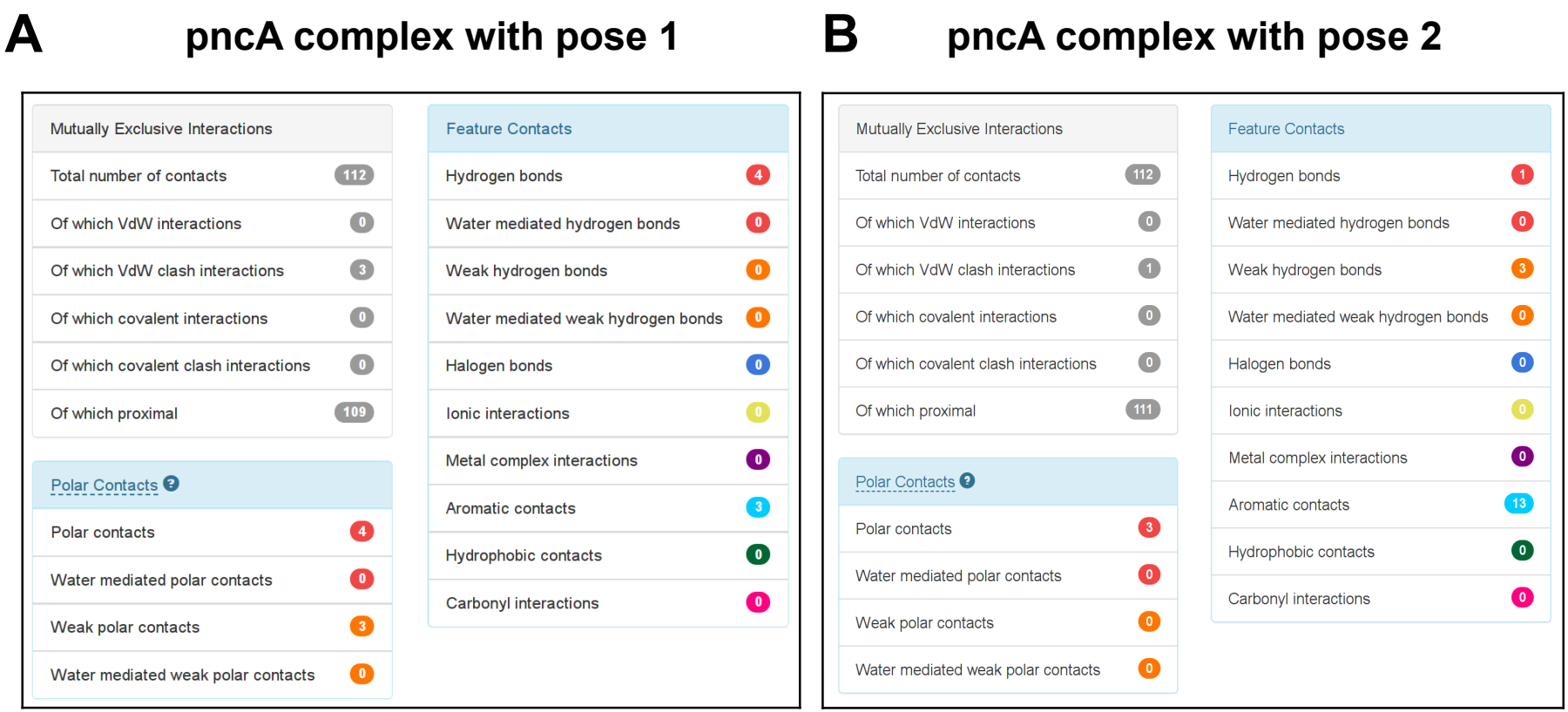
Supplementary Figure 1. Docking of Pyrazinamide (PZA) within pncA

Configuration of nine PZA poses returned by Autodock Vina located within the binding cavity, exploiting confirmations around the one rotatable bond in PZA. (**A)** Poses 1, and 3-6 with orientation of PZA ring towards the ring of tryptophan (W68), while **(B)** Poses 2, 8 and 9 showing the orientation of the PZA ring away from tryptophan. Residues marked in green participate in hydrogen bonding, residues in yellow form the catalytic triad, residues in teal (and blue) are involved in substrate binding, while residues in purple are involved in the iron centre. The figure is rendered using Chimera (version 1.14).



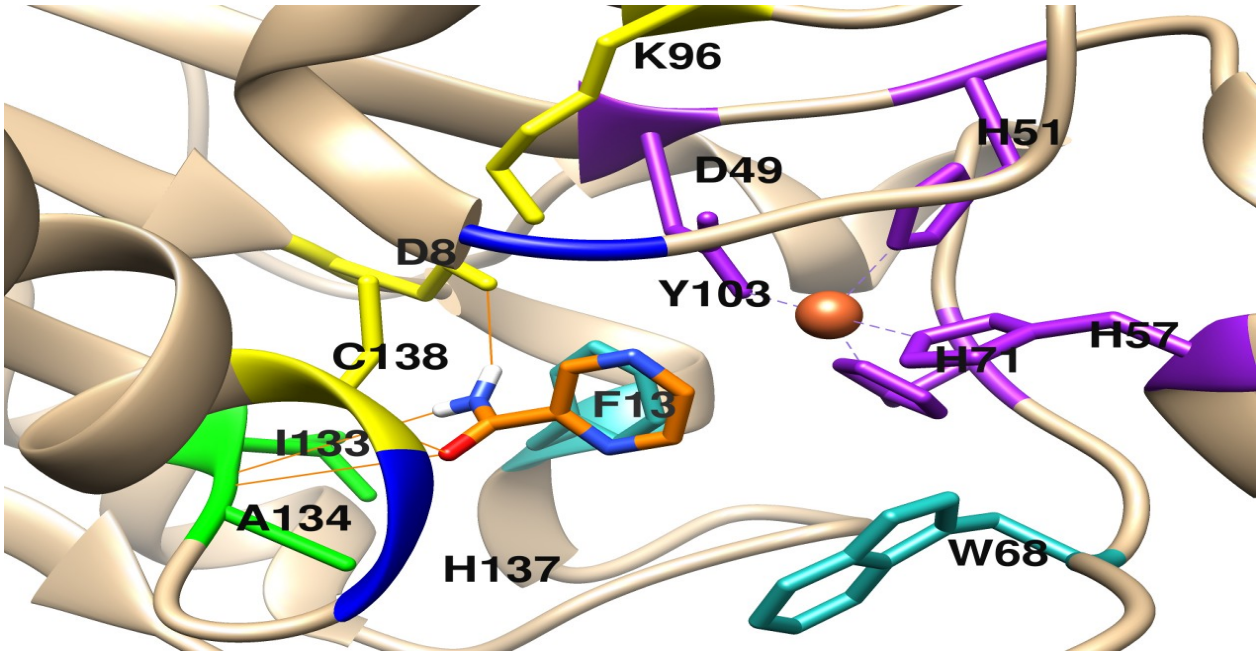
Supplementary Figure 2. Comparing PZA Poses 1 and 2 in relation to docking

(A) Comparison of poses 1 and 2 returned from Autodock Vina highlighting the differing orientation of the ring between the two poses. **(B)** Pose 1 resembles closely to the docking performed in the recent case report published (Karmakar et al., 2018), while pose 2 is closely aligned with the proposed binding cavity by the authors of the pncA crystal structure (Petrella et al., 2011).



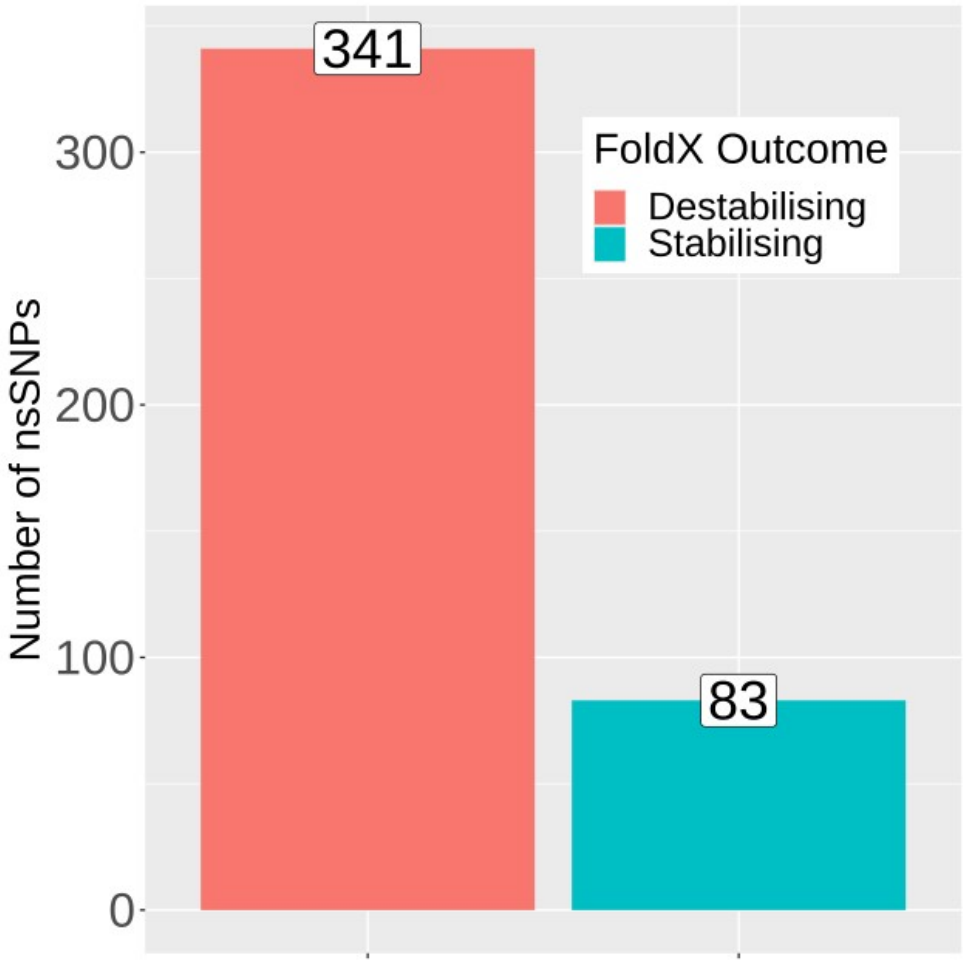
Supplementary Figure 3. Molecular interactions for PZA Poses 1 and 2

Arpreggio analyses showing molecular interactions between PZA pose1 (**A)** and **(B)** pose 2 reporting differences between hydrogen bonds, aromatic contacts, polar contacts and Van der Waals interactions. Screenshot from Arpgeggio web server (Jubb et al., 2017).

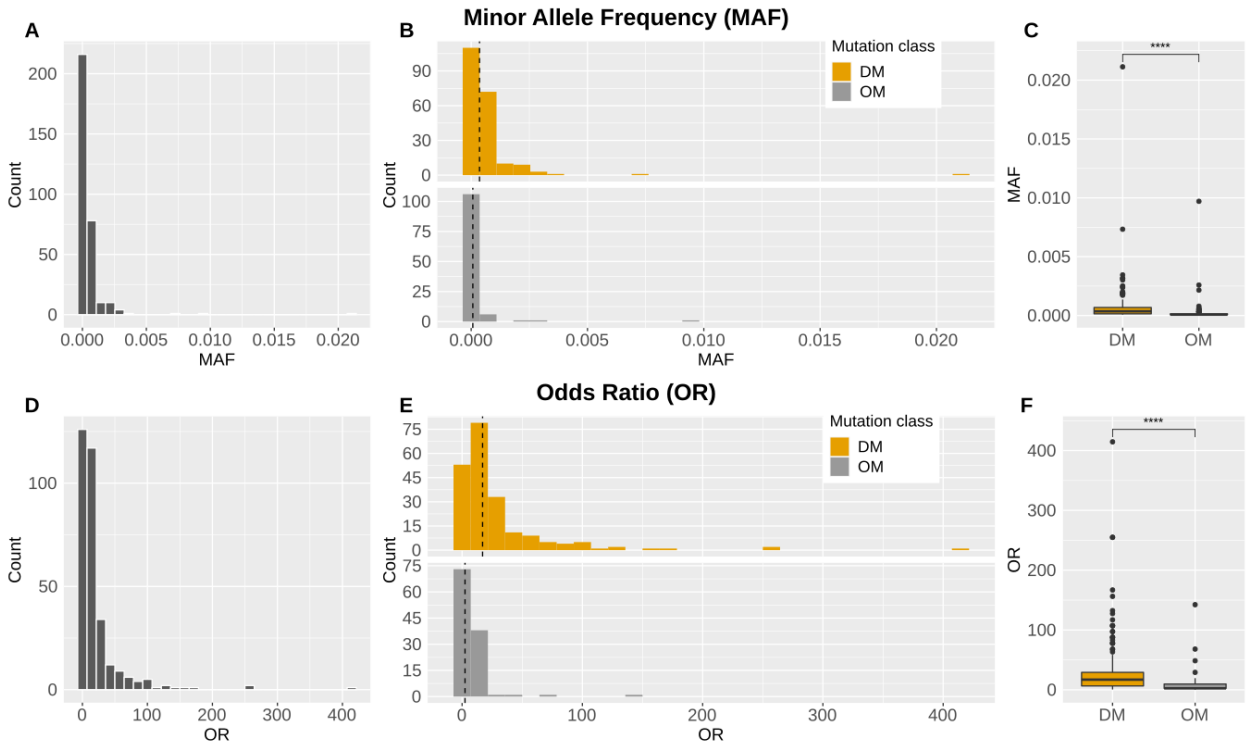


Supplementary Figure 4. Molecular docking of PZA with pncA

Protein-ligand complex formed by pncA with pose 1 of PZA after docking. Residues marked in yellow form the catalytic triad, residues in teal and blue are involved in substrate binding, while residues in purple are involved in the iron centre. Residues marked in green participate in hydrogen bonding, with hydrogen bonds between PZA and D8, I133, A134 and C138 are shown in orange. The figure is rendered using UCSF Chimera (version 1.14).

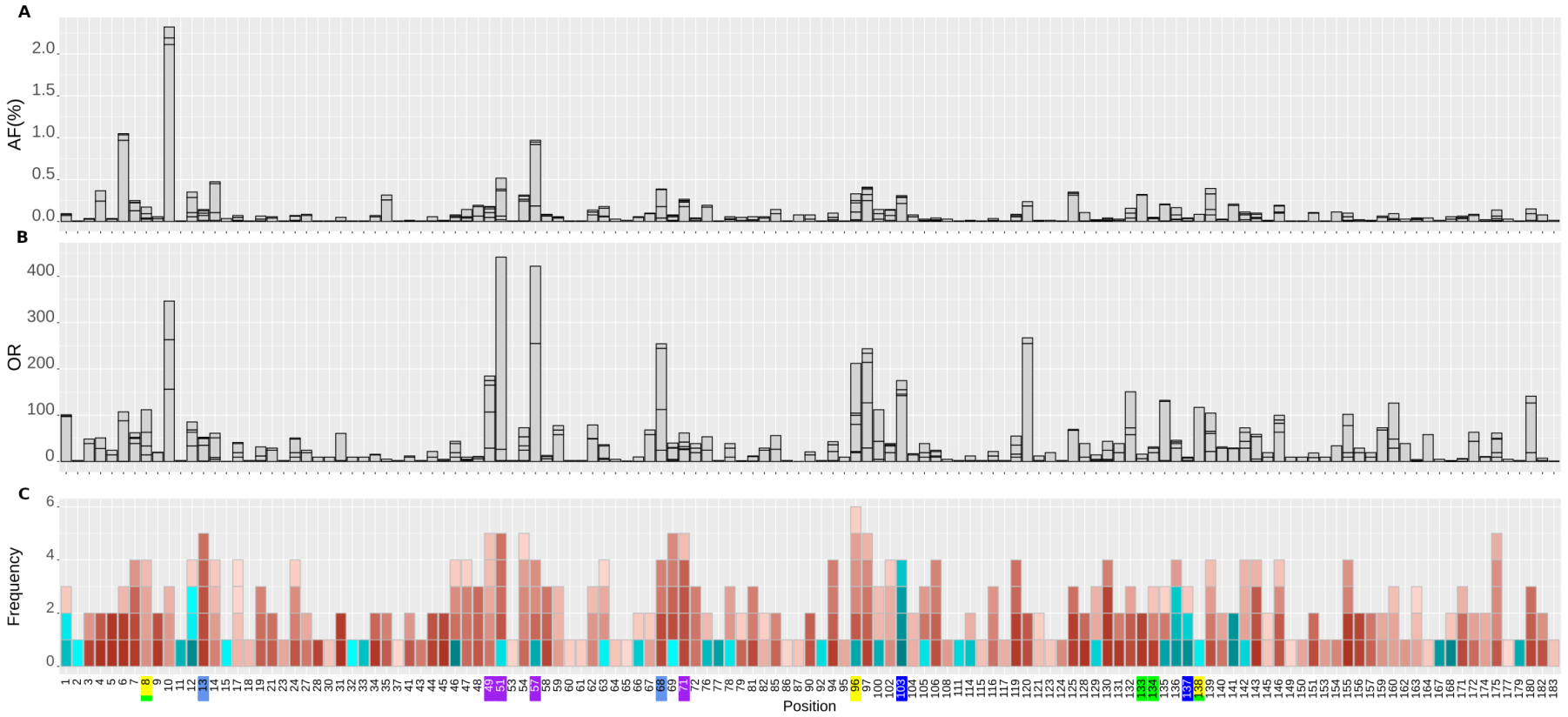


Supplementary Figure 5. Barplot of mutations with protein stability effect according to FoldX.
Number of mutations (SNPs) categorised as destabilising (n=341) and stabilising (n=83). The figure is generated using R statistical software (version 4.0.2).



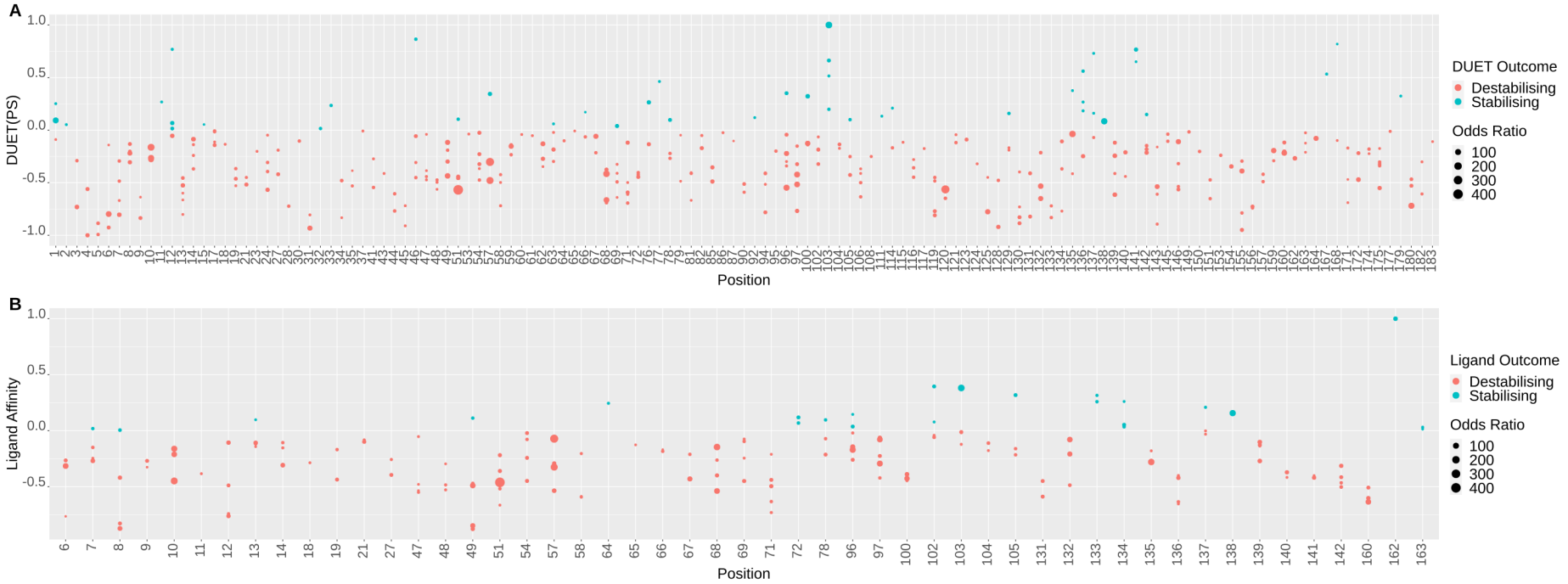
Supplementary Figure 6. Frequency distribution of Minor Allele Frequency (MAF) and Odds Ratio (OR) for pncA SNP mutations.

MAF and OR were calculated for a total of 322 SNPs. The top panel relates to Minor Allele frequency where **(A)** Histogram of MAF, **(B)** Histogram of MAF according to mutation class as either DM (associated with pyrazinamide resistance coloured in orange) or OM (not associated with pyrazinamide drug resistance coloured in grey). Dashed lines indicate median. **(C)** Box plot comparing MAF between DM and OM mutations. The bottom panel relates to Odds Ratio where **(D)** Histogram of OR, (**E)** Histogram of OR according to mutation class: 'DM' in orange and 'OM' in grey. Dashed lines represent median, **(F)** Boxplot comparing OR between DM and OM mutations. Wilcoxon rank-sum (unpaired) test was used to compare DM and OM mutations, and significance indicated as ****P<0.0001. The figure is generated and statistical analysis performed using R statistical software (version 4.0.2).



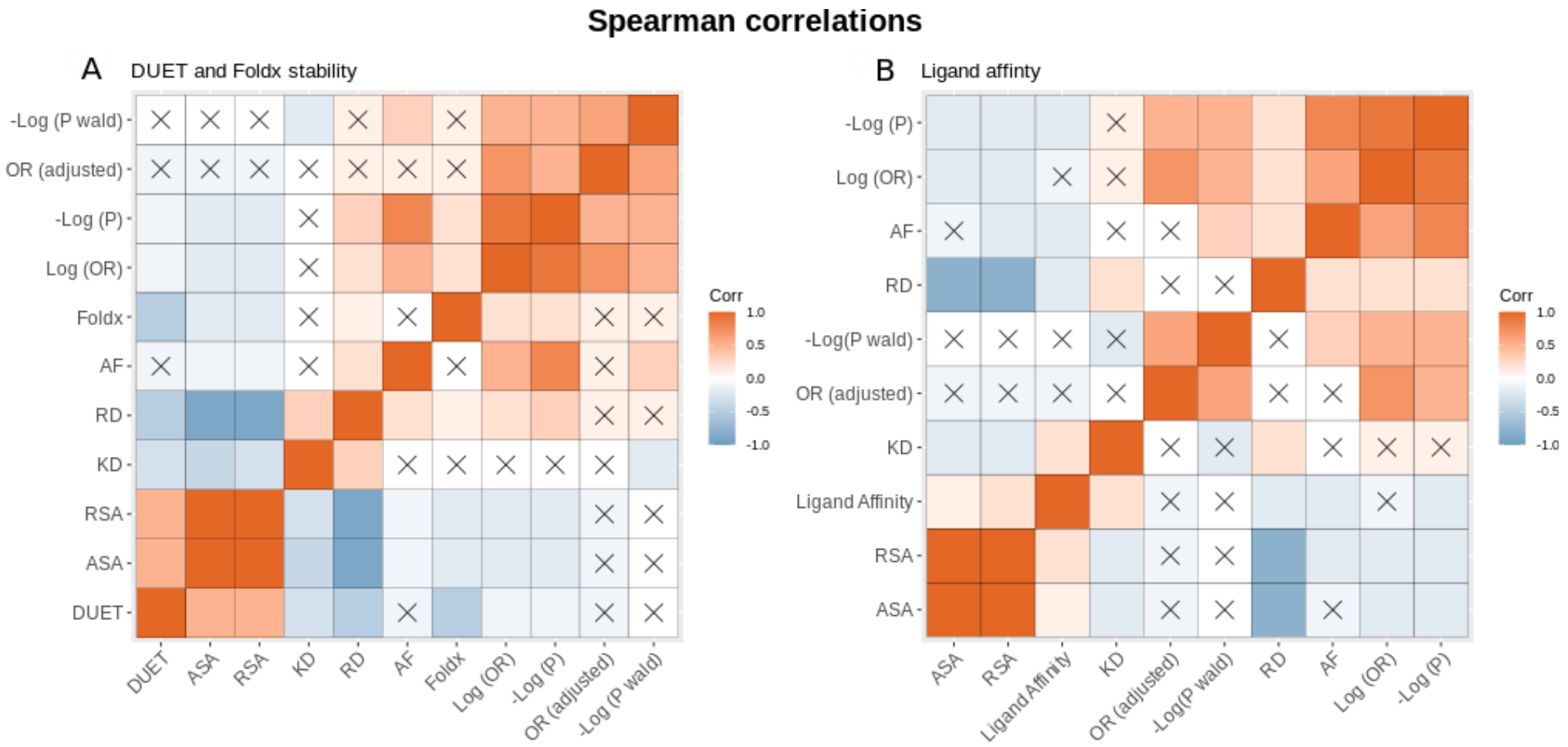
Supplementary Figure 7. Allele Frequency (AF), Odds Ratio (OR) and DUET effects of SNPs within pncA

Barplot showing 322 mutations associated with AF, OR and SNP diversity by position highlighting the prominent positions in terms of AF. OR and frequency of SNPs within pncA. The horizontal axis shows the mutational positions within pncA and are coloured as green (residues involved in hydrogen bonding with PZA) yellow (catalytic triad), blue and teal (substrate binding), purple(iron centre).The vertical axis shows **(A)** cumulative AF associated with one or more mutations at that position, **(B)** cumulative OR associated with one or more mutations at the given position and **(C)** the frequency of SNPs at mutational position within pncA. The red and the blue bars denote destabilising (n=279) and stabilising (n=43) mutations for a total of 322 mutations according to DUET. Destabilising mutations are depicted in red and stabilising mutations in blue, where colour intensity reflects the extent of effect, ranging from -1 (most destabilising) to +1 (most stabilising). The figure is generated using R statistical software (version 4.0.2).



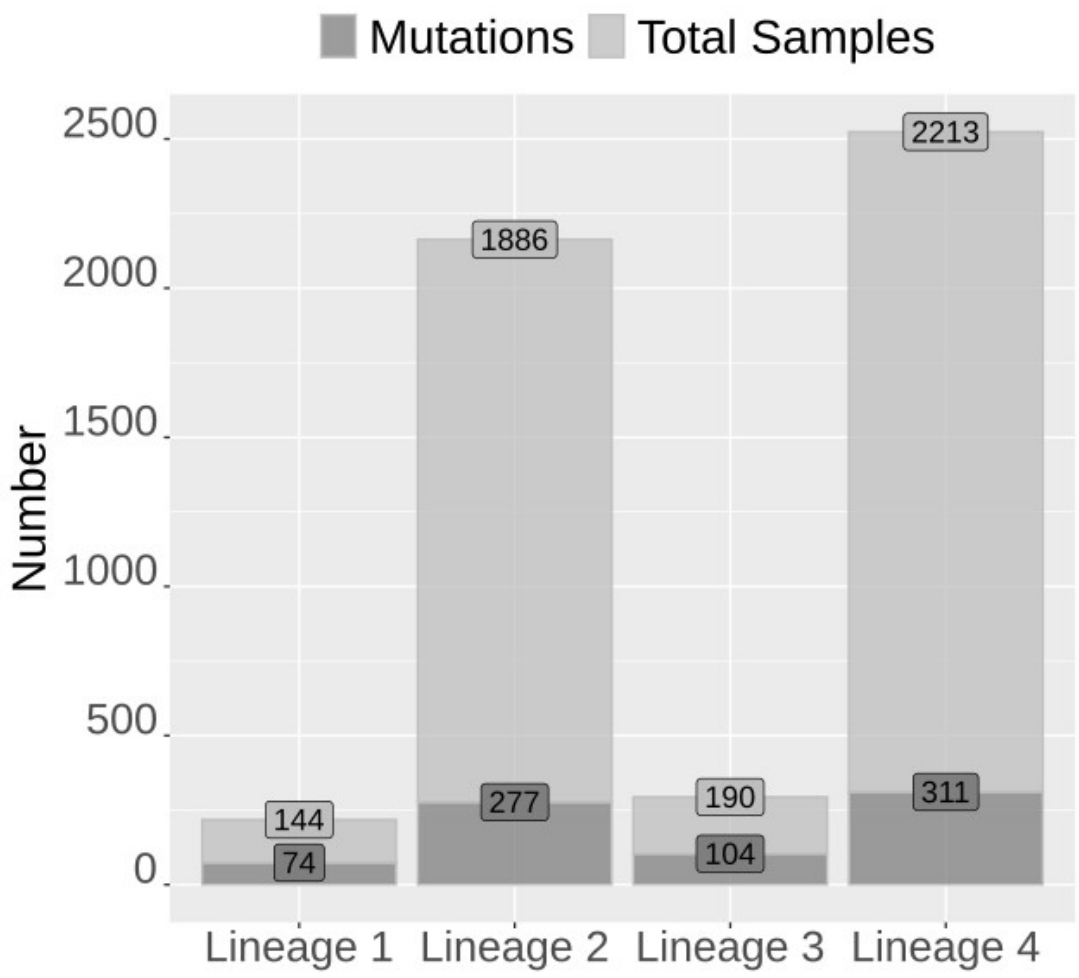
Supplementary Figure 8. Comparing stability effects of SNPs with GWAS measures of Odds Ratio (OR)

Bubble plot displaying the relationship between OR with **(A)** DUET Protein stability and **(B)** Ligand affinity corresponding to and 322 and 160 mutations respectively. The horizontal axis shows the mutational positions within pncA and the vertical axis shows protein stability effects ranging from -1 (most destabilising) to +1 (most stabilising). Each dot represents a unique mutation at that position, with the colour corresponding to destabilising (red) and stabilising (blue) mutations, while the size of the dot is proportional to the OR of that mutation. The figure is generated using R statistical software (version 4.0.2).



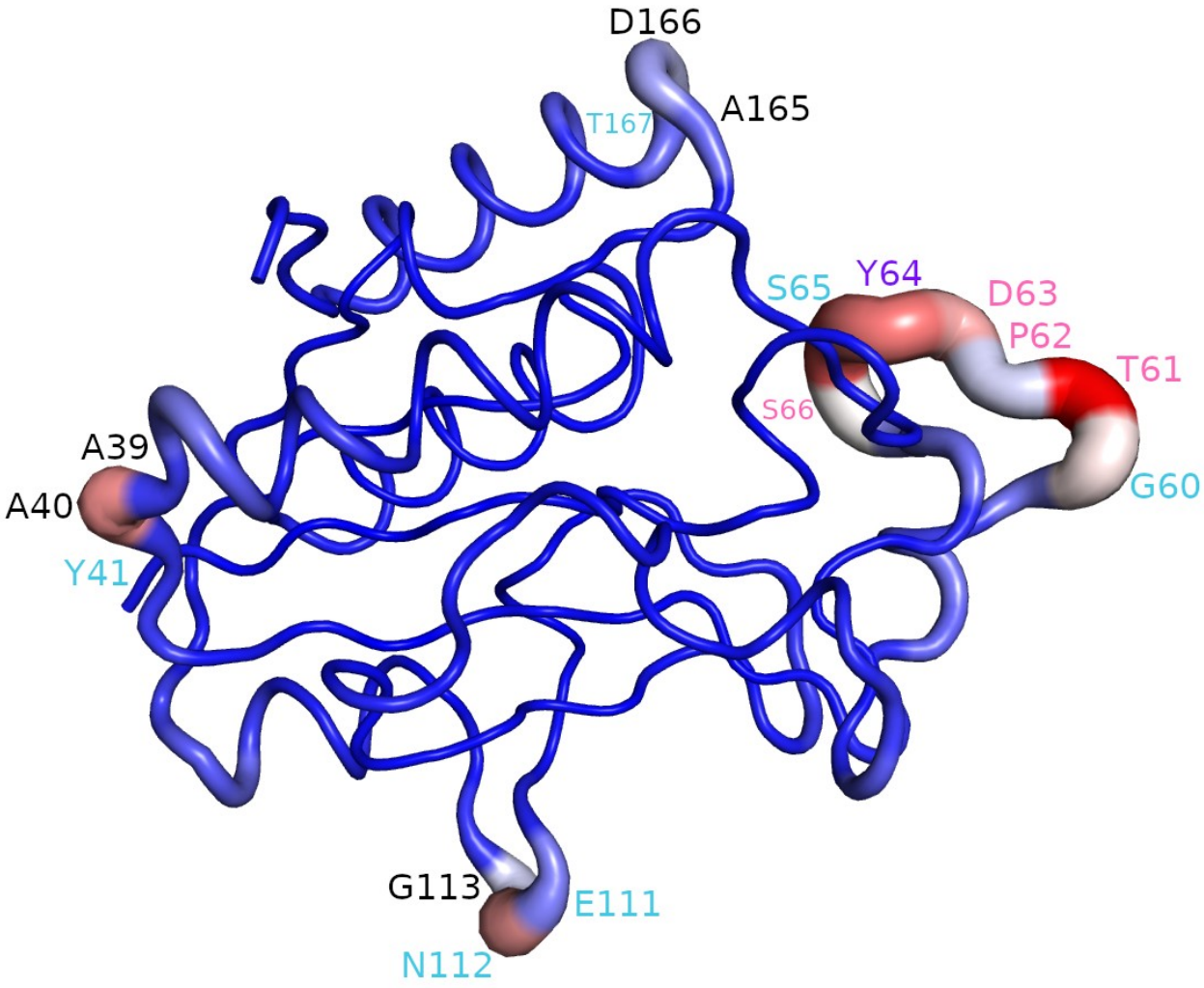
Supplementary Figure 9. Correlation of protomer stability and ligand affinity effects with GWAS and structural parameters.

Pairwise Spearman correlations between Foldx stability, DUET stability, Allele Frequency (AF), negative log P-value (-Log(P)), Log Odds Ratio (OR), adjusted Log OR (accounting for sample relatedness in GWAS analysis), negative log P-value from Wald test corresponding to the adjusted OR, along with structural parameters of accessible (ASA) and relative (RSA) surface area, KD (hydrophobicity values based on the Kyte and Doolittle scale) and RD (Residue Depth). The parameters are ordered using hierarchical clustering. Squares marked with an 'X' indicate statistical insignificance (P>0.05). Part **(A)** shows correlations with DUET and Foldx stability values, for a total of 424 SNPs, while **(B)** shows correlations with Ligand affinity for a 201 SNPs. The figure is generated using R statistical software (version 4.0.2).



Supplementary Figure 10. Barplot of total samples and mutations within Mtb Lineages

The total number of samples along with the number of mutations associated with PZA resistance within the 4 Mtb Lineages. The dark grey bars show the number mutations, while the light grey bar show the total number of samples within each lineage. Lineage 1 has 74 mutations out of 144 samples, Lineage 2 has 277 mutations out of 1886 samples, while lineages 3 and 4 have 104 and 311 mutations out of 190 and 2213 number of samples respectively. The figure is generated using R statistical software (version 4.0.2).



Supplementary Figure 11. Protein fluctuation analysis of pncA structure (3PL1) based on Normal Mode Analysis (NMA).

Sites associated with fluctuation as depicted by NMA. The magnitute of the fluctuation is represented by thin to thick tube coloured blue (low), white (moderate) and red (high). The corresponding wild-type residues (using the standard one-letter code) at these sites are labelled and coloured according to the mutational effects of one or more nsSNPs at these sites: Drug resistant mutations (DM) are coloured purple, Other mutations (OM) appear in blue, while sites linked to mutations belonging to either category are coloured in pink. Sites associated with no nsSNPs in our study are depicted in black. The NMA analysis and figure is generated from the DynaMut web server. Abbreviations used: pncA: pyrazinamidase.

Supplementary Table 1 (part of all supplementary material) for this article can be found online at: https://www.frontiersin.org/ar- ticles/10.3389/fmolb.2021.619403/full#supplementary-material

Lineage comparisons	**All mutations** (n=4433)		
KS test to compare distributions	P-value	Adj P-value	*Adj P-value signif*
Lineage 1 vs Lineage 2	1.48E-03	8.87E-03	**
Lineage 1 vs Lineage 3	5.33E-03	3.20E-02	*
Lineage 1 vs Lineage 4	9.69E-05	5.81E-04	***
Lineage 2 vs Lineage 3	4.46E-04	2.68E-03	**
Lineage 2 vs Lineage 4	<2.2E-016	<0.0001	****
Lineage 3 vs Lineage 4	8.09E-03	4.86E-02	ns

Lineage comparisons	**DM** (n=3565) vs **OM** (n=868)		
KS test to compare distributions	P-value	Adj P-value	*Adj P-value signif*
Lineage 1	7.31E-08	2.92E-07	****
Lineage 2	<0.0001	<0.0001	****
Lineage 3	2.97E-01	1.19E+00	ns
Lineage 4	2.55E-13	1.02E-12	****

Supplementary Table 2: Kolmogorov-Smirnoff (KS) test reporting the statistical differences in distributions between Mtb lineages when assessed based on Protein stability (DUET outcome). Lineage comparisons were performed for all mutations, and between mutations associated with PZA drug resistance (DM) and other mutations (OM). s). Adj. P-values: Bonferroni adjusted P-values, n=number of samples, ns = not significant, Adj. P-value *signif:* Statistical significance thresholds used are *P<0.05, **P<0.01, ***P<0.001, ****P<0.0001.

Chapter 4

EmbB-ethambutol results

4.1 Background

4.1.1 Mechanism of action of Ethambutol

Ethambutol (EMB) is a drug used in the treatment of tuberculosis as a combination therapy with isoniazid, rifampicin and pyrazinamide. The main target for EMB is an arabinosyltransferase, termed *embB,* and to a lesser extent the other genes (*embC and embA*) in the *embABC* operon.[1,2] The arabinosyltransferases (EmbA, EmbB, and EmbC) are enzymes involved in the polymerisation of arabinogalactan, an essential component of the mycobacterial cell wall. The *M. tuberculosis* cell wall is a highly complex structure enriched with lipids and carbohydrates, consisting of three distinct layers: peptidoglycan, arabinogalactan and mycolic acids **(Figure 1A)**. The mycolic acids are covalently linked to the peptidoglycan via an arabinogalactan network.[3] The unique composition of the mycobacterial cell wall with its low permeability is responsible for its pathogenicity and virulence, and help the bacteria evade host immune response.[4] with lipid-mediated defence mechanisms.[5] Ethambutol is bacteriostatic against actively growing TB bacilli. It works by disrupting the arabinogalactan synthesis by inhibiting the enzyme arabinosyl transferase required for the cell wall synthesis.[6] This leads to increased cell wall permeability, allowing the drug to further diffuse into the *M. tuberculosis* cells. Once inside the cell, EMB prevents formation of the cell wall component arabinogalactan and lipoarabinomannan, the latter being a crucial virulence factor as well as a key component in modulating the host-pathogen interaction.[7,8] An overview of the mechanism of action of EMB is shown in **Figure 1B**.

4.1.2 Active site description and EMB resistance

While mutations in the embABC operon are responsible for EMB resistance, the majority of EMB resistance in clinical isolates occurs due to mutations in the *embB* gene.[1,10,11] **(Figure 1B)**. The structural characterisation of EMB in complex with *M. tuberculosis* EmbB was recently investigated using cryoelectron microscopy and X-ray crystallography, generating structures of mycobacterial EmbA-EmbB and $EmbC_2$ in the presence of their donor (decarpaneyl phosphate, DPA) and acceptor (arabinan) substrates, as well EMB. The overall complex is a hetero-trimer, with EmbA-EmbB forming a hetero-dimer complex, while EmbC forms a symmetric homo-dimer. The hetero-dimer complex is stabilised by the presence of cardiolipin (CDL), while the presence of a calcium (Ca^{2+}) ion is responsible for the structural stability of EmbC.[12] Consistent with clinical drug resistance studies, the authors demonstrate that EMB preferentially binds to EmbB and EmbC rather than EmbA,[1,10] with a high degree of similarity between the binding modes.[11] EMB competes with donor DPA, and ac-

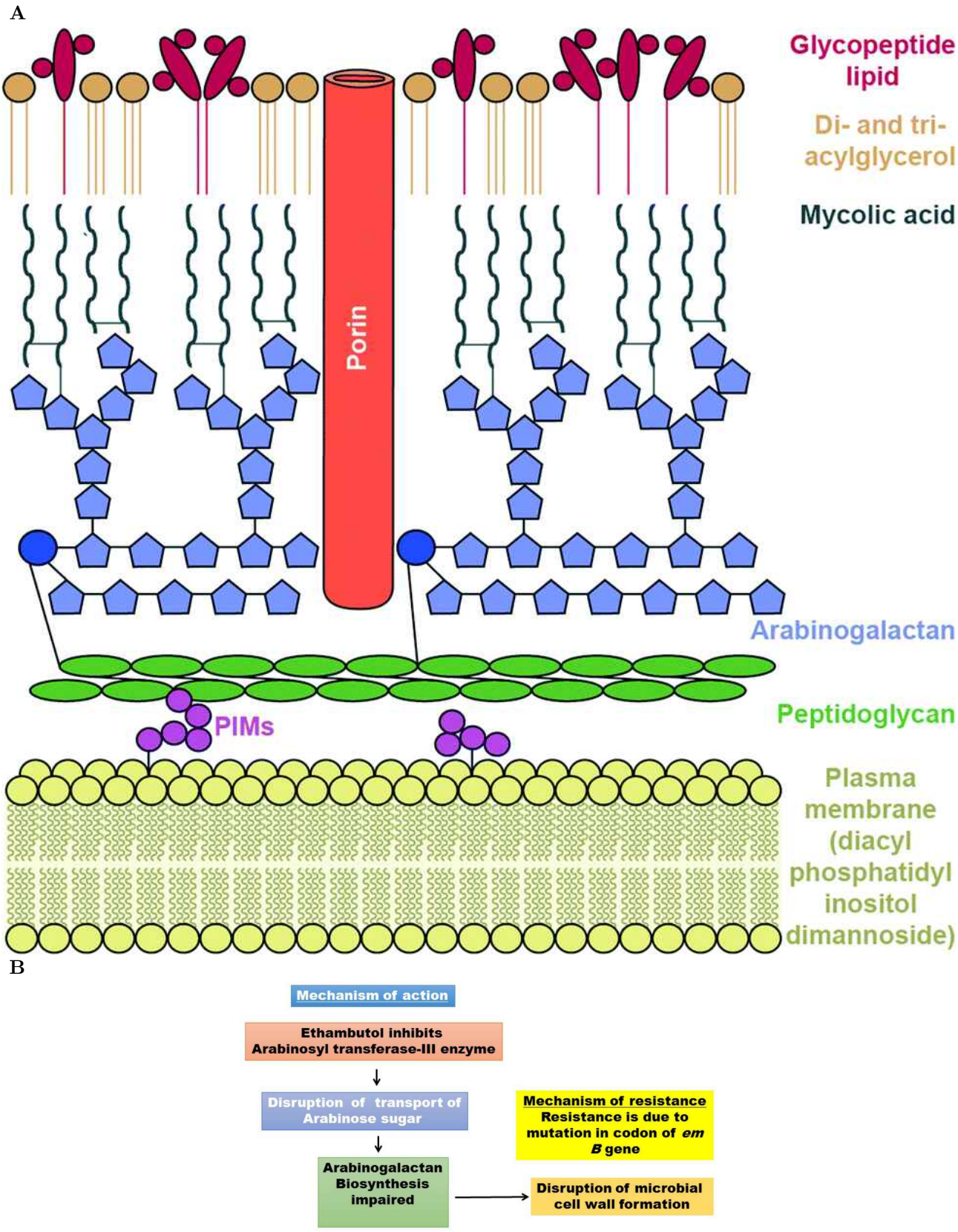


Figure 1: Cell wall components of *M. tuberculosis* and mechanism of action for ethambutol
A) The complex cell wall components of *M. tuberculosis*. The peptidoglycan and arabinogalactan form the cell wall, while mycolic acids and glycolipids form part of the outer membrane. Together these form the mycobacterial cell envelope. Figure adapted from Sylvie Garneau-Tsodikova and Kristin J. Labby, 2016,[7] **B)** An overview of the mechanism of action and resistance in *M. tuberculosis* for ethambutol. Figure adapted from Sheikh, *et. al.*.[9]

ceptor arabinose substrates for binding to the EmbB and EmbC subunits.[11] The active site of EmbB with EMB bound is characterised by interactions with residues D299, Y302, I303, E327, M306, W592, H594, W988, and W1028 **(Figure 2A)**.[11] Residue D299 forms three electrostatic interactions with EMB, while the two hydroxyl (OH) groups of EMB form hydrogen bonds with residues H594 and E327 **(Figure 2A)**.[11] Residues I303, M306, W592, and W1028 form van der Waals interaction while residues W988 and Y302 form pi-cation interactions with EMB.

Sites G406 and Q497 are considered resistant hotspots along with site M306, a conserved site in all EmbB proteins. Mutations M306V and M306I are shown to be favoured by resistant clinical isolates due to reduced binding affinity for EMB.[13] Mutations at site M306 also result in disruption of its surrounding interactions network involving residues Y302 and E327 which in turn are involved in interacting with EMB. Residues G406 and Q497, despite being further away from EMB (>10Å) disrupt surrounding interaction networks involving residues E328 and E327, where the latter is involved in EMB interaction. Structural insight into mutations I289M and I289F also revealed resistance development as a result of steric hindrance without affecting enzymatic activity.[11] Additionally, all possible interacting residues for EMB, DPA, CDL, and Ca^{2+} ion were identified using Arpeggio, PLIP, and LigPlus tools. **Figure 2B** shows the hetero-trimer complex of the embABC (PDB-ID: 7BVF), with interacting residues for EMB, and other interacting partners: DPA, CDL and Ca^{2+} ion.

Interactions in EmbB

Molecular interactions with residues in EmbB, EMB, and interacting partners (DPA CDL, and Ca^{2+} ion) were identified using LigPlus, PLIP and Arpeggio tools, resulting in a total of thirty-four interacting residues shared among some of the interacting partners:

- Fourteen residues at sites 298, 299, 302, 303, 306, 318, 327, 334, 403, 445, 592, 594, 988, and 1028 were identified to be interacting with EMB.
- Twenty-nine residues at sites 299, 322, 329, 330, 403, 435, 438, 439, 442, 445, 446, 449, 452, 455, 486, 489, 490, 493, 506, 509, 510, 513, 514, 515, 587, 589, 590, 592, 595 were identified to be interacting with DPA.
- Twenty-eight residues at sites 456, 457, 460, 461, 521, 525, 533, 537, 554, 558, 568, 569, 572, 573, 575, 576, 579, 580, 582, 583, 586, 601, 605, 616, 658, 661, 662, 665) were identified to be interacting with CDL
- Nine residues at sites 847, 853, 854, 952, 954, 955, 956, 959, 960 were identified to be interacting with the Ca^{2+} ion.

An overview of the EmbB-EMB structural complex with all interactions shown in **Figure 2C**.

Information gathered in this manner was used to identify all possible interactions for key binding partners (EMB, DPA, CDL, Ca^{2+} ion) to curate the active site residues for EmbB protein. This was done to help visualise active sites in relation to mutational diversity, EMB resistance, and to inform downstream ML analysis.

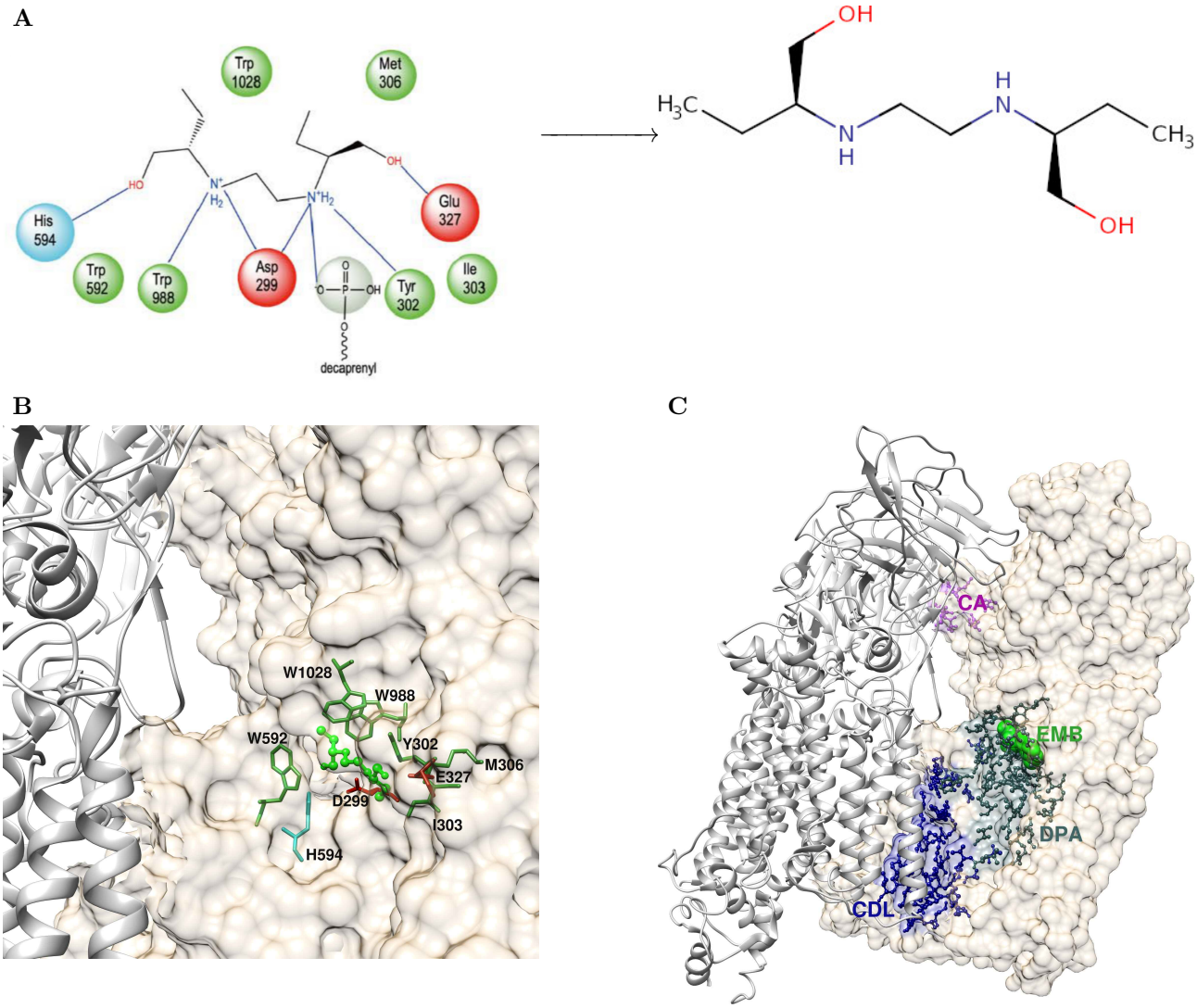


Figure 2: Active site description of *M. tuberculosis* EmbB and its interacting partners

Overall description of the EmbB-EMB complex and its interacting partners (PDB-ID: 7BVF). **A)** Active site residues in EmbB with the chemical structure of EMB indicated on the right. Figure adapted from Zhang, *et. al.*,[11] **B)** The residues indicated in part A coloured accordingly are indicated in the structure of EmbB-EMB complex, **C)** hetero-trimer EmbB-EMB complex with EmbB (chain B) indicated as surface representation in tan colour, while chains A and B are shown as grey ribbons. All interactions for binding partners indicated: EMB with its interacting residues appear in green; DPA and its interacting residues are shown in dark slate grey; CDL and its interacting residues appear in navyblue; Ca^{2+} ion and its interacting residues shown in purple. Abbreviations used: EMB: ethambutol, DPA: decaprenyl-phosphate-arabinose, CDL: cardiolipin, Ca^{2+} ion: calcium ion.

4.2 Structural and genomic insights into ethambutol resistance

4.2.1 Mutational landscape of EmbB

Sites with multiple SAVs (hotspots) are located away from the active site, with most active site residues displaying single mutations

A total of 858 SAVs were found in the protein coding region of *embB* (Genomic id: Rv3795, coding region: 4246514-4249810), and appear distributed across the protein (**Figure 3**), with mutations present in 570 unique positions in EmbB **(Figure 4**).

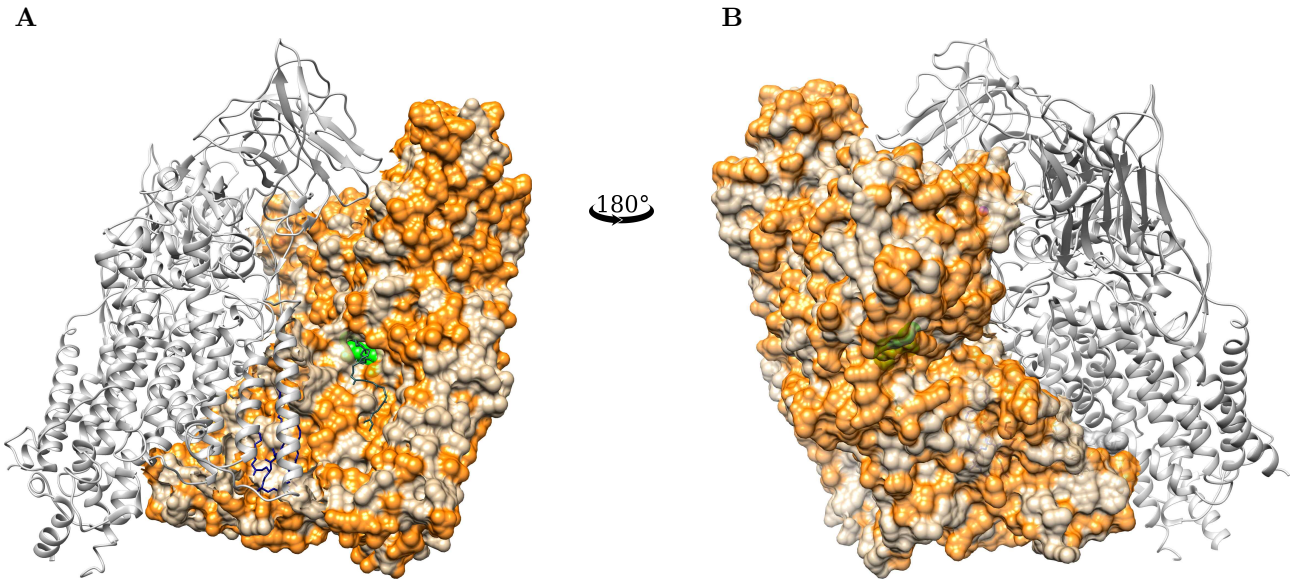


Figure 3: Mutational landscape of *M. tuberculosis* EmbB

An overview of all mutational sites on *M. tuberculosis* EmbB chain B (PDB-ID: 7BVF) appearing as surface representation in tan colour with chains A and P appearing as grey ribbons. Panels **A)** and **B)** are opposing representations (rotated 180°) of EmbB, with EMB shown in green as spheres in the binding pocket. The figure is generated using UCSF Chimera version 1.14. Abbreviations used: SAV: single amino acid variation, EMB: ethambutol.

A minority (5/14 for EMB, 11/29 for DPA, 3/9 for Ca^{2+} ion) of residues in the active site excluding those interacting with CDL at the PPI **(Figure 2)** displayed SAVs. Sites beyond the active site showed multiple SAVs with a maximum of 6 SAVs at three non-active site residues: F330, G406, and Q497 (**Figure 4**). A majority (19/28) of residues involved in CDL interactions were however associated with SAVs, budding resistant hotspots being most prominent at the PPI **(Figure 4**, sites marked in navy blue**)**. Mapping mutations by positions in EMB (**Figure 4**) highlight the following:

Prominent mutation hotspots not involving the active site

- Single or budding resistant hotspots: None
- Hotspots with four mutations: G305, D311, A356, E378, S500, K511, V744
- Hotspots with five mutations: A201
- Hotspots with six mutations: D328, G406 and Q497

Sites with EMB interactions associated with a maximum of 4 SAVs (sites marked in green)

- Single mutation: N318, H334, and Q445
- Budding resistant hotspots: I303
- Hotspots with four mutations: M306

Sites with DPA interactions associated with a maximum of 4 SAVs (sites marked in dark slate grey)

- Single mutation: V435, Q445, I489, V493, T506, L449, V452
- Budding resistant hotspots: A438 and A510
- Hotspots with three mutations: A439
- Hotspots with four mutations: F330

Sites with CDL interactions associated with a maximum of 3 SAVs (sites marked in navy blue)

- Single mutation: A457, R573, G576, A601, L558, G569, M582, M586, P616, P661, K662
- Budding resistant hotspots: E521, V554, R568, M575, G580, S658, G665
- Hotspots with three mutations: V456

Sites with Ca^{2+} ion interactions associated with a maximum of 2 SAVs (sites marked in purple)

- Single mutation: Q854, T956
- Budding resistant hotspots: Q853

Resistant hotspot sites D328, G406, and Q497 display the highest (6 SAVs each) number of SAV mutations in EmbB, despite these sites not being directly involved with EMB binding. Resistance is thought to arise due to disruption in the EMB interaction network **(Figure 2)**. For example, it is thought that SAVs at site 497 cause conformational changes that affect E327, one of the EMB binding sites.[14] Similarly SAVs at site G406 and D328 may also affect drug binding by disrupting underlying molecular interactions causing protein conformational changes.[14] It appears that resistant hotspots

not involving EMB binding residues, affect drug binding indirectly, by disrupting the network of EMB binding residues or by inducing local protein conformational changes, which reduce EMB binding affinity.[11]

Of the two active site residues, M306 and F330 displaying four SAVs each, the amino acid property only changed for a single mutation at these sites. For example, when considering mutations M306I, M306L, M306T, and M306V, only mutation M306T changed the property of the amino acid from hydrophobic amino acid in the wild-type (M) to a polar residue (T) while all other mutations (M306I, M306L, and M306V) retained the hydrophobic amino acid property similar to the wild-type **(Figure 4**, site marked in green**)**. Similarly, residue F330 involved with DPA interactions, displaying multiple SAVs (F330S, F330V, F330I, and F330L) also retained the hydrophobic wild-type amino acid property (F) for all mutations except F330S **(Figure 4**, site marked in dark slate grey**)**. The majority (55%, n=474) of the mutational effects resulted in electrostatic changes.

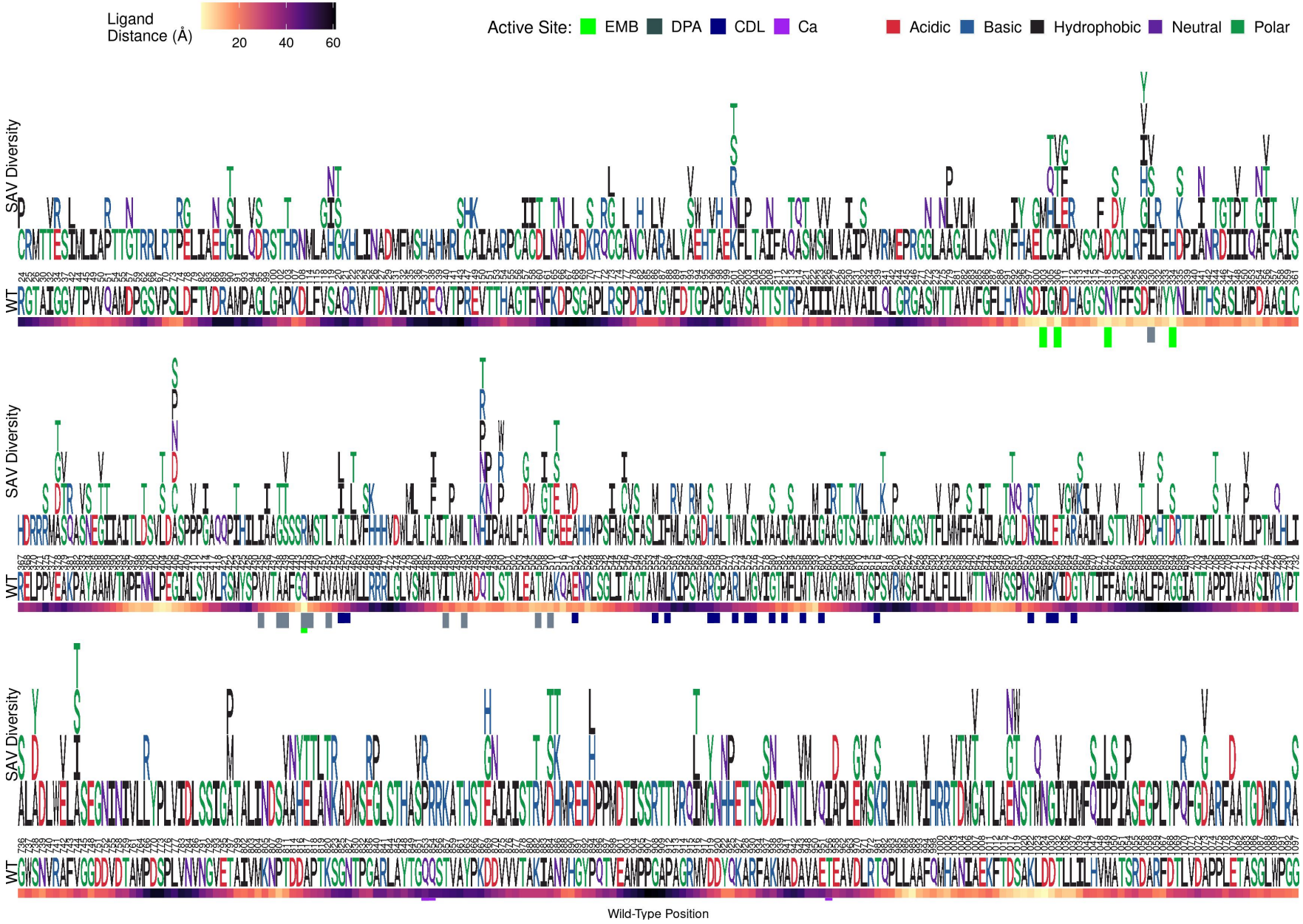

Ligand Distance (Å)
20
40
60
Active Site: EMB DPA CDL Ca
Acidic Basic Hydrophobic Neutral Polar
SAV Diversity
WT
Wild-Type Position

Figure 4: Sites associated with SAVs in *M. tuberculosis* EmbB
Logo plot showing 570 unique sites/positions associated with 858 SAVs in *M. tuberculosis* EmbB. The horizontal axis shows the wild-type positions associated with SAVs in EmbB and the vertical axis shows all the mutant residues observed in our data highlighting SAV diversity at a given site. Residues are coloured according to the amino acid (aa) property where acidic aa appear in red, basic aa appear in blue, hydrophobic aa in black, neutral aa in purple, and polar aa in darkgreen. The structural positions associated with SAVs in EmbB are indicated on the horizontal axis. The wild-type (WT) residues also coloured according to aa property appear under the respective position markings. The heat bar underneath the WT residues indicate the distance of that position from EMB according to the magma colour gradient where light yellow indicates sites closer to EMB (ligand distance in Angstroms). The positions are further annotated to reflect active site residues involved in interactions with EMB (green), DPA (dark slate grey), CDL (navy blue), and Ca^{2+} ion (purple). The figure is generated using R statistical software version 4.0.2, ggplot2 package. Abbreviations used: SAV: single amino acid variation, EMB: ethambutol, DPA: decaprenyl-phosphate-arabinose, CDL: cardiolipin, Ca^{2+} ion: calcium ion.

4.2.2 Mutational outcome from protomer stability changes and evolutionary conservation

Mutational consequences are destabilising for protomer stability without affecting protein function

Most mutations had a destabilising effect on the overall protomer stability when assessed using different computational tools **(Figure 5A-D)**, with mCSM-DUET estimating over 90% (n=790) as destabilising, followed by ~84% (n=721) predicted as destabilising by DeepDDG, and Dynamut2 predicting ~80% (n=684) as destabilising mutations. FoldX predicted the fewest with just over 60% (n=524) of mutations as destabilising. From an evolutionary conservation perspective, over 60% of the mutations were predicted to have a non-deleterious impact (effect) on protein function indicated by PROVEAN and SNAP2 scores. PROVEAN estimated 62% (n=532) SAVs with neutral effect **(Figure 5E)** while SNAP2 predicted 66% (n=570) SAVs with neutral effect **(Figure 5F)**. The mentioned computational estimates were independently run for all 858 SAVs, without assessing for agreement among them.

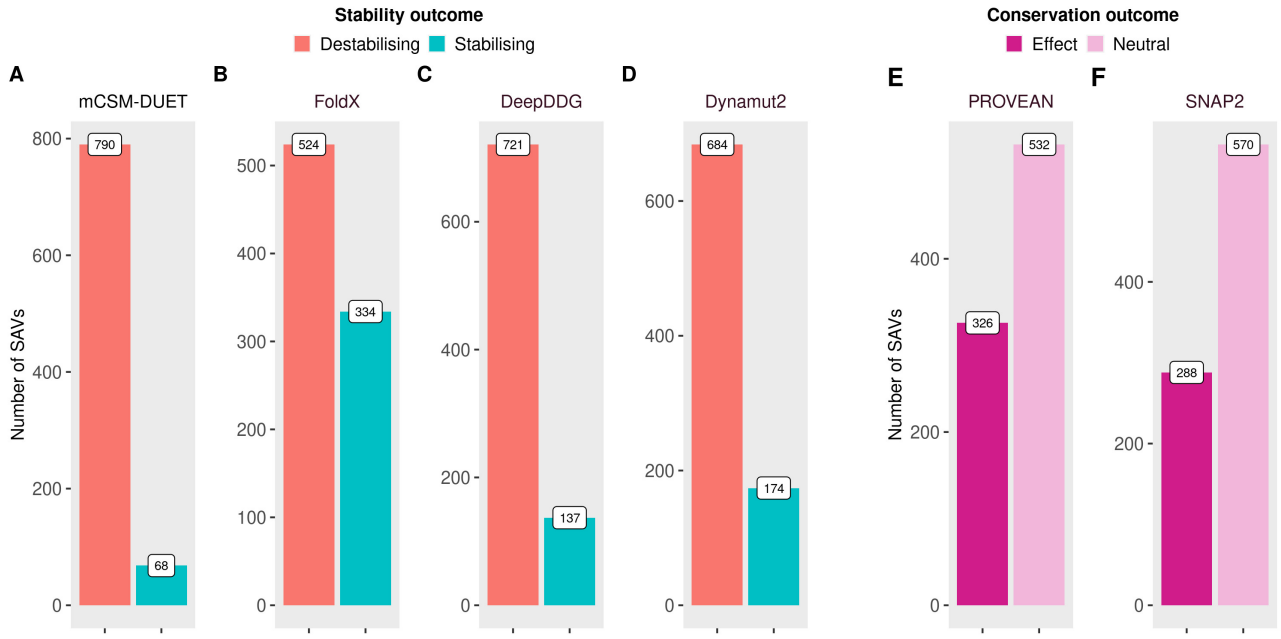


Figure 5: Protein stability outcome of SAVs in *M. tuberculosis* EmbB
Mutational impact on overall protein stability and evolutionary conservation changes for 858 SAVs, **A-D)** Barplots showing number of SAVs categorised as destabilising (red) or stabilising (blue) according to protein stability changes (ΔΔG Kcal/mol) as measured by four computational tools: mCSM-DUET, FoldX, DeepDDG, and Dynamut2, **E-F)** Number of SAVs categorised as Effect/Deleterious (magenta) or Neutral (pink) according to evolutionary conservation changes estimated by computational tools: PROVEAN, and SNAP2. The figures are generated using R statistical software version 4.0.2, ggplot2 package. Abbreviations used: ΔΔG: change in Gibbs free energy, SAV: single amino acid variation.

Evolutionary and structure-based predictors provide different insights into understanding mutational impact. Mutational impact in this context is considered to be its effect on protein stability, drug binding affinity, other binding affinities such as PPI or nucleic acid, and functional effects arising

from protein sequence variations. The first three mutational consequences are assessed by structure based predictors relying on the 3D structure of a protein, while the last is assessed by sequence based predictors relying mainly on evolutionary conservation trends across many proteins using multiple sequence alignments. The sequence based predictors are aimed at predicting pathogenicity or change of molecular function, structure based tools rely on estimating variant effects in relation to structure damage, corresponding to stability changes, as protein stability is considered the basic characteristic affecting function, activity, and regulation. Predictors such as ConSurf are able to use both structural and sequence information to identify important functional regions conserved in proteins. A variant classified as 'deleterious' to protein conservation may display gain-of-function in the presence of a drug through optimised protein stability. Thus, different methodological strategies benefit from complementary information when assessing specific proteins.

Sites involving EMB and DPA were mostly destabilising while those interacting with CDL were mostly stabilising

When assessing the impact on protomer stability changes due to mutations, the estimates from all four tools employed: mCSM-DUET, FoldX, DeepDDG, and Dynamut2 were considered together and averaged to provide a consensus mutational effect **(Figure 6G)**. While most (n=645) mutational effects were destabilising, most sites with 2 or more mutations showed mixed stability effects. Of such mutations, sites interacting with CDL at the PPI, mutations V456, V554, S658, and S659 were purely stabilising in their impact for all mutations observed (except V456 displaying 3 SAVs, others were budding resistant hotspots) **Figure 7**, sites marked in navy blue**)**. The impact of mutations were mainly stabilising (2 out of 3) for residues involved in Ca^{2+} ion interactions with site T956 showing only mildly destabilising impact **(Figure 7**, sites marked in purple**)**. Where all four mutations at site M306 involved in EMB interaction were associated with moderate-strong destabilising effects, residues nearby, at sites 305 and 311, each had 3 out of 4 mutations with stabilising effects **(Figure 7**, top panel, near the first two sites marked green**)**. Similarly, another interesting site 328 near to the bound EMB showed 6 SAVs, of which 4 were stabilising **(Figure 7**, top panel).

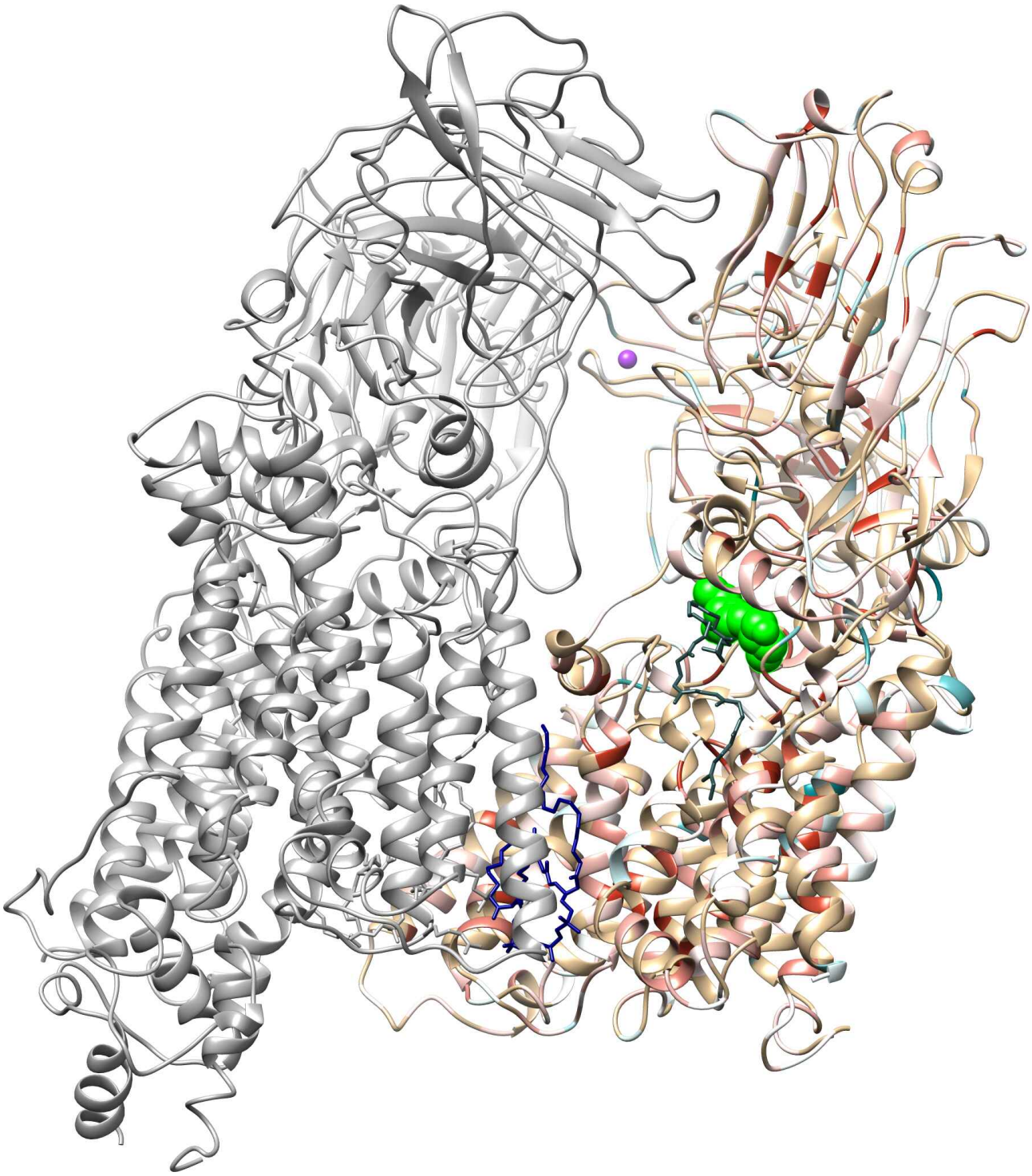


Figure 6: Average protein stability effects of SAVs mapped onto the *M. tuberculosis* EmbB protein structure

The protein stability changes (ΔΔG Kcal/mol) of SAV mutations measured by mCSM-DUET, FoldX, DeepDDG, and Dynamut2 were averaged and mapped onto EmbB sites (appearing as tan coloured ribbon). Destabilising mutational sites are depicted in red and stabilising mutational sites appear in blue, where the colour intensity reflects the extent of effect, ranging from -1 (most destabilising) to +1 (most stabilising). EMB is shown in green spheres in the binding pocket, while other binding partners are coloured as sticks in dark slate grey (DPA), navy blue (CDL), and Ca^{2+} ion (purple). The figure is rendered using UCSF Chimera version 1.14. Abbreviations used: ΔΔG: change in Gibbs free energy, SAV: single amino acid variation, EMB: ethambutol, DPA: decaprenyl-phosphate-arabinose, CDL: cardiolipin, Ca^{2+} ion: calcium ion.

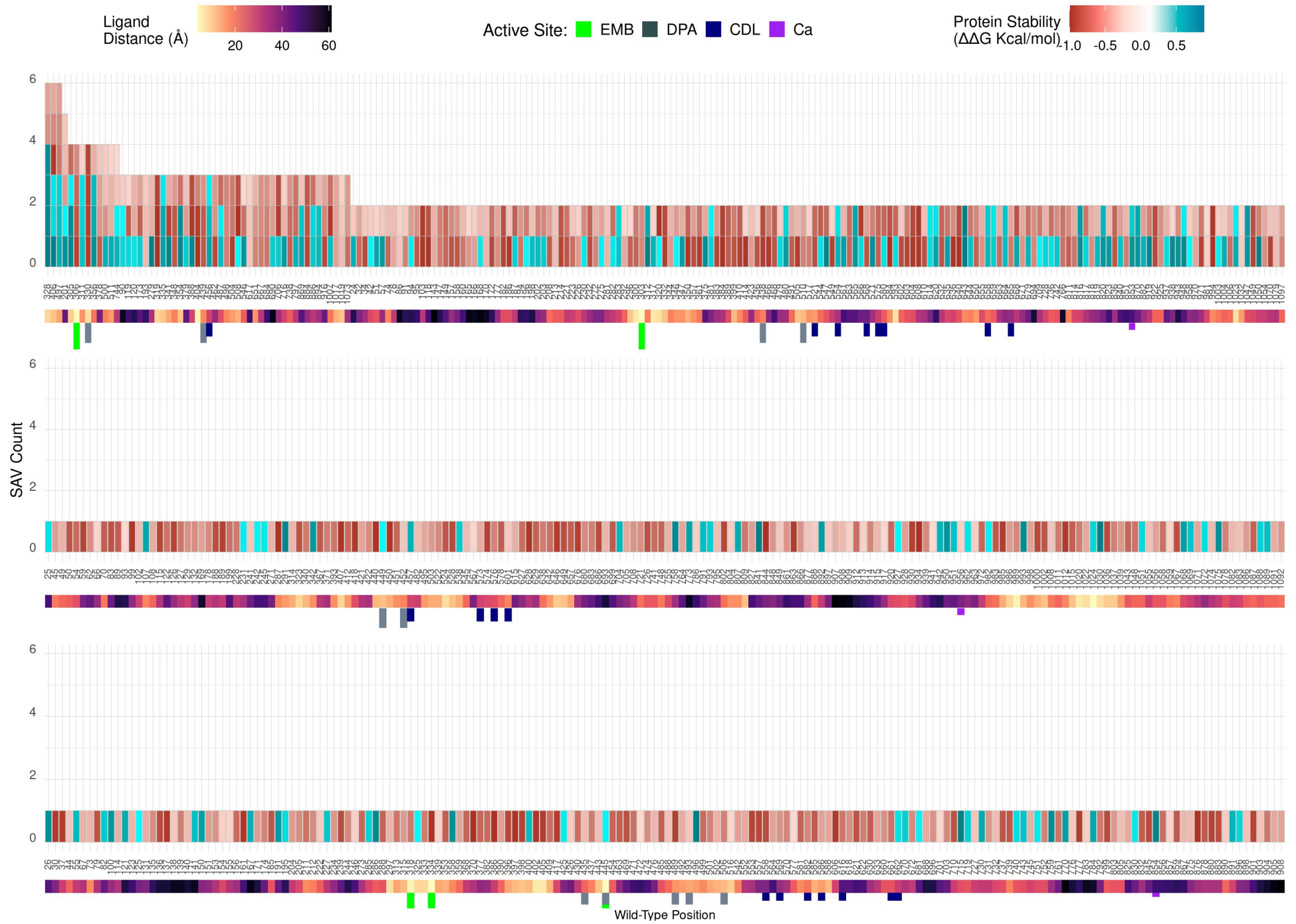

Ligand Distance (Å)
20
40
60
Active Site:
EMB
DPA
CDL
Ca
Protein Stability (ΔΔG Kcal/mol)
-1.0
-0.5
0.0
0.5
SAV Count
Wild-Type Position

Figure 7: Average protein stability effect for individual SAVs occurring in *M. tuberculosis embB*
Barplot showing the number of single amino acid variation (SAV) mutation at each position in EmbB coloured by the average protein stability effect, where the horizontal axis shows the wild-type positions associated with SAVs, and the vertical axis shows the number of SAVs at that position. The horizontal axis is ordered to highlight wild-type positions with the highest number of SAVs. For a given position, each corresponding SAV is coloured by the average protein stability effect calculated across estimates (ΔΔG Kcal/mol) from mCSM-DUET, FoldX, DeepDDG, and Dynamut2. The structural positions associated with SAVs in EmbB are indicated on the horizontal axis. The heat bar underneath the positions indicates the distance of that position from EMB according to the magma colour gradient where light yellow indicates sites closer to EMB (ligand distance in Angstroms). The positions are further annotated to reflect active site residues involved in interactions with: EMB (green), DPA (dark slate grey), CDL (navy blue), and Ca^{2+} ion (purple). The figure is generated using R statistical software version 4.0.2, ggplot2 package. Abbreviations used: ΔΔG: change in Gibbs free energy, SAV: single amino acid variation, EMB: ethambutol, DPA: decaprenyl-phosphate-arabinose, CDL: cardiolipin, Ca^{2+} ion: calcium ion.

4.2.3 Mutational consequences on affinity changes and prominent mutational effects

Mutations decrease binding affinity of EMB while increasing affinity at the PPI

Only 5% (n=47) of SAVs inducing changes in EMB binding affinity were within 10Å of EMB. These mutations occurred at 23 distinct sites, with most sites showing single mutations. Of these, over 90% (n=44) had a destabilising effect on EMB binding affinity as measured by mCSM-lig and all 47 mutations were destabilising when measured by mmCSM-lig **(Figure 8A** top panel, Appendix Table 4.A.1**)**. When the 23 mutational sites with their average effect on binding affinity were mapped onto the EmbB chain B, these showed mild to moderate destabilising mutational consequences **(Figure 8A** bottom panel. Analysing the PPI of EmbB highlighted 14% (n=121) of mutations to be within 10Å of the PPI as measured by mCSM-PPI2, with 65% (n=79) of mutational effects being destabilising (**Figure 8B** top panel, Appendix Table 4.B.1). Interestingly, sites at the PPI showed mixed mutational effects on binding affinity and were distributed throughout the entire interface **(Figure 8B** bottom panel**)**.

Of the total 570 unique sites in EmbB displaying SAVs, 62% (n=355) of these sites harboured single mutations, followed by 28% (n=162) sites presenting as budding resistant hotspots, followed by 40 sites displaying 3 mutations, 9 sites displaying 4 mutations, 1 site showing 5 mutations, and 3 sites showing a maximum of 6 mutations **(Figure 8C** top panel**)**.

The most prominent effects on EMB binding were from reduced affinity (destabilising effect), contributed by 38 mutations **(Figure 8C**, yellow text boxes, and bottom panel**)** at surrounding sites. Similarly, the dimer interface of EmbB was principally affected by affinity changes at the interface with the majority (n=22) of the effects resulting in stabilising i.e. increasing PPI affinity (**Figure 8C**, pink text boxes, and bottom panel). All other sites were affected largely (but not exclusively) by desta-

bilising mutations for EmbB protein structure (**Figure 8C**, blue and red text boxes, and bottom panel).

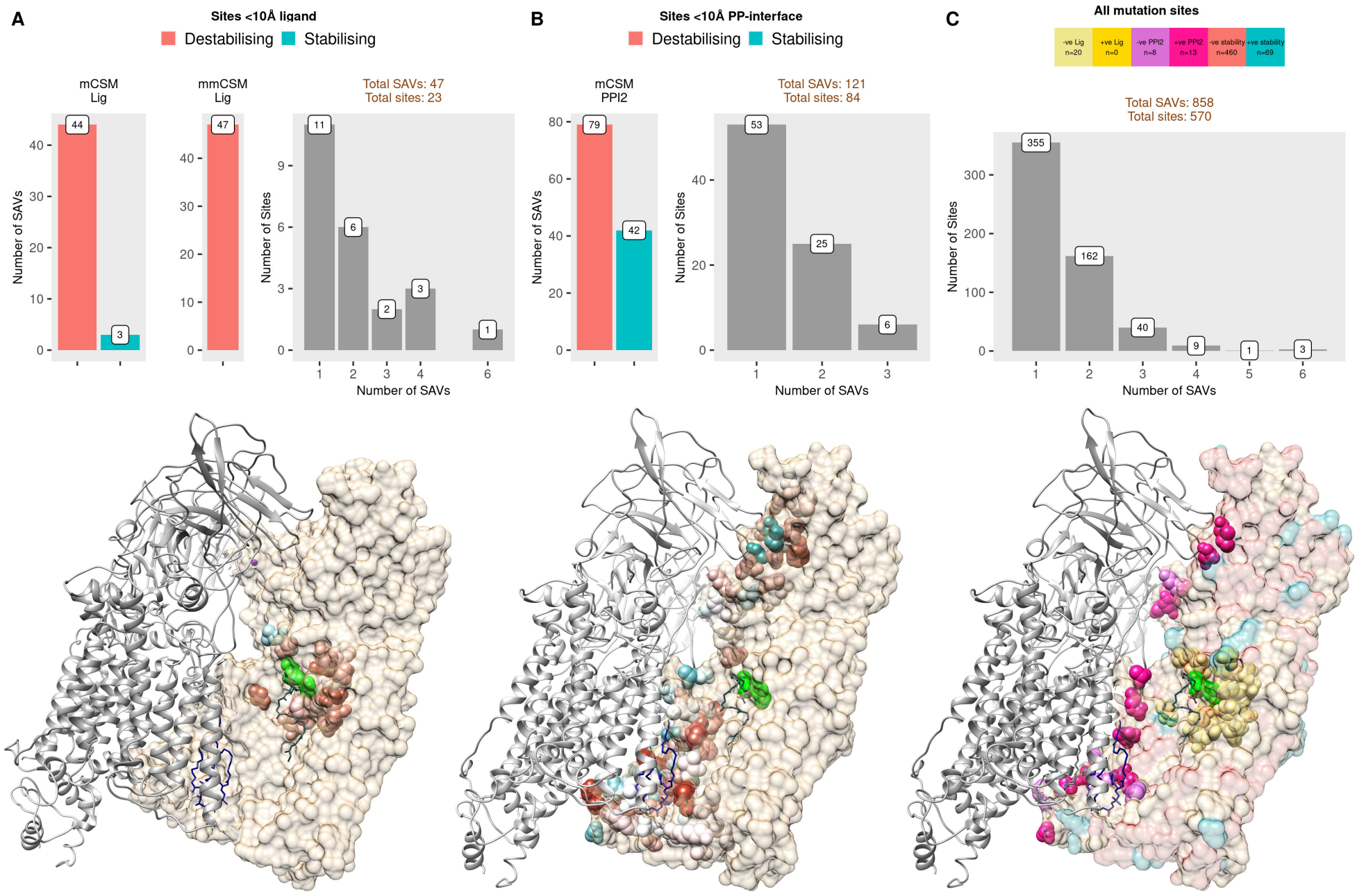

A
Sites <10Å ligand
Destabilising
Stabilising
mCSM Lig
mmCSM Lig
Total SAVs: 47
Total sites: 23
Number of SAVs
Number of Sites
B
Sites <10Å PP-interface
mCSM PPI2
Total SAVs: 121
Total sites: 84
C
All mutation sites
-ve Lig n=20
+ve Lig n=0
-ve PPI2 n=8
+ve PPI2 n=13
-ve stability n=460
+ve stability n=69
Total SAVs: 858
Total sites: 570

Figure 8: Mutational impact on EMB binding affinity, protein-protein interaction on EmbB, and sites with the most prominent mutational effects within *M. tuberculosis* EmbB
The top panel displays barplots showing the mutational outcome of affinity changes and their corresponding site frequency, while the bottom panel shows the corresponding mutational impact mapped onto EmbB (chain B, PDB-ID: 7BVF) appearing in tan colour, while chains A and P are shown as grey ribbons. EMB is shown in green as spheres in the binding site. Other binding partners are indicated as sticks: DPA in dark slate grey, and CDL is shown in navy blue. **A)** Mutational impact on EMB binding affinity (log fold change) from mCSM-lig and mmCSM-lig where 47 mutations, corresponding to 23 sites within 10Å of EMB, **B)** Mutational impact on protein-protein (PP) binding affinity ($\Delta\Delta$G) for 121 mutations, corresponding to 84 sites within 10Å of the PPI. For both parts A) and B), red denotes destabilising mutational sites while blue denotes stabilising mutational sites, and the colour intensity reflects the extent of the effect ranging from -1 (most destabilising) to +1 (most stabilising), **C)** Most prominent mutational effect for all 858 SAVs (corresponding to 570 sites) prioritised in order of increasing effect size: mCSM/mmCSM-lig, mCSM-PPI2, followed by overall stability changes where brighter colours indicate stabilising effects. Sites marked in yellow indicate changes due to ligand (EMB) binding affinity with lighter yellow indicating destabilising changes, pink areas indicate prominent changes due to PPI affinity where bright pink indicates stabilising while light pink areas indicate destabilising effects. All other sites are coloured by overall stability changes where blue denotes stabilising and red denotes destabilising effects. The corresponding number of mutation sites contributing to the different effect types are indicated in the text box at the top, and coloured accordingly. The barplots figures are generated using R statistical software version 4.0.4, ggplot2 package. The structure figures are generated using Chimera version 1.14. Abbreviations used: Å: Angstroms, $\Delta\Delta$G: change in Gibbs free energy in kcal/mol, SAV: single amino acid variation, EMB: ethambutol, DPA: decaprenyl-phosphate-arabinose, CDL: cardiolipin.

4.2.4 Mutational association with EMB resistance and flexibility

Most mutations occur in the variable regions with resistant mutation sites associated with lower flexibility relative to sites with sensitive mutations.

Mutational association with resistance according to aggregate DST data showed only a minority (14%, n=127) of mutations as resistant. Mutational sites on EmbB were mapped onto the 3D structure to highlight sites with exclusively resistant (red), sensitive (blue) and sites displaying both resistant and sensitive mutations (purple). For EmbB, there were 54 sites with exclusively resistant mutations, 46 sites with both resistant and sensitive mutations, while 470 sites with exclusively sensitive mutations **(Figure 9A)**.

While there were some resistant mutations close to EMB and the PPI, mutations were not restricted to these areas, with resistant mutations occurring as far away as 57Å from the drug **(Figure 9A)**. This is perhaps due to EmbB being part of a larger multimeric complex, such that sites far away in EmbB are closer to the binding site on another chain of the protein complex. As such, interactions in a multimeric protein become important to consider when understanding mutational association with resistance. ConSurf scores are calculated for each site on the protein, and range from 1 (rapidly evolving, variable sites) to 9 (slowly evolving, conserved sites). The resistant mutational sites surrounding EMB were not found to be located in the conserved regions of EmbB, as defined by ConSurf, with sensitive mutational sites predominantly located in the variable regions of EmbB **(Figure 9A** left and right panels**)**. Exclusively resistant mutation sites **(Figure 9A)** were not restricted in the conserved regions

of EmbB **(Figure 9B** left panel**)** with most mutations (n=206) occurring in the highly variable regions (ConSurf score 1) of EmbB **(Figure 9B** right panel**)**.

The local flexibility in EmbB in relation to EMB resistance was also analysed with thickness of the ribbon/tube (thinthick) indicating the extent of flexibility. Normal mode analysis (from Dynamut2) of the protein component of EmbB-EMB complex showed that overall EmbB displayed low-to-moderate flexibility (**Figure 10** left panel**)**. Visual inspection highlighted that sites with sensitive mutations were comparatively more flexible that those with resistant mutations. **(Figure 10** left panel**)**. The most prominent sites displaying the highest flexibility (thickest tubes) were: T643 (site with both resistant and sensitive mutations), M423 (site with exclusively resistant mutations), while G645 and I649 were sites with no SAVs. Similarly, regions surrounding EMB, as well as DPA and CDL were associated with low flexibility **(Figure 10** right panel**)**.

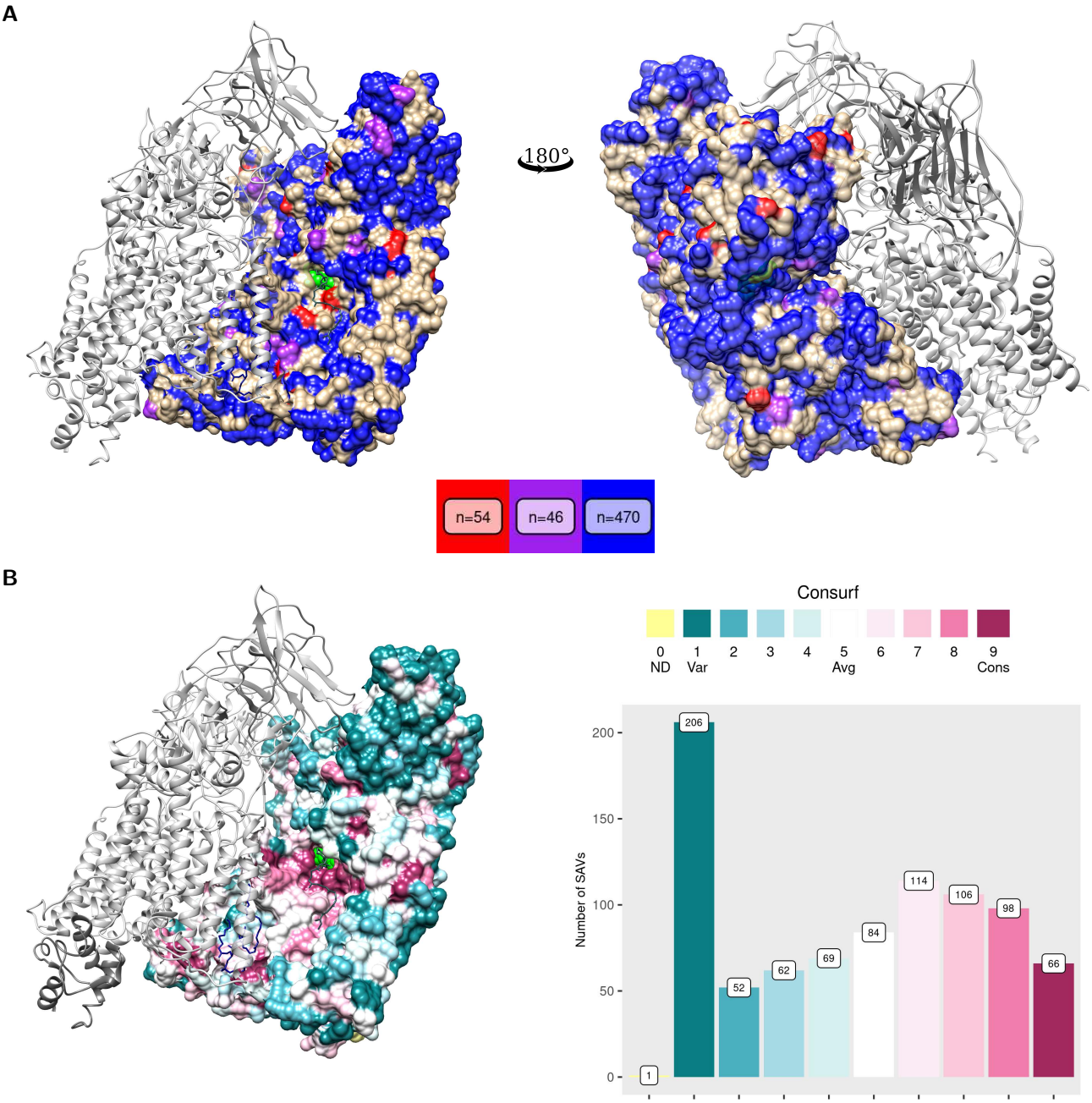


Figure 9: Mutational association with ethambutol resistance and evolutionary conservation in *M. tuberculosis* EmbB

Mutational landscape of *M. tuberculosis* EmbB according to different measures where **A)** All sites associated with SAVs on EmbB (chain B, PDB-ID: 7BVF) shown as surface representation in tan colour, along with chains A and P appearing as grey ribbons. EMB is shown in green either represented as spheres or ball-and-stick to aid visibility. The left panel shows all mutational sites associated with resistant (red, n=54 sites), sensitive (blue,n=470 sites), while common sites with both resistant and sensitive mutations appear in purple (n=46). The corresponding right panel depicts the structure rotated by 180°, **B)** Left panel shows EmbB chain B coloured according to ConSurf scores where maroon indicates conserved sites and teal indicates variable sites. EMB appears in green in the conserved binding pocket. Yellow areas reflect sites with uncertainty due to insufficient data for ConSurf score calculation. The barplot on the right panel shows the number of mutations associated with ConSurf values that range from 1 (variable) in teal to 9 (conserved) in maroon, where 0 denotes insufficient data/not defined (ND). The barplot figures are generated using R statistical software version 4.0.4, ggplot2 package. All structure figures were generated using UCSF Chimera version 1.14. Abbreviations used: SAV: single amino acid variation, EMB: ethambutol, DPA: decaprenyl-phosphate-arabinose, CDL: cardiolipin, Ca^{2+} ion: calcium ion.

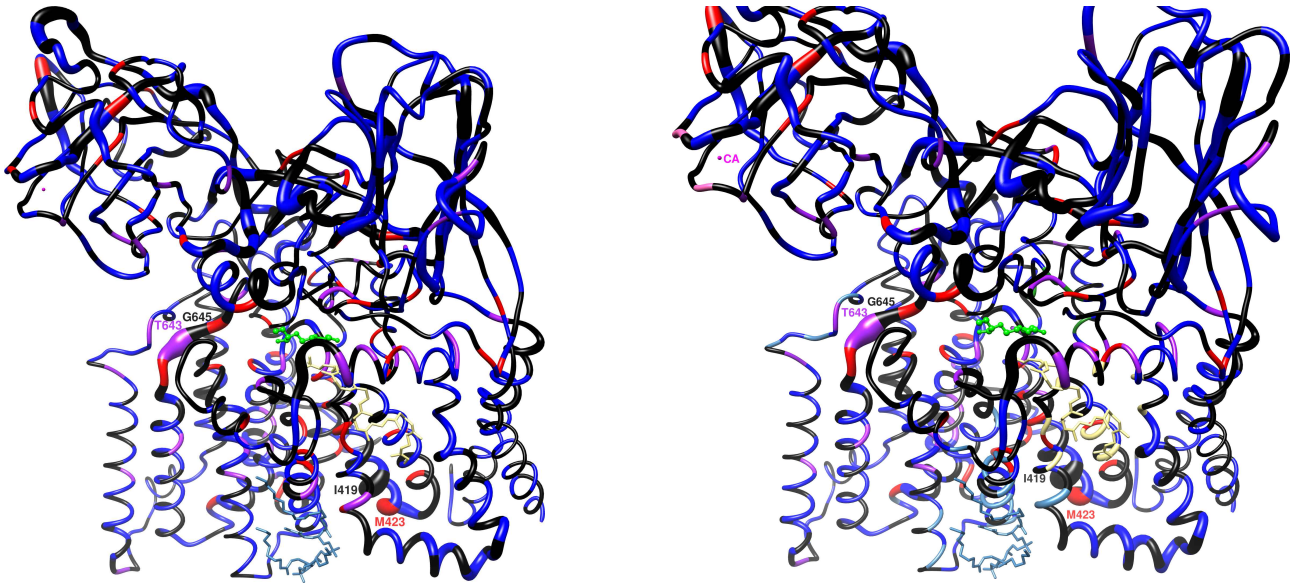


Figure 10: Mutational association with ethambutol resistance and local protein flexibility of *M. tuberculosis* EmbB

Mutational landscape of *M. tuberculosis* EmbB according to flexibility in EmbB according to normal mode analysis (NMA), measuring atomic deformation according to protein dynamics to denote flexibility associated at sites in EmbB. The magnitude of flexibility is represented from thin (low flexibility) to thick (high flexibility) tubes. Left panel: The tubes are further coloured to show mutational association with EMB resistance, red: resistant sites, blue: sensitive sites, purple: shared sites, black: sites with no SAVs, where sites with the highest flexibility (thickest tubes) are labelled according to the wild-type residues using the standard one-letter amino acid code. Right panel: Slightly zoomed in view to indicate EMB, DPA, CDL, and Ca^{2+} ion as well as their interacting residues sites in green, light yellow, steel blue, and light pink respectively. Similar to the left panel, the residues are labelled to indicate sites with the highest flexibility (thickest tube) according to the standard one-letter amino acid code. All structure figures were generated using UCSF Chimera version 1.14. Abbreviations used: SAV: single amino acid variation, EMB: ethambutol, DPA: decaprenyl-phosphate-arabinose, CDL: cardiolipin, Ca^{2+} ion: calcium ion.

4.2.5 Relating mutational frequency and biophysical and evolutionary conservational changes

Correlation analysis was performed to understand the relationship between frequently occurring mutations as assessed by MAF and their association with stability (mCSM-DUET, FoldX, DeepDDG, Dynamut2), conservation (ConSurf, SNAP2, PROVEAN) and affinity changes (mCSM-lig/mmCSM-lig, and mCSM-PPI2), distance to ligand (Lig-Dist) and protein-protein interface (PPI-Dist). A combined analysis with all mutations, as well as separately for resistant (R) and sensitive (S) mutations was undertaken **(Figures 11 and 12)**. Analyses focused on determining the strength of association without regard for the direction of the association due to dissimilarity of threshold criteria used by the various estimators.

Frequently occurring sensitive mutations were weakly related to protomer stability changes while frequently occurring resistant mutations were moderately associated with decreasing distance from EMB

Frequently occurring mutations were weakly related to protomer stability changes according to DeepDDG (ρ_{R+S}=0.14, P<0.001), but not associated with FoldX (ρ_{R+S}=-0.07, P<0.05), Dynamut2 and

mCSM-DUET (ρ_{R+S}=0, P>0.05) **(Figure 11)**. The weak associations with DeepDDG estimates were driven by sensitive mutations (ρ_S=0.19, P<0.001 for sensitive mutations) suggesting that frequently occurring sensitive mutations were weakly associated with protomer stability changes **(Figure 11)**. Frequently occurring resistant mutations were moderately associated with decreasing distance from EMB (ρ_R~0.3, P<0.01), and weakly related to PPI (ρ_R=0.22, P<0.05) **(Figure 11)**. As expected, mCSM-DUET and Dynamut2 were strongly correlated as these tools share common methodology (ρ_{R+S}=0.82, P<0.001), while other computational tools showed weak to strong associations amongst their predicted estimates, overall as well as for resistant and sensitive mutation groups individually (0.3< ρ_{R+S}≤0.8, P<0.001) **(Figure 11)**. Of note, the negative sign associated with FoldX correlations with other predictors is due to the inverse criteria used by these tools (Chapter 2: Methods).

Frequently occurring resistant mutations were weakly associated with evolutionary rate, while frequently occurring sensitive mutations were weakly associated with protein functional effects

Frequently occurring resistant mutations were weakly related to evolutionary conservation estimates (ρ_R=-0.19, P<0.05) **(Figure 12** left panel**)**, while frequently occurring sensitive mutations were weakly associated with predicted protein functional effects according to SNAP2 (ρ_S=-0.24, P<0.001) and PROVEAN (ρ_S=0.21, P<0.001). There was good agreement (moderate to strong association) between estimates across the three conservation estimators, overall (ρ_{R+S}≥0.5, P<0.001) and in the mutation groups ($\rho_{R/S}$>0.4, P<0.001) (**Figure 12** left panel**)**.

Frequently occurring mutations were weakly related to EMB affinity changes

Frequently occurring resistant mutations were weakly related to EMB binding affinity (mCSM- and mmCSM-lig) (ρ_R≥-0.24, P<0.01), and not associated with changes in PPI binding (mCSM-PPI2) ($\rho_{R/S}$<-0.1, P>0.05) **(Figure 12** right panel**)**. As expected, estimates from mCSM- and mmCSM-lig were highly correlated, both, overall and in the mutation groups (ρ>0.8, P<0.001) due to shared underlying methodology **(Figure 12** right panel**)**.

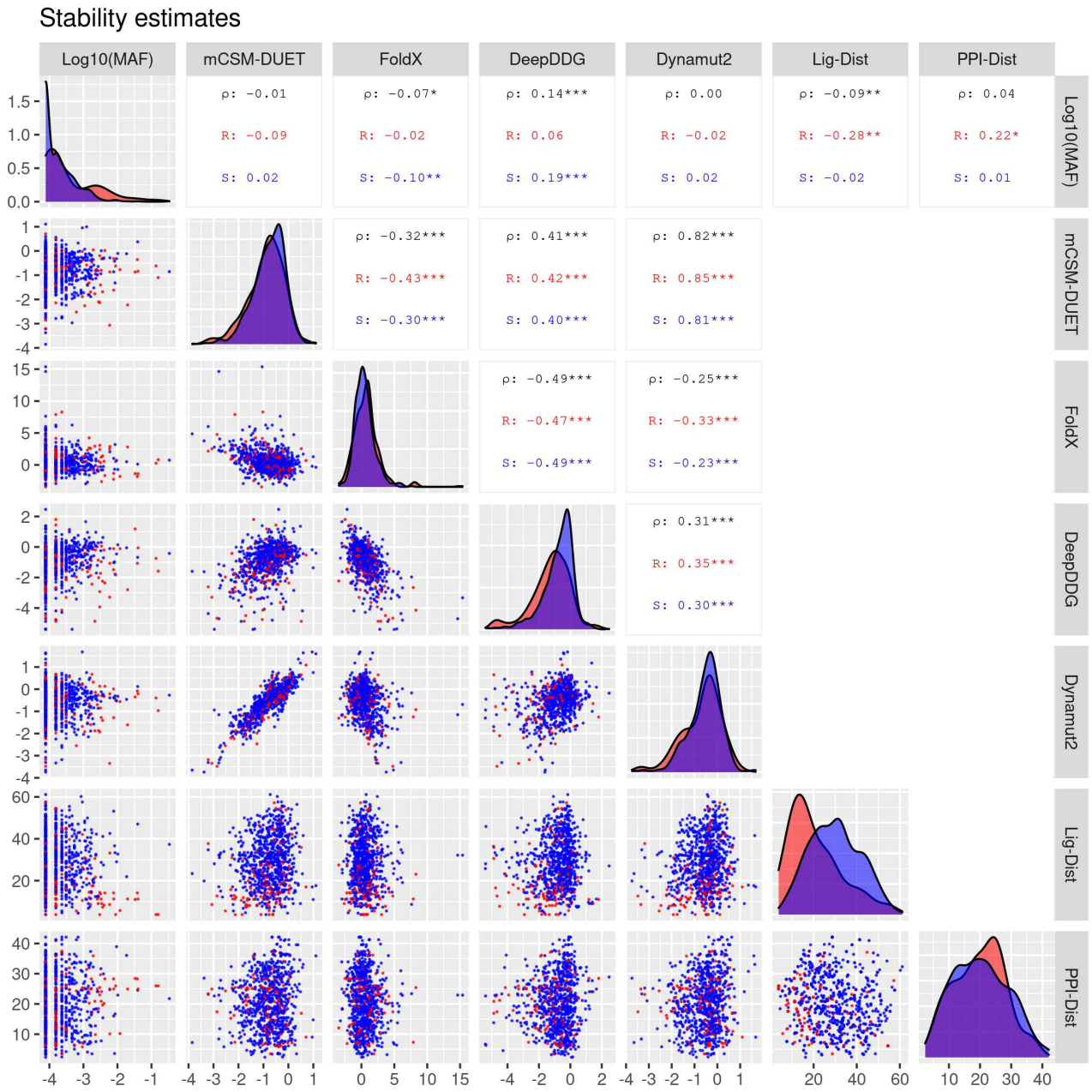


Figure 11: Correlation of protein stability changes and genomics measures
Pairwise correlations between minor allele frequency (MAF), protein stability changes ($\Delta\Delta$G) estimated using DUET, FoldX, DeepDDG, and Dynamut2, and distance to EMB, and the dimer interface for 858 SAVs. The upper panel in both plots include the pairwise Spearman (ρ) correlation values along with their statistical significance (.P<0.10, *P<0.05, **P<0.01, ***P<0.001). Three correlation values appear in each plot where black denotes the overall correlation with both resistant (R) and sensitive (S) mutations, while red denotes correlation estimates for resistant mutations, and blue denotes correlation estimates for sensitive mutations. The points in the lower panel represent SAVs, where red dots denote resistant mutations and blue represent sensitive mutations individually. The diagonal in each plot displays the density distribution of the corresponding parameter split by the two mutation groups. The figure is generated using R statistical software version 4.0.4, ggplot2 package. Abbreviations used: Å: Angstroms, $\Delta\Delta$G: change in Gibbs free energy in Kcal/mol, SAV: single amino acid variation, Lig-Dist: distance to ligand in Å, PPI-Dist: distance to protein-protein interface in Å, EMB: ethambutol.

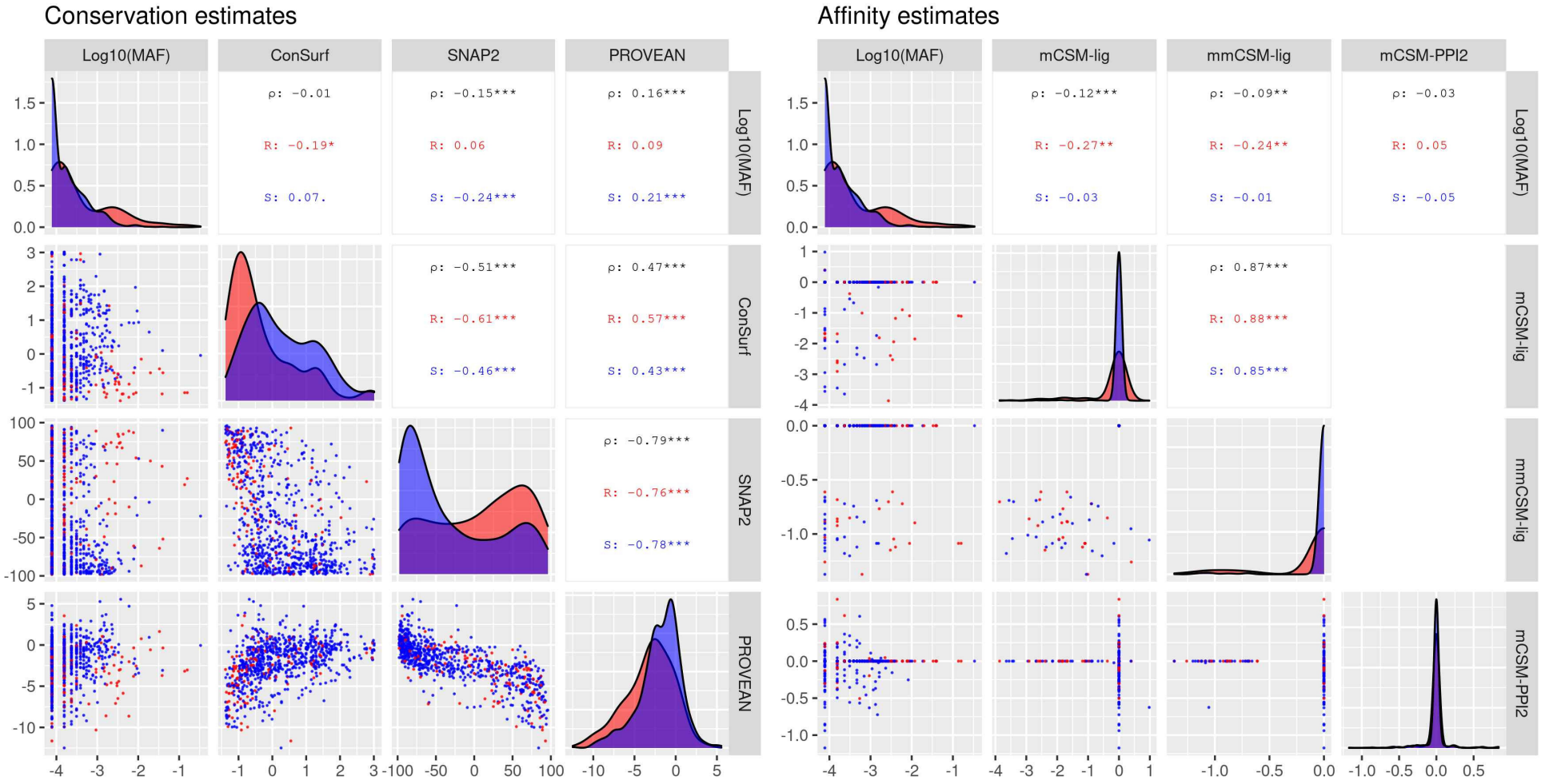


Figure 12: Correlation of evolutionary conservation, affinity changes, and genomics measures

Pairwise correlations of evolutionary conservation, affinity changes, and genomic measure of minor allele frequency (MAF) for 858 SAVs. **Left panel:** Evolutionary conservation predictors: ConSurf, SNAP2, and PROVEAN, **Right panel:** EMB binding affinity changes estimated as log fold change (mCSM-lig and mmCSM-lig) of 47 SAVs lying within 10Å of EMB, and protein-protein (PP) affinity changes (ΔΔG) measured using mCSM-PPI2 of 121 SAVs lying within 10Å of the dimer interface. All corresponding affinity measures for mutations located more than 10Å of EMB, and the PPI were given a value of 0 to allow complete SAVs to be used for analysis, while respecting the distance threshold for the respective tools. The upper panel in both plots include the pairwise Spearman (ρ) correlation values along with their statistical significance (.P<0.10, *P<0.05, **P<0.01, ***P<0.001). Three correlation values appear in each plot where black denotes the overall correlation with both resistant (R) and sensitive (S) mutations, while red denotes correlation estimates for resistant mutations, and blue denotes correlation estimates for sensitive mutations. The points in the lower panel represent SAVs, where red dots denote resistant mutations and blue represent sensitive mutations individually. The diagonal in each plot displays the density distribution of the corresponding parameter split by the two mutation groups. The figure is generated using R statistical software version 4.0.4, ggplot2 package. Abbreviations used: Å: Angstroms, ΔΔG: change in Gibbs free energy in Kcal/mol, SAV: single amino acid variation, Lig-Dist: distance to ligand in Å, PPI-Dist: distance to protein-protein interface in Å, EMB: ethambutol.

4.2.6 Comparing resistant and sensitive mutations

Resistant mutations occur less frequently, are more conserved and tend to be destabilising for protomer stability, located closer to EMB binding site without affecting binding affinity

Resistant mutations were more destabilising for changes in overall protomer stability compared with sensitive mutations but only according to DeepDDG (P<0.001) **(Figure 13C)**, and not according to mCSM-DUET, DeepDDG, and Dynamut2 (P>0.05) **(Figures 13A, 139C, 139D)**. Furthermore, frequently occurring mutations are also less likely to be resistant (P<0.0001) **(Figure 13E)**. Resistant mutations were closer to EMB binding site (P<0.001) **(Figure 13F)** without affecting binding affinity (P>0.05) **(Figure 13K, 13L)**. Resistant mutations were also not close to the PPI (P>0.05) **(Figure 13G)** and did not result in changes in PPI affinity (P>0.05) **(Figure 13M)**. Resistant mutations are conserved and predicted to result in deleterious effects (P<0.001) **(Figures 13H, 13I and 13J)**.

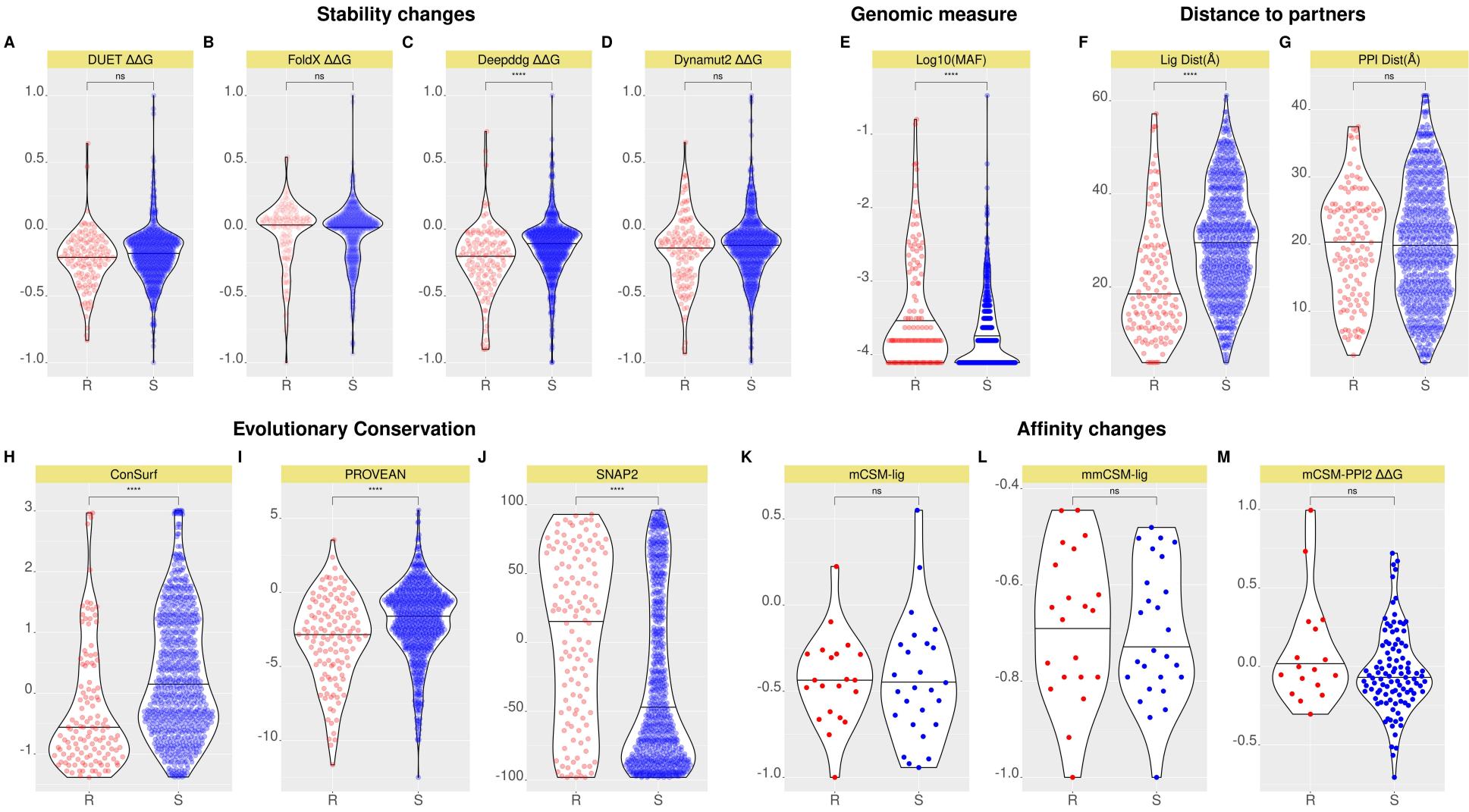

Stability changes
Genomic measure
Distance to partners
A
DUET ΔΔG
B
FoldX ΔΔG
C
Deepddg ΔΔG
D
Dynamut2 ΔΔG
E
Log10(MAF)
F
Lig Dist(Å)
G
PPI Dist(Å)
Evolutionary Conservation
Affinity changes
H
ConSurf
I
PROVEAN
J
SNAP2
K
mCSM-lig
L
mmCSM-lig
M
mCSM-PPI2 ΔΔG
ns
****
R
S

Figure 13: Comparison of resistant (R) and sensitive (S) mutations
Violin plots showing the distribution of features related to structural properties, genomic measure, evolutionary conservation for 858 SAVs. For affinity changes related to ligand (EMB) binding affinity measured by mCSM- and mmCSM-lig, only those mutations within 10Å of EMB (n=47) were considered. Similarly, for protein-protein (PP) affinity changes measured by mCSM-PPI2, only those mutation within 10Å of the PPI (n=121) were analysed. Mutations were grouped either as resistant (R, n=127) or sensitive (S, n=731) and were compared using the Wilcoxon rank-sum (unpaired) test, with statistical significance indicated as: *P < 0.05, **P < 0.01, ***P < 0.001, ****P < 0.0001). Mutations in the resistant group appear as red dots, while those in the sensitive group appear as blue dots, and the horizontal line in the violin plots display the median value. The two mutations groups were compared based on **A-D)** Stability changes ($\Delta\Delta$G) estimated from four computational tools: mCSM-DUET, FoldX, DeepDDG and Dynamut2, **E)** genomic measure of average mutational occurrence (Log10MAF), **F-G)** Distance to ligand (Lig-Dist) and Distance to the PPI (PPI-Dist), **H-J)** Evolutionary conservation measured by ConSurf (<0: Conserved, >0: Variable), PROVEAN (>-2.5: Neutral, < -2.5: Deleterious) and SNAP2 (<=0: Neutral, >0:Effect) computational tools, **K-L)** Comparison of EMB binding affinity changes from mCSM-lig and mmCSM-lig measured as log fold change for R (n=21) and S (n=26) mutations, and those for **M)** PP binding affinity changes (mCSM-PPI2) measured as $\Delta\Delta$G for R (n=17) and S (n=104) mutations. The figure is generated using R statistical software version 4.0.4. Abbreviations used: Å: Angstroms, $\Delta\Delta$G: change in Gibbs free energy in Kcal/mol, SAV: single amino acid variation, ns: not-significant, EMB: ethambutol, MAF: minor allele frequency, Lig-Dist: distance to ligand in Å, PPI-Dist: distance to protein-protein Interface in Å, R: resistant mutations, S: sensitive mutations.

4.2.7 Associating mutations with Odds Ratio and extreme effects

Mutations with high OR are not restricted to EMB active site

Based on DST data available for 614 (out of 858) SAVs, mutational association with resistance was further estimated using Odds Ratio (OR), with values above 1 suggesting association with EMB resistance. The higher the OR, the greater the likelihood of a given mutation being resistant. The majority (89%, n=549/614) of mutations were predicted to be associated with EMB resistance, much higher than observed in our data (14%, n=127/858). An overview of mutations associated with resistance show that sites with high OR are more prominent in areas close to EMB or residues interacting with DPA, Ca^{2+} ion, and to a lesser extent for CDL, although sites far away from EMB also show mutations with high OR **(Figure 14)**. Residues interacting with EMB as well as those surrounding it (positions between 303 and 306, and those between 319 and 334) were associated with OR>5 indicating the importance of sites surrounding EMB with detrimental consequences of mutations at these sites on EMB binding. Similarly, the region 405-409 also contained mutations with OR>10. All interacting partners for DPA except A439S were associated with OR>1 with Q445R, F330S, F330L showing OR>10. Both mutations at the Ca^{2+} ion interacting site: Q853R and Q853P were associated with OR>5. Sites at the dimer interface with CDL interacting residues showed less prominent OR compared with EMB, DPA and Ca^{2+} ion interacting sites **(Figure 14)** suggesting the important role of interface residues in maintaining the overall protein complex. Overall, these findings suggest the importance of sites surrounding, and distal to, EMB, along with interface residues. As such, resistance development appears to be mediated by compensatory mutational effects that alleviate fitness costs

at these important sites.

The top 10 mutations with the highest association with resistance were not related to residues interacting with EMB or other binding partners, with the exception of V456A (a site at the PPI interacting with CDL, OR=52.54, P<0.001). The SAV with the highest OR was at a hotspot site, Q497H (OR=52.64, P<0.0001) located >10Å from the EMB binding site and the PPI (**(Table 1)**). These were followed by A409P (OR=48.40), E405D (OR=38.18), G406A (OR=35.75), D328Y (OR=28.63). All P-values <0.0001. The next 4 in the list with OR>26 were T642A, N129D, R173L, and G1087D, but these were not statistically significant (adjusted P>0.05) **(Figure 14)**. Of the 38 mutations which occurred within 10Å of EMB, ∼29% (n=11) of mutations showed significant association with resistance (OR>1, P<0.05). The closest most significant mutation associated with resistance was Y334H (OR=7.90, P<0.0001), followed by M306V, M306L, and M306T (OR=22.87, OR=10.41, OR=10.34 respectively, P<0.0001) **(Figure 14)**.

However, when analysing the 84 mutations occurring within 10Å of the PPI, only 2% (n=2) of mutations were significantly associated with resistance. These were V456A (mentioned above) which was also the mutation with the second highest OR in the dataset, followed by T642A (OR=26.27, P<0.01). These results indicate that the burden of resistant mutations is less at the PPI of EMB compared with sites in proximity to EMB **(Figure 14**, Appendix Tables 4.A.1 and 4.B.1**)**.

Mutations with extreme effects are mainly located away from the active site

The most frequently occurring mutation, E378A (MAF ∼34%), as well as the most destabilising and stabilising mutations for PPI affinity (F676S and P690L respectively) are not involved with interactions in the active site related to EMB, DPA, CDL or Ca^{2+} ion **(Table 1)**. The single exception is mutation N318D (involved in EMB binding, and a single mutation at the site), which was responsible for the most destabilising effect on EMB binding affinity **(Table 1)**. This suggests that mutations with extreme effects in EMB do not involve the active site or binding partner residues, underscoring the importance of these sites. Consequently, mutational effects at key sites are thought to be milder, conferring local fitness advantages.

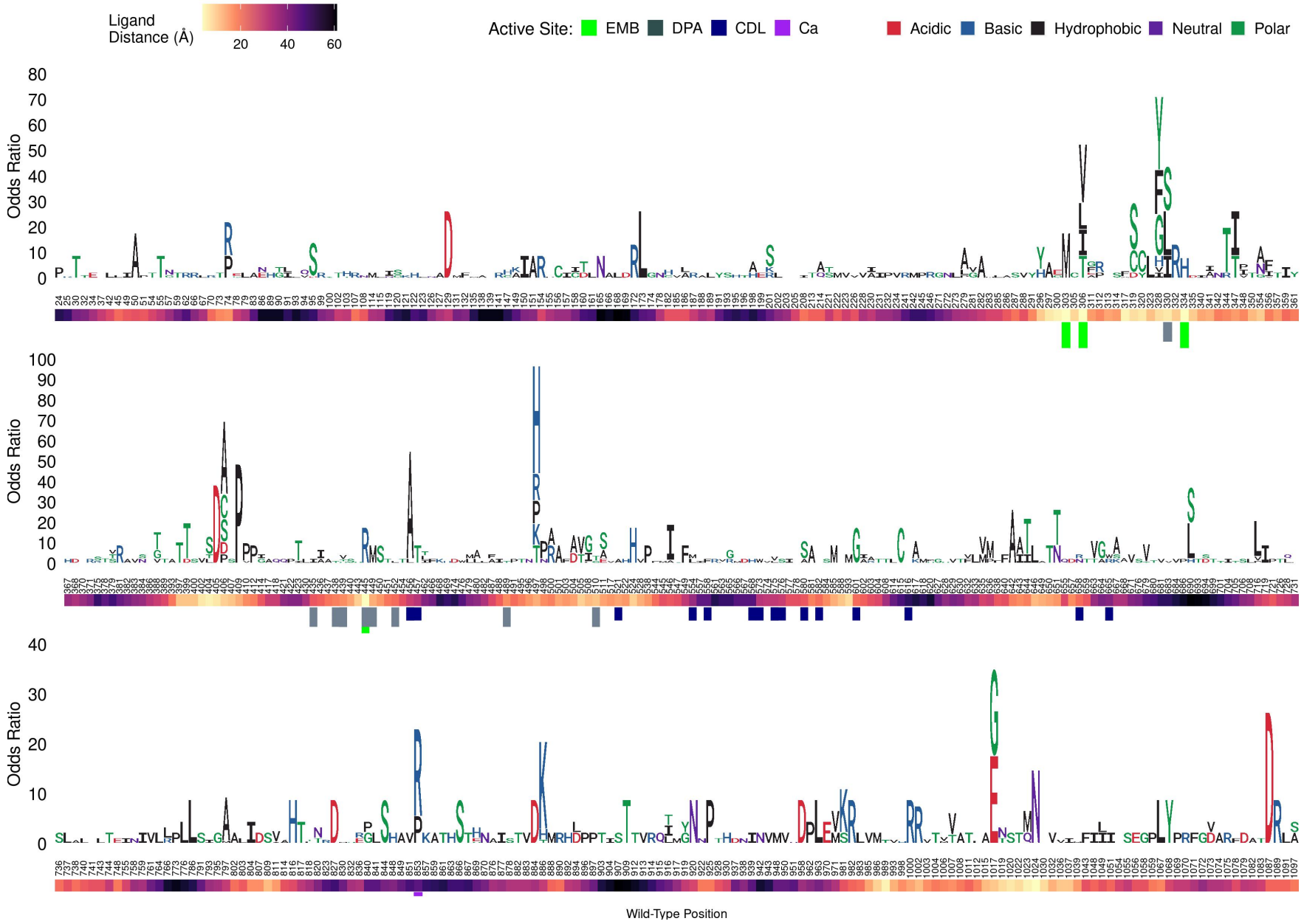

Ligand Distance (Å)
20
40
60
Active Site: EMB DPA CDL Ca
Acidic Basic Hydrophobic Neutral Polar
Odds Ratio
Odds Ratio
Odds Ratio
Wild-Type Position

Figure 14: Logo plot showing mutational sites and their association with resistance according to Odds Ratio

Logo plot showing 614 SAVs by mutational site according to their association with EMB resistance calculated using Odds Ratio (OR). The vertical axis represents the OR where letters denote mutant residues which are proportional to their corresponding OR, highlighting the most resistant mutation at each site and overall. The mutant residues are coloured according to the amino acid (aa) properties as denoted where red denotes acidic aa, basic aa appear in blue, hydrophobic aa in black, neutral aa in purple, and polar aa in darkgreen. The structural positions associated with SAVs with OR are indicated on the horizontal axis. The heat bar underneath the positions indicate the distance of that position from EMB according to the magma colour gradient where light yellow indicates sites closer to EMB (ligand distance in Angstroms). The positions are further annotated to reflect residues involved in interactions with EMB (green), DPA (dark slate grey), CDL (navy blue), and Ca^{2+} ion (purple). The figure is generated using R statistical software version 4.0.2, ggplot2 package. Abbreviations used: SAV: single amino acid variation, EMB: ethambutol, DPA: dcaprenyl-phosphate-arabinose, CDL: cardiolipin, Ca^{2+} ion: calcium ion.

Mutation	Mutational effect	Mutational effect value	Lig-Dist (Å)	PPI-Dist (Å)	Interacting partner
Q497H	Mutations with the highest OR	OR = 52.64	14.14	28.28	none
E378A	Most frequent mutation	MAF (%) = 33.52	37.34	21.77	none
V508G	Most Destabilising for protomer	ΔΔG = -0.56	12.65	17.39	none
H312P	Most Stabilising for protomer	ΔΔG = 0.50	17.61	35.90	none
N318D	Most Destabilising for EMB binding affinity	Log fold change = -0.90	4.66	22.51	EMB
F676S	Most Destabilising for PPI affinity	ΔΔG = -1.17	33.90	5.82	none
P690L	Most Stabilising for PPI affinity	ΔΔG = 0.84	54.47	6.13	none

Table 1: Mutations with extreme effects

Mutations (SAVs) with extreme effects related to Odds Ratio (OR), mutational frequency (MAF), stability and affinity changes. For affinity changes only mutations within 10Å of EMB for EMB binding affinity, and Protein-Protein (PP) interface for PPI affinity were considered. The protomer stability changes are the average effect of all four estimates (mCSM-DUET, FoldX, DeepDDG and Dynamut2) combined, and the EMB binding affinity changes are the average effect of the two mCSM based tools (mCSM-lig and mmCSM-lig) combined. Changes in PP affinity correspond to estimates from mCSM-PPI. The estimated effects were categorised as Destabilising (log fold affinity change/ΔΔG<0) and Stabilising (log fold affinity change/ΔΔG>0). Abbreviations used: ΔΔG: change in Gibbs free energy in Kcal/mol, MAF: minor allele frequency, SAV: single amino acid variation, Lig-Dist: distance to ligand, PPI-Dist: distance to protein-protein interface, EMB: ethambutol.

4.2.8 Relating lineage and protomer Stability

Lineages 1, 2 and 4 show comparable sample contribution while lineage 3 displays the highest SAV diversity

Around 44% of samples (n=15,526) consisted of SAVs in the protein coding region of EmbB, where 13,535 samples contributed to the four main *M. tuberculosis* lineages (Lineages 1-4). Most samples with EmbB mutations belonged to lineage 4 (n=5,276), followed by lineage 1 (n=4,013), lineage 2 (n=3,554) and finally by lineage 3 (n=692) with the smallest number of samples **(Figure 15A)**. However, lineage 3 was high in its SAV diversity (19%, n=189), followed by approximately similar SAV diversity for lineages 4 and 2 (8%, n=430 vs. 7%, n=241 SAVs respectively), with lineage 1 showing the least diversity of around 5% (n=189) **(Figure 15B)**.

Average stability distribution for mutations across the lineages was in the milder stability change estimates ($\Delta\Delta$G between +/- 0.3 Kcal/mol), highlighting that the mutational effects on stability for EmbB are not extreme **(Figure 15C)**. Resistant mutations for all lineages showed prominent peaks around mildly destabilising protomer stability changes ($\Delta\Delta G \sim$ -0.3 Kcal/mol). Resistant mutations were multimodal (>=3) for all lineages with two peaks around mildly destabilising ($\Delta\Delta$G -0.2 Kcal/mol and -0.3 Kcal/mol), with an additional peak around the mildly stabilising ($\Delta\Delta G \sim$0.2 Kcal/mol) protomer stability changes **(Figure 15C)**.

Sensitive mutations showed multiple peaks in all lineages except lineage 1. Lineage 1 displayed a single distinct peak for sensitive mutations with mildly destabilising effect on average protomer stability changes (-0.25 Kcal/mol$<\Delta\Delta G<0$). Lineage 2 showed a similar peak (-0.25 Kcal/mol$<\Delta\Delta G<0$ for sensitive mutations but spanned a wider range of stability estimates (-0.6 Kcal/mol$<\Delta\Delta G<$0.3 Kcal/mol). Lineage 3 however displayed a more distinct peak for sensitive mutations towards the mildly stabilising ($0<\Delta\Delta G<$0.25 Kcal/mol) protomer stability estimates, Lineage 4 showed showed a similar peak like lineage 3 for sensitive stabilising mutations with an additional peak around the mildly destabilising (-0.1 Kcal/mol$<\Delta\Delta G<0$) **(Figure 15C)**. Overall distributions for protomer stability changes were significantly different between all lineages (adjusted P<0.0001), as well as in lineages between resistant and sensitive mutation (adjusted P<0.0001) (Appendix Table 4.C.1).

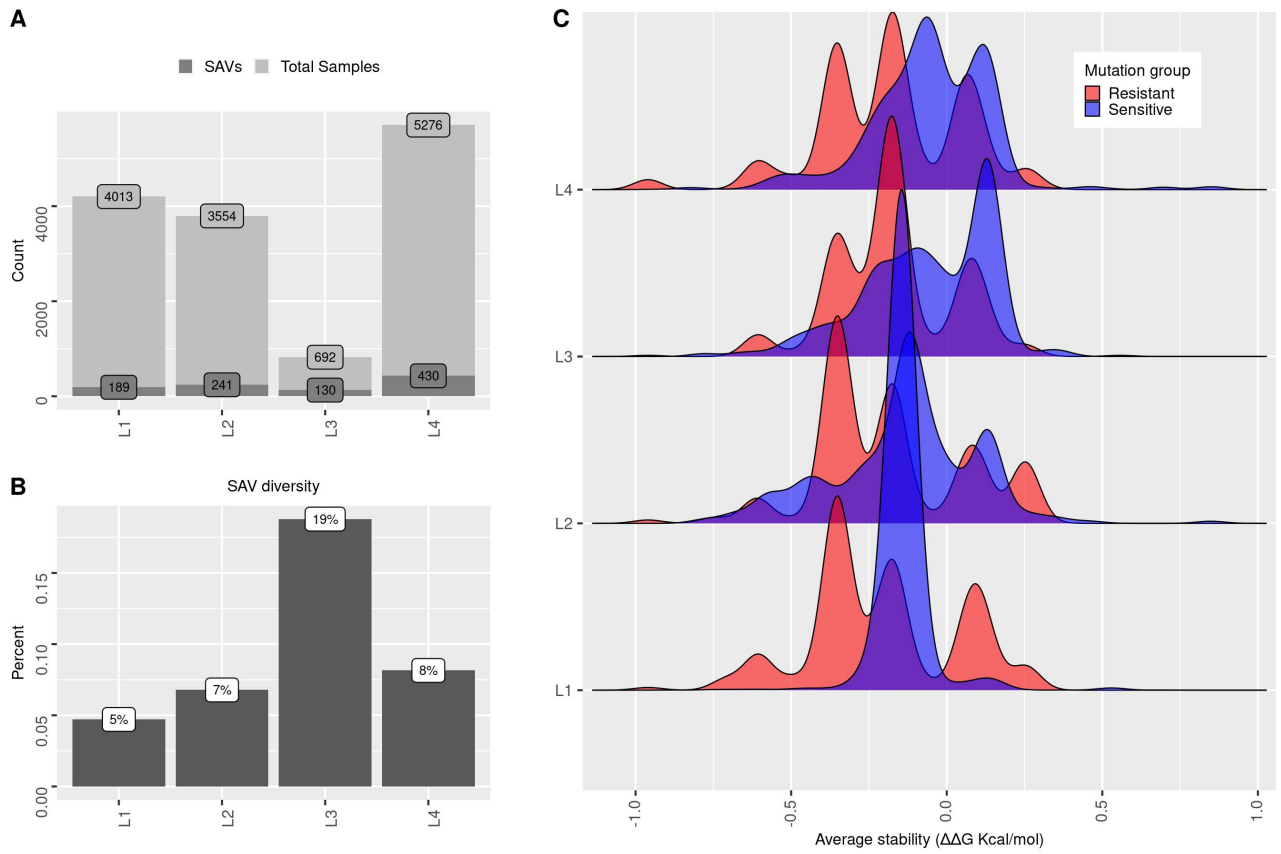


Figure 15: Lineage and protomer stability distribution
Total number of samples (n=13,535) along with the number of mutations associated with EMB resistance in the four *M. tuberculosis* lineages (L1-L4). **A)** The dark grey bars show the number of mutations (SAVs), while the light grey bar show the total number of samples in each lineage, **B)** Mutational diversity in each lineage, **C)** Density distribution of lineages according to protein stability changes (ΔΔG). Estimates from four different computational tools: mCSM-DUET, FoldX, DeepDDG,and Dynamut2 were combined to calculate the average mutational stability impact for each SAV. The horizontal axis shows the average stability values (-1: highly destabilising and +1: highly stabilising) further coloured by mutational association with EMB resistance: Red denotes resistant mutations (n=127, from 6,878 samples) and blue indicates sensitive mutations (n=731, from 6,657 samples). The figure is generated using the R statistical software version 4.0.4. Abbreviations used: ΔΔG: change in Gibbs free energy, SAV: single amino acid variation, EMB: ethambutol.

4.3 Chapter summary

With only a few active site residues having SAV mutations, and the majority of those being single SAV, sites involved with EMB binding are conserved and have limited tolerance for mutational heterogeneity. Most mutational consequences destabilise protomer stability without affecting function, and occur in variable regions, reinforcing its essential role in EMB binding. Due to EMBs competitive binding with the natural substrate (DPA), mutational effects involving EMB and DPA follow closely, e.g. M306 involved in EMB binding, and F330 involved in DPA binding, both display 4 SAVs with destabilising effects and the highest associations with EMB resistance. Similarly, EMB binding sites I303, and DPA interacting residues A438 and A510, each show 3 mutations with destabilising effects. Contrary to this, sites with CDL interactions tend to show the reverse, where most mutational impact is stabilising

for the PPI. This highlights that the resistance development in EmbB is a balanced interplay of the molecular interactions of the overall hetero-trimer complex, where flexibility in EmbB likely playing a role, though the mechanism is unclear. Resistance in EmbB appears underestimated in DST (14%), with GWAS inference predicting over 80% of mutations as resistant. As such, resistance hotspots can be located away from EMB with limited impact on its binding affinity.

References

[1] Qing Sun et al. "Mutations within embCAB Are Associated with Variable Level of Ethambutol Resistance in Mycobacterium Tuberculosis Isolates from China". In: *Antimicrobial Agents and Chemotherapy* 62.1 (Jan. 2018), e01279–17. ISSN: 1098-6596. DOI: 10.1128/AAC.01279-17.

[2] S. V. Ramaswamy et al. "Single Nucleotide Polymorphisms in Genes Associated with Isoniazid Resistance in Mycobacterium Tuberculosis". In: *Antimicrobial Agents and Chemotherapy* 47.4 (Apr. 2003), pp. 1241–1250. DOI: 10.1128/AAC.47.4.1241-1250.2003.

[3] E. Lederer. "The Mycobacterial Cell Wall". In: *Pure and Applied Chemistry. Chimie Pure Et Appliquee* 25.1 (1971), pp. 135–165. ISSN: 0033-4545. DOI: 10.1351/pac197125010135.

[4] N. Rastogi. "Recent Observations Concerning Structure and Function Relationships in the Mycobacterial Cell Envelope: Elaboration of a Model in Terms of Mycobacterial Pathogenicity, Virulence and Drug-Resistance". In: *Research in Microbiology* 142.4 (May 1991), pp. 464–476. ISSN: 0923-2508. DOI: 10.1016/0923-2508(91)90121-p.

[5] Jéssica D. Petrilli et al. "Differential Host Pro-Inflammatory Response to Mycobacterial Cell Wall Lipids Regulated by the Mce1 Operon". In: *Frontiers in Immunology* 11 (2020). ISSN: 1664-3224.

[6] *Ethambutol*. Vol. 88. Handbook of Anti-Tuberculosis Agents. Mar. 1, 2008. 102-105.

[7] Sylvie Garneau-Tsodikova and Kristin J. Labby. "Mechanisms of Resistance to Aminoglycoside Antibiotics: Overview and Perspectives". In: *MedChemComm* 7.1 (2016), pp. 11–27. DOI: 10.1039/C5MD00344J.

[8] Isabelle Vergne, Martine Gilleron, and Jérôme Nigou. "Manipulation of the Endocytic Pathway and Phagocyte Functions by Mycobacterium Tuberculosis Lipoarabinomannan". In: *Frontiers in Cellular and Infection Microbiology* 4 (2014), p. 187. ISSN: 2235-2988. DOI: 10.3389/fcimb.2014.00187.

[9] Bashir A. Sheikh et al. "Development of New Therapeutics to Meet the Current Challenge of Drug Resistant Tuberculosis". In: *Current Pharmaceutical Biotechnology* 22.4 (Mar. 2021), pp. 480–500. ISSN: 13892010. DOI: 10.2174/1389201021666200628021702.

[10] Florence Brossier et al. "Molecular Analysis of the embCAB Locus and embR Gene Involved in Ethambutol Resistance in Clinical Isolates of Mycobacterium Tuberculosis in France". In: *Antimicrobial Agents and Chemotherapy* 59.8 (July 16, 2015), pp. 4800–4808. DOI: 10.1128/AAC.00150-15.

[11] Lu Zhang et al. "Structures of Cell Wall Arabinosyltransferases with the Anti-Tuberculosis Drug Ethambutol". In: *Science (New York, N.Y.)* 368.6496 (June 12, 2020), pp. 1211–1219. ISSN: 1095-9203. DOI: 10.1126/science.aba9102.

[12] Luke J. Alderwick et al. "The C-Terminal Domain of the Arabinosyltransferase Mycobacterium Tuberculosis EmbC Is a Lectin-Like Carbohydrate Binding Module". In: *PLOS Pathogens* 7.2 (Feb. 24, 2011), e1001299. ISSN: 1553-7374. DOI: 10.1371/journal.ppat.1001299.

[13] Betzaida Cuevas-Córdoba et al. "Mutation at embB Codon 306, a Potential Marker for the Identification of Multidrug Resistance Associated with Ethambutol in Mycobacterium Tuberculosis". In: *Antimicrobial Agents and Chemotherapy* 59.9 (Aug. 14, 2015), pp. 5455–5462. DOI: 10.1128/AAC.00117-15.

[14] Precious Bwalya et al. “Characterization of embB Mutations Involved in Ethambutol Resistance in Multi-Drug Resistant Mycobacterium Tuberculosis Isolates in Zambia”. In: *Tuberculosis* 133 (Mar. 1, 2022), p. 102184. ISSN: 1472-9792. DOI: 10.1016/j.tube.2022.102184.

Appendix for Chapter 4

4.A Mutations close to ethambutol

Mutation	Lig-Dist (Å)	mCSM-lig affinity	mCSM-lig outcome	mmCSM-lig affinity	mmCSM-lig outcome	MAF (%)	Odds Ratio	P-value	Adjusted P-value	Adjusted P-value significance
E405D	7.31	-1.94	Destabilising	-1.05	Destabilising	0.37	38.18	<0.0001	<0.0001	****
D328Y	9.03	-2.53	Destabilising	-0.61	Destabilising	0.34	28.63	<0.0001	<0.0001	****
M306V	3.81	-1.11	Destabilising	-1.09	Destabilising	15.84	22.87	<0.0001	<0.0001	****
Y319S	8.59	-0.9	Destabilising	-0.72	Destabilising	0.59	18.47	<0.0001	<0.0001	****
Q445R	4.47	-2.47	Destabilising	-0.89	Destabilising	0.07	17.5	0.01	0.09	ns
I303M	3.77	-2.62	Destabilising	-0.92	Destabilising	0.02	17.49	0.03	0.33	ns
D328F	9.03	-2.57	Destabilising	-0.68	Destabilising	0.02	17.49	0.03	0.33	ns
F330L	8.13	-1.68	Destabilising	-0.89	Destabilising	0.02	17.49	0.03	0.33	ns
F330S	8.13	-0.38	Destabilising	-0.85	Destabilising	0.03	17.49	0.03	0.33	ns
D328G	9.03	-3.87	Destabilising	-0.7	Destabilising	0.27	16.13	<0.0001	<0.0001	****
D1024N	5.71	-1.85	Destabilising	-0.9	Destabilising	1.19	14.65	<0.0001	<0.0001	****
M306L	3.81	-1.1	Destabilising	-1.09	Destabilising	0.89	10.42	<0.0001	<0.0001	****
M306I	3.81	-1.1	Destabilising	-1.09	Destabilising	13.66	10.33	<0.0001	<0.0001	****
Y319C	8.59	-1.81	Destabilising	-0.77	Destabilising	0.16	8.76	<0.0001	0.01	**
P404S	5.51	-1.01	Destabilising	-1.37	Destabilising	0.06	8.75	0.09	0.7	ns
M306T	3.81	0.4	Stabilising	-1.26	Destabilising	0.01	8.74	0.19	0.76	ns
F320C	7.78	-1.81	Destabilising	-1.03	Destabilising	0.01	8.74	0.19	0.76	ns
F330I	8.13	-1.68	Destabilising	-0.61	Destabilising	0.01	8.74	0.19	0.76	ns
Y334H	3.75	-2.39	Destabilising	-1.15	Destabilising	0.3	7.91	<0.0001	<0.0001	****
S297A	8.3	-1.18	Destabilising	-1.12	Destabilising	0.4	6.26	<0.0001	0.01	**
D328H	9.03	-2.68	Destabilising	-0.69	Destabilising	0.14	4.38	0.02	0.33	ns
V593M	7.77	-2.91	Destabilising	-0.86	Destabilising	0.02	4.37	0.34	1	ns

L1023M	7.57	-0.16	Destabilising	-1.09	Destabilising	0.16	4.37	0.16	0.76	ns
L402V	6.78	-1.74	Destabilising	-1.09	Destabilising	0.14	2.63	0.1	0.74	ns
D300E	6.95	-0.86	Destabilising	-1.16	Destabilising	0.01	2.19	1	1	ns
G305C	9.69	-1.57	Destabilising	-0.66	Destabilising	0.01	2.19	1	1	ns
S317F	7.35	-1.51	Destabilising	-0.9	Destabilising	0.01	2.19	1	1	ns
D328I	9.03	-3.41	Destabilising	-0.7	Destabilising	0.01	2.19	1	1	ns
D328V	9.03	-3.56	Destabilising	-0.69	Destabilising	0.01	2.19	1	1	ns
P404T	5.51	-1.06	Destabilising	-1.37	Destabilising	0.01	2.19	1	1	ns
A1020S	9.14	0.39	Stabilising	-0.95	Destabilising	0.01	2.19	1	1	ns
L1023Q	7.57	0.98	Stabilising	-1.06	Destabilising	0.01	2.19	1	1	ns
F320Y	7.78	-1.94	Destabilising	-1.05	Destabilising	0.02	2.19	1	1	ns
P404L	5.51	-2.68	Destabilising	-1.03	Destabilising	0.02	2.19	1	1	ns
Y319D	8.59	-0.89	Destabilising	-0.72	Destabilising	0.02	2.19	0.46	1	ns
D300G	6.95	-3.45	Destabilising	-1.04	Destabilising	0.02	1.09	1	1	ns
A989T	8.25	-0.55	Destabilising	-1.2	Destabilising	0.03	0.73	1	1	ns
S317A	7.35	-2.36	Destabilising	-1.12	Destabilising	0.05	0.55	1	1	ns
I303L	3.77	-2.16	Destabilising	-1.09	Destabilising	0.01	NA	NA	NA	ns
G305H	9.69	-1.85	Destabilising	-1.13	Destabilising	0.01	NA	NA	NA	ns
G305T	9.69	-0.96	Destabilising	-0.87	Destabilising	0.01	NA	NA	NA	ns
S325R	8.66	-1.91	Destabilising	-0.82	Destabilising	0.01	NA	NA	NA	ns
F330V	8.13	-1.67	Destabilising	-0.89	Destabilising	0.01	NA	NA	NA	ns
V593I	7.77	-2.95	Destabilising	-1.01	Destabilising	0.01	NA	NA	NA	ns
N318D	4.66	-3.65	Destabilising	-1.18	Destabilising	0.02	NA	NA	NA	ns
G305Q	9.69	-0.67	Destabilising	-0.74	Destabilising	0.04	NA	NA	NA	ns
Y333F	8.88	-2.14	Destabilising	-0.84	Destabilising	0.05	NA	NA	NA	ns

Table 4.A.1: Mutations close to EMB

Forty-seven single amino acid variation (SAV) mutations lying within 10Å of EMB and their corresponding ligand affinity changes (log fold change) measured by mCSM-Lig and mmCSM-lig. The estimated effect are categorised as Destabilising (log fold affinity change<0) and Stabilising ($\Delta\Delta G>0$). The genomic measures of minor allele frequency (MAF), Odds Ratio, P-values, and FDR adjusted P-values are shown. Statistical significance is indicated as: *P < 0.05, **P < 0.01, ***P < 0.001, ****P < 0.0001, ns: >0.05. The table is arranged by Odds Ratio to show mutation with the highest OR at the top for mutations close to EMB. Columns with NA indicate insufficient data to calculate Odds Ratio and P-values. Abbreviations used: FDR: false discovery rate, ns: not significant, EMB: ethambutol.

4.B Mutations close to the protein-protein interface

Mutation	PPI2-Dist (Å)	mCSM-PPI2 (ΔΔG)	mCSM-PPI2 outcome	MAF (%)	Odds Ratio	P-value	Adjusted P-value	Adjusted P-value significance
V456A	5.78	-0.2	Decreasing	0.09	52.55	<0.001	0	***
T642A	6.41	-0.51	Decreasing	0.05	26.27	<0.05	0.01	**
N129D	3.47	0.62	Increasing	0.02	26.25	0.01	0.1	ns
R524H	7.32	-0.03	Decreasing	0.02	17.49	0.03	0.33	ns
P690L	6.13	0.84	Increasing	0.02	17.49	0.03	0.33	ns
P690S	6.13	-0.1	Decreasing	0.02	17.49	0.03	0.33	ns
Q853R	7.11	0.25	Increasing	0.02	17.49	0.03	0.33	ns
A457T	6.92	0.24	Increasing	0.01	8.74	0.19	0.76	ns
T643A	9.02	-0.3	Decreasing	0.01	8.74	0.19	0.76	ns
T667A	7.04	-0.13	Decreasing	0.01	8.74	0.19	0.76	ns
V786L	9.92	0	Decreasing	0.01	8.74	0.19	0.76	ns
E958D	6.26	0.2	Increasing	0.01	8.74	0.19	0.76	ns
L585M	9.56	-0.29	Decreasing	0.02	8.74	0.19	0.76	ns
L635V	7.25	-0.37	Decreasing	0.02	8.74	0.19	0.76	ns
L636M	6.69	-0.59	Decreasing	0.13	8.74	0.19	0.76	ns
Q853P	7.11	0.04	Increasing	0.09	5.47	0.01	0.22	ns
M804I	7.01	-0.09	Decreasing	0.02	4.37	0.34	1	ns
I671V	7.76	-0.2	Decreasing	0.07	4.37	0.34	1	ns
L635M	7.25	-0.46	Decreasing	0.08	4.37	0.34	1	ns
T643I	9.02	-0.63	Decreasing	0.09	4.37	0.08	0.69	ns
F633L	6.83	-0.72	Decreasing	0.15	4.37	0.34	1	ns
W640F	7.24	-0.16	Decreasing	0.16	4.37	0.16	0.76	ns
L1023M	7.58	-0.02	Decreasing	0.16	4.37	0.16	0.76	ns
K107R	8.64	-0.06	Decreasing	0.01	2.19	1	1	ns
S119I	7.52	0	Increasing	0.01	2.19	1	1	ns
R122H	3.21	-0.24	Decreasing	0.01	2.19	1	1	ns
D127A	7.01	-0.34	Decreasing	0.01	2.19	1	1	ns
I132F	7.74	0.52	Increasing	0.01	2.19	1	1	ns
D178N	8.9	-0.3	Decreasing	0.01	2.19	1	1	ns
E521A	4.66	-0.94	Decreasing	0.01	2.19	1	1	ns
N522H	2.36	0.14	Increasing	0.01	2.19	1	1	ns
L528V	7.83	-0.18	Decreasing	0.01	2.19	1	1	ns
R568H	5.65	-0.2	Decreasing	0.01	2.19	1	1	ns
R573W	9.36	0.13	Increasing	0.01	2.19	1	1	ns
M575V	9.94	-0.39	Decreasing	0.01	2.19	1	1	ns
A627G	9.98	-0.05	Decreasing	0.01	2.19	1	1	ns

L632F	9.71	0.56	Increasing	0.01	2.19	1	1	ns
L636P	6.69	-0.07	Decreasing	0.01	2.19	1	1	ns
T667K	7.04	-0.06	Decreasing	0.01	2.19	1	1	ns
V668A	7.78	0.02	Increasing	0.01	2.19	1	1	ns
F676S	5.82	-1.17	Decreasing	0.01	2.19	1	1	ns
A684V	7.87	0.08	Increasing	0.01	2.19	1	1	ns
L686P	6.35	-0.41	Decreasing	0.01	2.19	1	1	ns
N807D	7.95	0.2	Increasing	0.01	2.19	1	1	ns
G851V	8.68	-0.21	Decreasing	0.01	2.19	1	1	ns
L1023Q	7.58	-0.63	Decreasing	0.01	2.19	1	1	ns
V456L	5.78	-0.12	Decreasing	0.02	2.19	1	1	ns
M462L	7.53	-0.27	Decreasing	0.02	2.19	1	1	ns
M462T	7.53	-0.18	Decreasing	0.02	2.19	1	1	ns
L466F	7.11	0.2	Increasing	0.02	2.19	1	1	ns
R468K	6.54	-0.08	Decreasing	0.02	2.19	1	1	ns
V554M	5.04	-0.5	Decreasing	0.02	2.19	1	1	ns
L558F	3.66	-0.01	Decreasing	0.02	2.19	1	1	ns
F584S	8.4	-0.38	Decreasing	0.02	2.19	1	1	ns
A630T	9.69	0.37	Increasing	0.02	2.19	1	1	ns
P690H	6.13	0.24	Increasing	0.02	2.19	1	1	ns
K882T	8.48	-0.05	Decreasing	0.02	2.19	1	1	ns
Q896P	7.86	-0.02	Decreasing	0.02	2.19	1	1	ns
V554L	5.04	0.24	Increasing	0.05	2.19	1	1	ns
T667S	7.04	0.11	Increasing	0.1	2.19	0.46	1	ns
L632V	9.71	-0.39	Decreasing	0.12	2.19	0.46	1	ns
K561R	3.51	0.04	Increasing	0.18	2.19	1	1	ns
A659T	9.86	0.05	Increasing	0.23	2.19	0.38	1	ns
L638F	4.22	0.34	Increasing	0.03	1.46	0.56	1	ns
A680T	9.65	0.22	Increasing	0.03	1.46	0.56	1	ns
I563V	9.56	-0.21	Decreasing	0.16	1.46	0.56	1	ns
M462I	7.53	-0.1	Decreasing	0.23	1.46	0.65	1	ns
D108N	7.07	-0.27	Decreasing	0.02	1.09	1	1	ns
A517V	6.24	0.04	Increasing	0.02	1.09	1	1	ns
L544F	7.51	0.24	Increasing	0.02	1.09	1	1	ns
M575L	9.94	-0.39	Decreasing	0.02	1.09	1	1	ns
A683V	8.73	0.19	Increasing	0.02	1.09	1	1	ns
R469H	9.2	-0.07	Decreasing	0.03	1.09	1	1	ns
Q121K	9.58	0.02	Increasing	0.11	1.09	1	1	ns
M557I	4.09	-0.28	Decreasing	0.13	1.09	1	1	ns
V131M	3.59	-0.25	Decreasing	0.82	1.09	0.78	1	ns

T126N	4.57	-0.19	Decreasing	0.02	0.73	1	1	ns
S119N	7.52	-0.1	Decreasing	0.11	0.73	1	1	ns
F115L	9.37	-0.56	Decreasing	0.03	0.55	1	1	ns
V123L	9.21	-0.16	Decreasing	0.05	0.55	1	1	ns
A547S	9.9	0.19	Increasing	0.05	0.55	1	1	ns
I563L	9.56	-0.37	Decreasing	0.13	0.55	1	1	ns
M582I	8.47	-0.25	Decreasing	0.05	0.44	0.59	1	ns
V668I	7.78	-0.16	Decreasing	0.33	0.44	0.41	1	ns
V456I	5.78	-0.01	Decreasing	0.01	NA	NA	NA	ns
L463V	4.26	-0.58	Decreasing	0.01	NA	NA	NA	ns
L466S	7.11	-0.4	Decreasing	0.01	NA	NA	NA	ns
R468H	6.54	-0.17	Decreasing	0.01	NA	NA	NA	ns
Q516E	6.81	0.12	Increasing	0.01	NA	NA	NA	ns
E521D	4.66	0.48	Increasing	0.01	NA	NA	NA	ns
A547V	9.9	0.07	Increasing	0.01	NA	NA	NA	ns
R568S	5.65	0.23	Increasing	0.01	NA	NA	NA	ns
M586I	3.7	0.03	Increasing	0.01	NA	NA	NA	ns
W621S	4.24	-0.26	Decreasing	0.01	NA	NA	NA	ns
W640S	7.24	-0.55	Decreasing	0.01	NA	NA	NA	ns
A659S	9.86	0.13	Increasing	0.01	NA	NA	NA	ns
M660I	3.42	-0.13	Decreasing	0.01	NA	NA	NA	ns
P661L	9.85	-0.13	Decreasing	0.01	NA	NA	NA	ns
K662E	8.34	-0.25	Decreasing	0.01	NA	NA	NA	ns
T670I	5.68	-0.16	Decreasing	0.01	NA	NA	NA	ns
I671M	7.76	-0.29	Decreasing	0.01	NA	NA	NA	ns
F672L	6.78	-0.85	Decreasing	0.01	NA	NA	NA	ns
A684D	7.87	0.55	Increasing	0.01	NA	NA	NA	ns
F688C	3.5	-0.86	Decreasing	0.01	NA	NA	NA	ns
G851S	8.68	-0.02	Decreasing	0.01	NA	NA	NA	ns
Q854R	6.45	0.23	Increasing	0.01	NA	NA	NA	ns
S856R	8.69	0.11	Increasing	0.01	NA	NA	NA	ns
K882R	8.48	-0.08	Decreasing	0.01	NA	NA	NA	ns
S119H	7.52	0.09	Increasing	0.02	NA	NA	NA	ns
R471H	5.17	-0.1	Decreasing	0.02	NA	NA	NA	ns
A684T	7.87	0.26	Increasing	0.02	NA	NA	NA	ns
K805N	3.09	-0.24	Decreasing	0.02	NA	NA	NA	ns
T956I	2.46	0.12	Increasing	0.02	NA	NA	NA	ns
E958A	6.26	-0.27	Decreasing	0.02	NA	NA	NA	ns
V125I	8.11	-0.36	Decreasing	0.03	NA	NA	NA	ns
A517E	6.24	0.61	Increasing	0.05	NA	NA	NA	ns

L544I	7.51	-0.3	Decreasing	0.05	NA	NA	NA	ns
T552A	8.64	-0.12	Decreasing	0.05	NA	NA	NA	ns
A553S	8.7	0.28	Increasing	0.05	NA	NA	NA	ns
F584C	8.4	-0.57	Decreasing	0.05	NA	NA	NA	ns
K561M	3.51	-0.35	Decreasing	0.07	NA	NA	NA	ns

Table 4.B.1: Mutations close to EmbB PPI
One hundred and twenty one single amino acid variation (SAV) mutations lying within 10Å of the Protein-Protein interface (PPI) and their corresponding PPI affinity changes ($\Delta\Delta$G) measured by mCSM-PPI2. The estimated effect are categorised as Destabilising ($\Delta\Delta$G<0) and Stabilising ($\Delta\Delta$G>0). The genomic measures of minor allele frequency (MAF), Odds Ratio, P-values, and FDR adjusted P-values are shown. Statistical significance is indicated as: *P < 0.05, **P < 0.01, ***P < 0.001, ****P < 0.0001, ns: >0.05. The table is arranged by Odds Ratio to show mutation with the highest OR at the top for mutations at the PPI. Columns with NA indicate insufficient data to calculate Odds Ratio and P-values. Abbreviations used: $\Delta\Delta$G: change in Gibbs free energy in Kcal/mol, FDR: false discovery rate, ns: not significant, EMB: ethambutol.

4.C Average stability comparisons for lineages

Lineage comparisons	**Samples (n)**	**Adjusted P-values**	***Adjusted P-values* Significance**
L1 vs L2	L1 (4013), L2 (3554)	<0.0001	****
L1 vs L3	L1 (4013), L3 (692)	<0.0001	****
L1 vs L4	L1 (4013), L4 (5276)	<0.0001	****
L2 vs L3	L2 (3554), L3 (692)	<0.0001	****
L2 vs L4	L2 (3554), L4 (5276)	<0.0001	****
L3 vs L4	L3 (692), L4 (5276)	<0.0001	****
Within Lineage comparisons			
L1: R vs S	R (n=209), S (n=3804)	<0.0001	****
L2: R vs S	R (n=3014), S (n=540)	<0.0001	****
L3: R vs S	R (n=486), S (n=206)	<0.0001	****

Table 4.C.1: Lineage comparisons for EmbB mutations
Kolmogorov-Smirnoff (KS) test reporting the statistical differences in distributions between *M. tuberculosis* lineages when assessed based on average stability changes ($\Delta\Delta$G) measured by mCSM-DUET, FoldX, DeepDDG, and Dynamut2. Lineage comparisons were performed for samples containing mutations associated with sensitivity (R: Resistant, S: Sensitive). These comparisons were performed for R and S samples between and within lineages. Statistical significance thresholds used are *P<0.05, **P<0.01, ***P<0.001, ****P<0.0001. Abbreviations used: $\Delta\Delta$G: change in Gibbs free energy in Kcal/mol, Adj. P-values: Bonferroni adjusted P-values, n=number of samples.

Chapter 5

GidB-streptomycin results

5.1 Background

5.1.1 Mechanism of action of streptomycin

Streptomycin (STR) was the first aminoglycoside antibiotic to be discovered, and the first antibiotic to be successfully used against TB.[1,2] As an aminoglycoside, STR is bactericidal in nature and works by interfering with ribosomal peptide/protein synthesis.[3,4] The mechanism of action of STR involves binding to the 30S subunit of bacterial ribosome at ribosomal S12 protein and 16S rRNA.[3,4] Different regions in the 16S rRNA including residues G526, G527 and the G530 hairpin loop are involved in binding to STR. In this manner, STR interferes with downstream protein synthesis by causing misreading of mRNA and thus halting protein synthesis altogether[5,6] **(Figure 1)**.

5.1.2 Streptomycin resistance in *M. tuberculosis*

The effectiveness of STR as a broad spectrum antibiotic (active against both Gram-positive and Gram-negative bacteria) has diminished largely due to prevailing and emerging drug resistance.[7] Rapid emergence of STR resistance in *M. tuberculosis* was quickly discovered from monotherapy regimens used initially.[8] Thereafter STR has only been used in combination with other drugs like isoniazid, rifampicin, and pyrazinamide for the treatment of active TB.

The main route to STR resistance is from mutations in the *rpsL* and *rrs* genes which together account for 60%-70% of STR resistance.[4,9] These genes respectively code for the ribosomal S12 protein and the 16S rRNA (located on the smaller subunit of the 30S subunit) ribosomal components respectively. However 20%-30% of strains do not show mutations in either gene,[10] where another gene, Glucose-inhibited division protein B (*gidB or gid*), previously known as Ribosomal RNA small subunit methyltransferase G (*rmsG*) is implicated in the development of low-level STR resistance with a minimum inhibitory concentration (MIC) of <32 μg/ml).[6,11] MIC is defined as the lowest concentration of a drug that prevents the visible *in vitro* growth of the microorganism being tested.[12]

The gene *gidB* encodes a conserved S-adenosylmethionine (SAM)-dependent 7-methylguanosine (M7G) methyltransferase. The SAM co-factor of GidB is known to methylate G527 in the 530 loop of the 16S rRNA, and as such is considered a mutational hotspot where GidB mutants lacking an M7G modification result in resistance to STR **(Figure 1)**.

For mutations in the *rpsL* gene, substitutions in codon 43 and 88 (K43R and K88R) have been the most commonly reported mutations associated with high-level STR resistance. For the *rrs* gene, the most common mutations occur around nucleotides 530 and 915. SAV mutations in *gidB* are increasingly

being identified and linked to STR resistance.[13] Interestingly, *gidB* mutations occurring together with mutations in *rpsL* and/or *rrs* mutations result in high level STR resistance.[6,11,14,15] SAV mutations G34E, P75T, G76R, L79S, E92K, L101F, G164D, E170D, R206L in *gidB* have also been linked to STR resistance and include residues involved in interactions with the SAM co-factor in the GidB-SAM complex. Also SAV mutations including G30R, W45C, W45S, H48Y, L49P, N52T, D67H, P75S, L79F, P84L, W148R, G164C have been linked to low-level resistance. Further, it has been shown that the sequential progression of low-high level resistance to STR has occurred due to existing *gidB* alterations in the genetic background.[5] In this chapter, SAVs in the GidB were investigated to understand the biophysical consequences of SAVs on the GidB-SAM and STR complex.

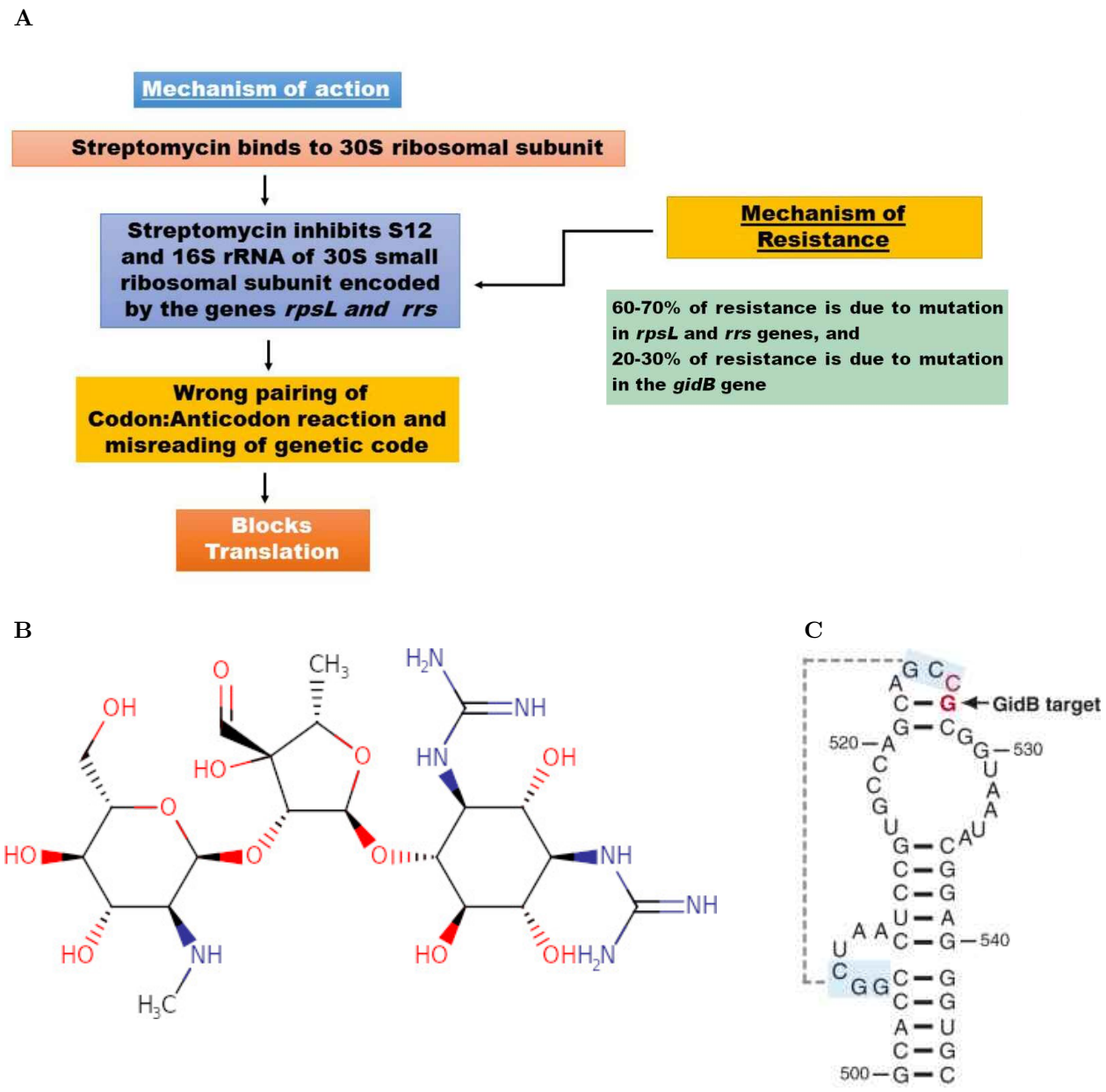


Figure 1: Mechanism of action and resistance for streptomycin and its chemical structure
A) An overview of the mechanism of action and resistance for streptomycin (STR), with 60-70% STR resistance is attributed to mutations within *rpsL and rrs* genes, while 20-30% STR resistance is due to mutations within *gidB*. Figure adapted from Sheikh, *et. al.*.[16] **B)** The chemical structure of STR displayed at the bottom left is sourced from DrugBank (ID: DB01082), **C)** Figure showing the target for GidB (G527 residue), on the secondary structure of the 530 loop region of *E. coli* 16SrRNA. Nucleotide in red is involved in STR binding, while those in cyan are involved in the formation of the pseudo-knot structure connected by the dotted line. Figure adapted from Okamoto, *et. al.*.[6]

5.1.3 Description of the GidB complex

As it is the GidB complex that interacts with STR, a model that included all co-factors and binding partners was required. As described above, SAM is considered an essential partner for GidB. Additionally, the structure of *T. Thermophilus* GidB (PDB-ID: 3G89) published in 2009[17] contained an Adenosine monophosphate (AMP) molecule bound together with SAM. The authors proposed that the site for AMP binding may serve as a potential binding site for RNA. Thus AMP was considered one of

the interacting partners of GidB. Similarly, in the absence of a crystal structure for the 30S ribosomal unit in *M. tuberculosis* bound with STR, the crystal structure of *T. thermophilus* 30S ribosomal subunit bound with streptomycin (PDB-ID: 4DR3)[18] was used to extract a 5nt RNA fragment (residues G526-G530) containing STR. Therefore the 5nt RNA fragment was also considered an interacting partner required to form the final GidB-complex. As an experimentally determined structure of GidB in *M. tuberculosis* was not publicly available until 2021 (PDB-ID: 7CFE, paper not yet published), the structure of *M. tuberculosis* GidB was modelled, followed by molecular docking with AMP, SAM, and 5nt-RNA (bound with STR) to form the complete GidB-AMP-SAM-RNA-STR complex (see Chapter 2: Methods) for use in all downstream analyses.

Interactions of GidB binding partners

Molecular interactions between GidB binding partners: AMP, RNA, and SAM, and those interacting with the drug STR were identified using LigPlus, PLIP and Arpeggio resulting in a total of forty interacting residues:

- Eighteen residues at sites 33, 34, 35, 36, 37, 38, 47, 48, 51, 94, 97, 137, 138, 139, 163, 164, 165, and 199 were identified to be interacting with the RNA fragment.
- Twenty-one residues were found to be interacting with the co-factor SAM present at sites 68, 69, 92, 93, 97, 117, 118, 119, 120, 136, 137, 138, 139, 140, 148, 218, 219, 220, 221, 222, and 223.
- Four residues at sites 123, 125, 213, and 214 were identified to be interacting with AMP.
- Four residues at sites 118, 148, 220, 223, and 224 were identified to be interacting with STR.

An overview of the GidB structural complex with all interactions identified is shown in **Figure 2**.

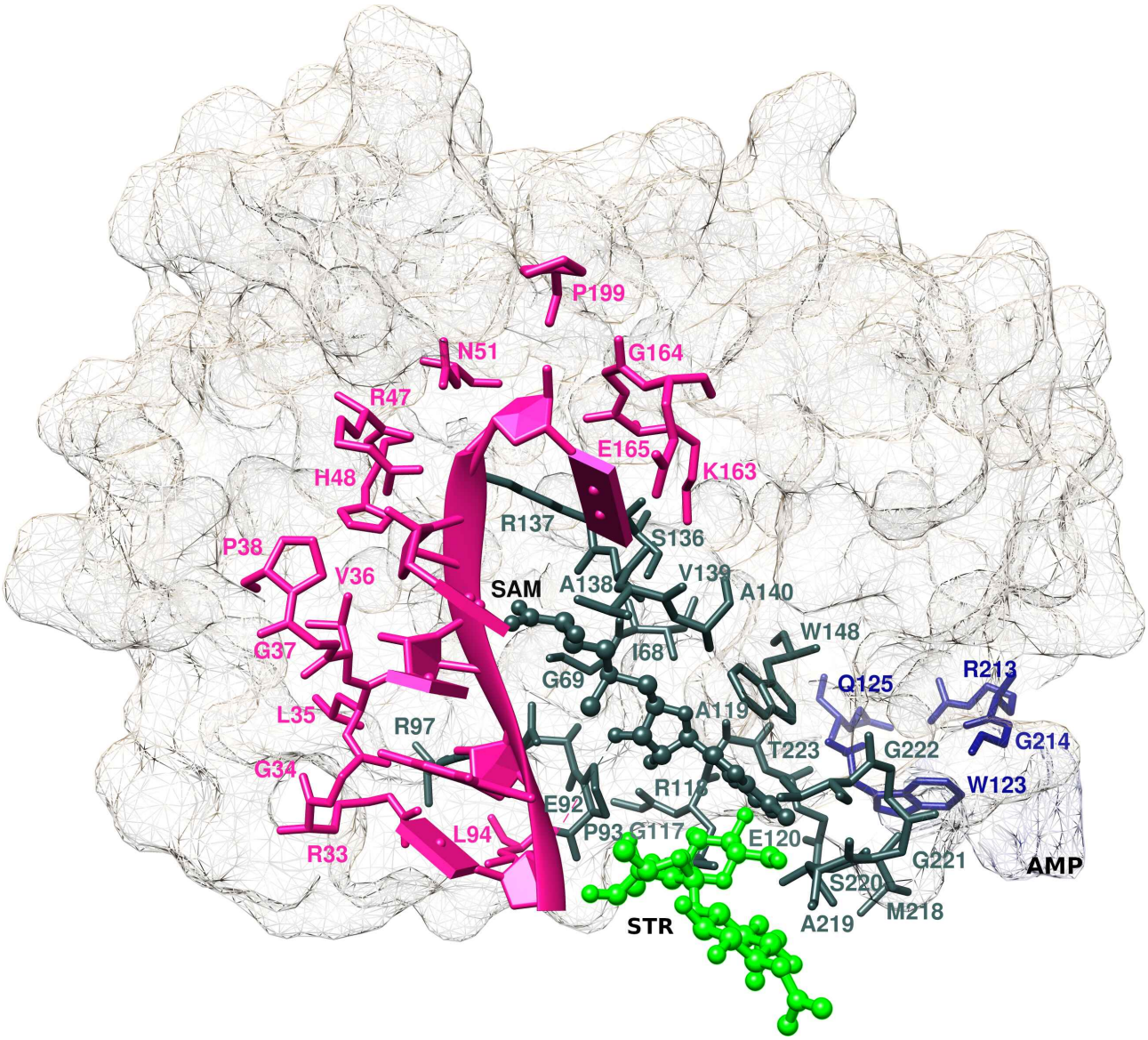


Figure 2: Description of *M. tuberculosis* GidB complex with all interacting partners: RNA, SAM, and AMP.

Overall description of GidB and its interacting partners and streptomycin (STR). The RNA fragment (G526-G530) is shown in deep pink with all its interacting residues shown in light pink as sticks. The co-factor SAM is shown as ball-and-stick in dark slate grey with its interacting residues shown in light grey as sticks. STR appears as green ball-and-stick with interacting residues shared with SAM. The surface of AMP and its interacting residues are indicated similarly in navy blue. Abbreviations used: SAM: S-adenosylmethionine, AMP: adenosine monophosphate.

5.2 Structural and genomic insights into streptomycin resistance

5.2.1 Mutational landscape of GidB

Multiple SAV mutations are distributed on GidB and include, but are not restricted to GidB binding partners

A total of 531 SAVs were found in the protein coding region of *gidB* (Genomic id: Rv3919, coding region:4407528-4408202), and appear distributed across the protein **(Figure 3)**, with mutations present in 201 unique positions for a maximum of 6 SAVs **(Figure 4)**.

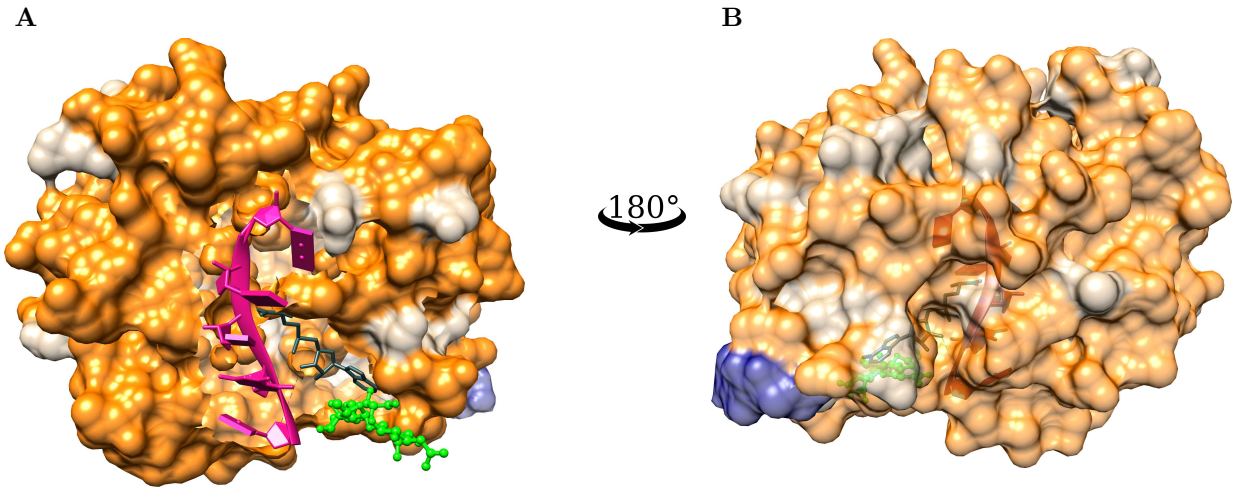


Figure 3: Mutational landscape of *M. tuberculosis* GidB
An overview of all mutational sites on *M. tuberculosis* GidB appearing as surface representation in tan colour. Sites associated with SAVs are coloured orange. Panels **A)** and **B)** are opposing representations (rotated 180°) of GidB, with STR shown in green as ball-and-stick. The RNA fragment (G526-G530) is shown in deep pink, co-factor SAM appears as dark slate grey sticks, while the surface of AMP is indicated in navy blue. The figure is generated using UCSF Chimera version 1.14. Abbreviations used: SAV: single amino acid variation, AMP: adenosine monophosphate, SAM: S-adenosylmethionine, STR: streptomycin.

Most residues interacting with GidB binding partners were associated with SAVs with a maximum of 6 SAVs at a single site. RNA interacting residue 165, SAM interacting residue 222 and AMP interacting residue 123 were not associated with SAVs **(Figure 2** and **Figure 4)**. While sites with multiple SAVs were not restricted to GidB binding partners **(Figure 4)**, most residues (33/40) interacting with GidB binding partners exhibited multiple SAVs, with mutant residues altering the corresponding wild-type amino acid property **(Figure 4)**. Mapping mutations by position in GidB highlights the following:

Sites with RNA interactions associated with a maximum of 6 SAVs (sites marked in deep pink)

- Single mutation: L94
- Budding resistant hotspots: R33, V36 and N51

- Hotspots with three mutations: P38, R97, V139, P199
- Hotspots with four mutations: L35, G37 and K163
- Hotspots with five mutations: G34, R47, R137 and A138
- Hotspots with six mutations: H48 and G164

Sites with SAM interactions associated with a maximum of 5 SAVs (sites marked in dark slate grey)

- Single mutation: I68, A219, S220, T223, E120
- Budding resistant hotspots: A140, M218, G221
- Hotspots with three mutations: P93, R97, R118, S136, V139, W148
- Hotspots with four mutations: G69, G117, A119
- Hotspots with five mutations: E92, R137, A138

Sites with AMP interactions associated with a maximum of 3 SAVs (sites marked in navy blue)

- Single mutation: Q125, R213
- Budding resistant hotspots: None
- Hotspots with three mutations: G214

Sites with STR interactions associated with a maximum of 3 SAVs (sites marked in green)

- Single mutation: S220, T223
- Budding resistant hotspots: A224
- Hotspots with three mutations: R118, W148

The majority (56%, n=298) of the mutational effects resulted in electrostatic changes.

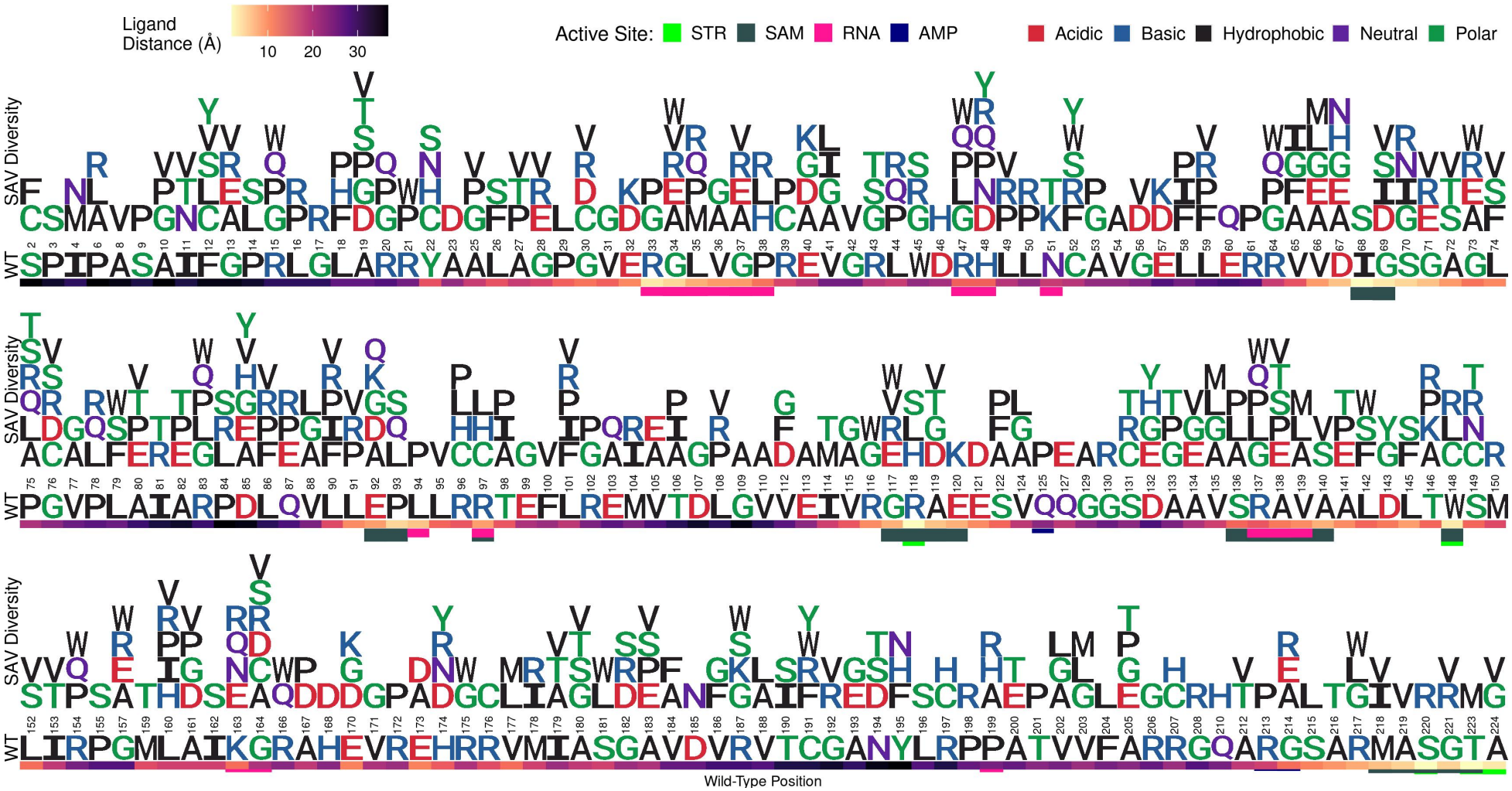


Figure 4: Sites associated with SAVs in *M. tuberculosis* GidB protein

Logo plot showing 201 unique sites/positions associated with 531 SAVs in *M. tuberculosis* GidB. The horizontal axis shows the wild-type positions associated with SAVs in GidB and the vertical axis shows all the mutant residues observed in our data highlighting SAV diversity at any given site. Residues are coloured according to the amino acid (aa) property where acidic aa appear in red, basic aa appear in blue, hydrophobic aa in black, neutral aa in purple, and polar aa in dark green. The structural positions associated with SAVs in GidB are indicated on the horizontal axis. The wild-type (WT) residues also coloured according to aa property appear under the respective position markings. The heat bar underneath the WT residues indicate the distance of that position from STR according to the magma colour gradient where light yellow indicates sites closer to STR (ligand distance in Angstroms). The positions are further annotated to reflect active site residues involved in interactions with binding partners: STR (green), RNA fragment (deep pink), SAM (dark slate grey), and AMP (navy blue). The figure is generated using R statistical software version 4.0.2, ggplot2 package. Abbreviations used: SAV: single amino acid variation, AMP: adenosine monophosphate, SAM: S-adenosylmethionine, STR: streptomycin.

5.2.2 Mutational outcome from protomer stability changes and evolutionary conservation

Mutational consequences are destabilising for protomer stability and have deleterious impact on protein function

Most mutations had a destabilising effect on the overall protomer stability when measured by different computational tools **(Figure 5A-D)**, with DeepDDG estimating 93% (n=495) as destabilising, followed by Dynamut2 (n=460) and mCSM-DUET (n=450) estimating about 85% mutations as destabilising, followed by FoldX predicting 80% (n=425) mutations as destabilising. From an evolutionary conservation perspective, most mutations were predicted to result in a deleterious impact (effect) on protein function indicated by PROVEAN and SNAP2 scores. PROVEAN estimated 75% (n=396) **(Figure 5E)** and SNAP2 estimated nearly 67% (n=355) SAVs to result in a deleterious impact **(Figure 5F)**.

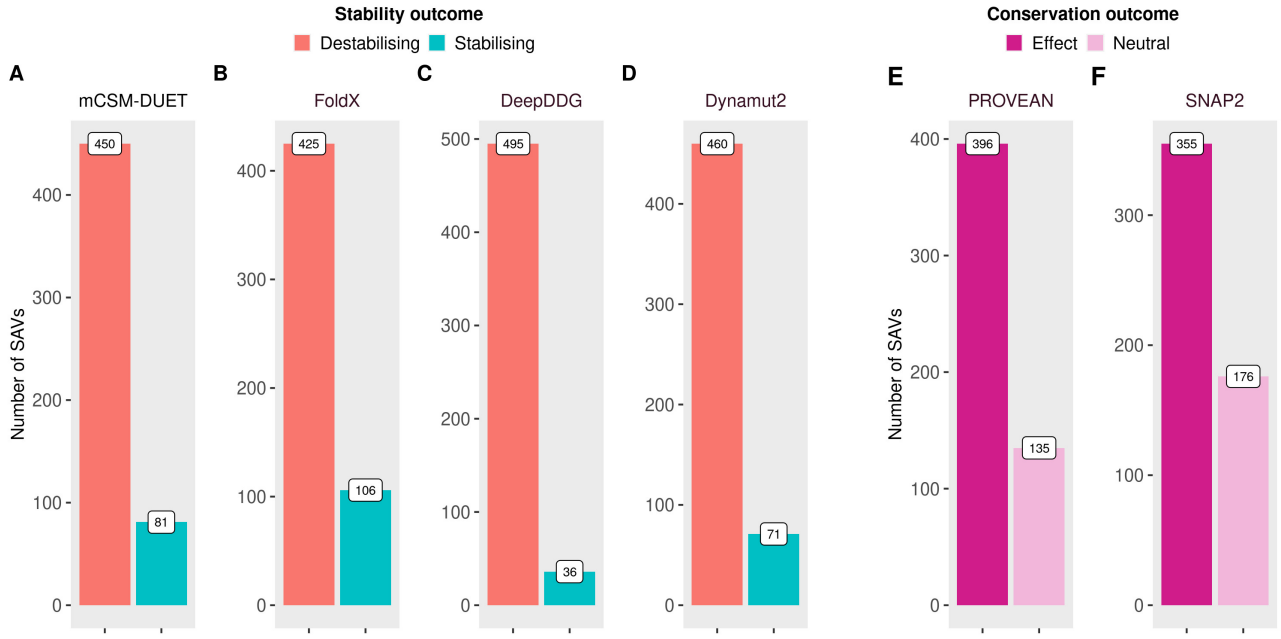


Figure 5: Protein stability outcome of SAVs in *M. tuberculosis* GidB
Mutational impact on overall protein stability and evolutionary conservation changes for 531 SAVs, **A-D)** Barplots showing number of SAVs categorised as destabilising (red) or stabilising (blue) according to protein stability changes (ΔΔG Kcal/mol) as measured by four computational tools: mCSM-DUET, FoldX, DeepDDG, and Dynamut2, **E-F)** Number of SAVs categorised as Effect/Deleterious (magenta) or Neutral (pink) according to evolutionary conservation changes estimated by computational tools: PROVEAN, and SNAP2. The figures are generated using R statistical software version 4.0.2, ggplot2 package. Abbreviations used: ΔΔG: change in Gibbs free energy, SAV: single amino acid variation.

Evolutionary and structure-based predictors provide different insights into understanding mutational impact. Mutational impact in this context is considered to be its effect on protein stability, drug binding affinity, other binding affinities such as PPI or nucleic acid, and functional effects arising from protein sequence variations. The first three mutational consequences are assessed by structure

based predictors relying on the 3D structure of a protein, while the last is assessed by sequence based predictors relying mainly on evolutionary conservation trends across many proteins using multiple sequence alignments. The sequence based predictors are aimed at predicting pathogenicity or change of molecular function, structure based tools rely on estimating variant effects in relation to structure damage, corresponding to stability changes, as protein stability is considered the basic characteristic affecting function, activity, and regulation. Predictors such as ConSurf are able to use both structural and sequence information to identify important functional regions conserved in proteins. A variant classified as 'deleterious' to protein conservation may display gain-of-function in the presence of a drug through optimised protein stability. Thus, different methodological strategies benefit from complementary information when assessing specific proteins.

Sites interacting with AMP, and those distal to STR and RNA have stabilising mutational consequences

When assessing the impact on protomer stability changes due to mutations, the estimates from all four tools employed: mCSM-DUET, FoldX, DeepDDG, and Dynamut2, were considered together and averaged to provide a consensus mutational effect **(Figure 6)**. While most (n=410) mutational effects were destabilising for overall protomer stability, mutations at sites interacting with AMP predominantly resulted in stabilising effects **(Figure 6** and **Figure 7**, sites marked in navy blue**)**. The impact of mutations on SAM, RNA, and STR interacting sites were overall destabilising for protomer stability with the exception of sites 139 and 199 where all mutational impact resulted in stabilising effect **(Figure 7**, sites marked in pink and dark slate grey **)**. Sites distal to STR and RNA exhibited predominantly stabilising mutational impact **(Figure 6)** with mutational sites 6, 39, 87, 122, 139, 166, 181, 194, 199, 218, resulting in all mutational effects being stabilising, and sites 12 (5 SAVs) and 85 (6 SAVs) with all but one mutations being stabilising **(Figure 7)**.

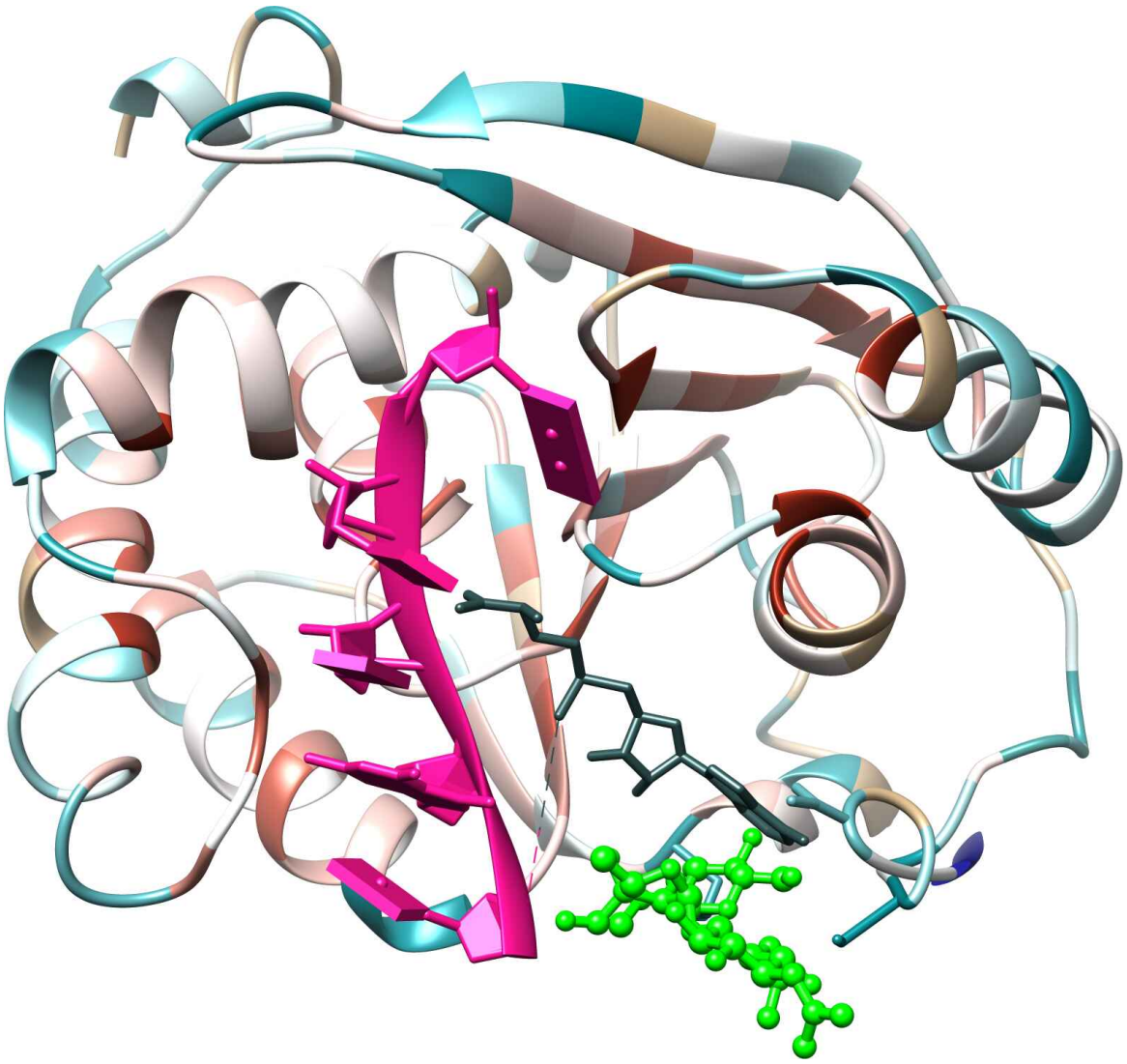


Figure 6: Average protein stability effects of SAVs mapped onto the *M. tuberculosis* GidB protein structure

The protein stability changes (ΔΔG Kcal/mol) of SAV mutations measured by mCSM-DUET, FoldX, DeepDDG, and Dynamut2 were averaged and mapped onto GidB positions (appearing as tan coloured ribbon). Destabilising mutational sites are depicted in red and stabilising mutational sites appear in blue, where colour intensity reflects the extent of effect, ranging from -1 (most destabilising) to +1 (most stabilising). STR is shown in green as ball-and-stick, RNA fragment (G526-G530) is shown in deep pink, co-factor SAM appears as dark slate grey sticks, while AMP is indicated on the ribbon representation as navy blue. The figure is rendered using UCSF Chimera version 1.14. Abbreviations used: ΔΔG: change in Gibbs free energy, SAV: single amino acid variation, AMP: adenosine monophosphate, SAM: S-adenosylmethionine, STR: streptomycin.

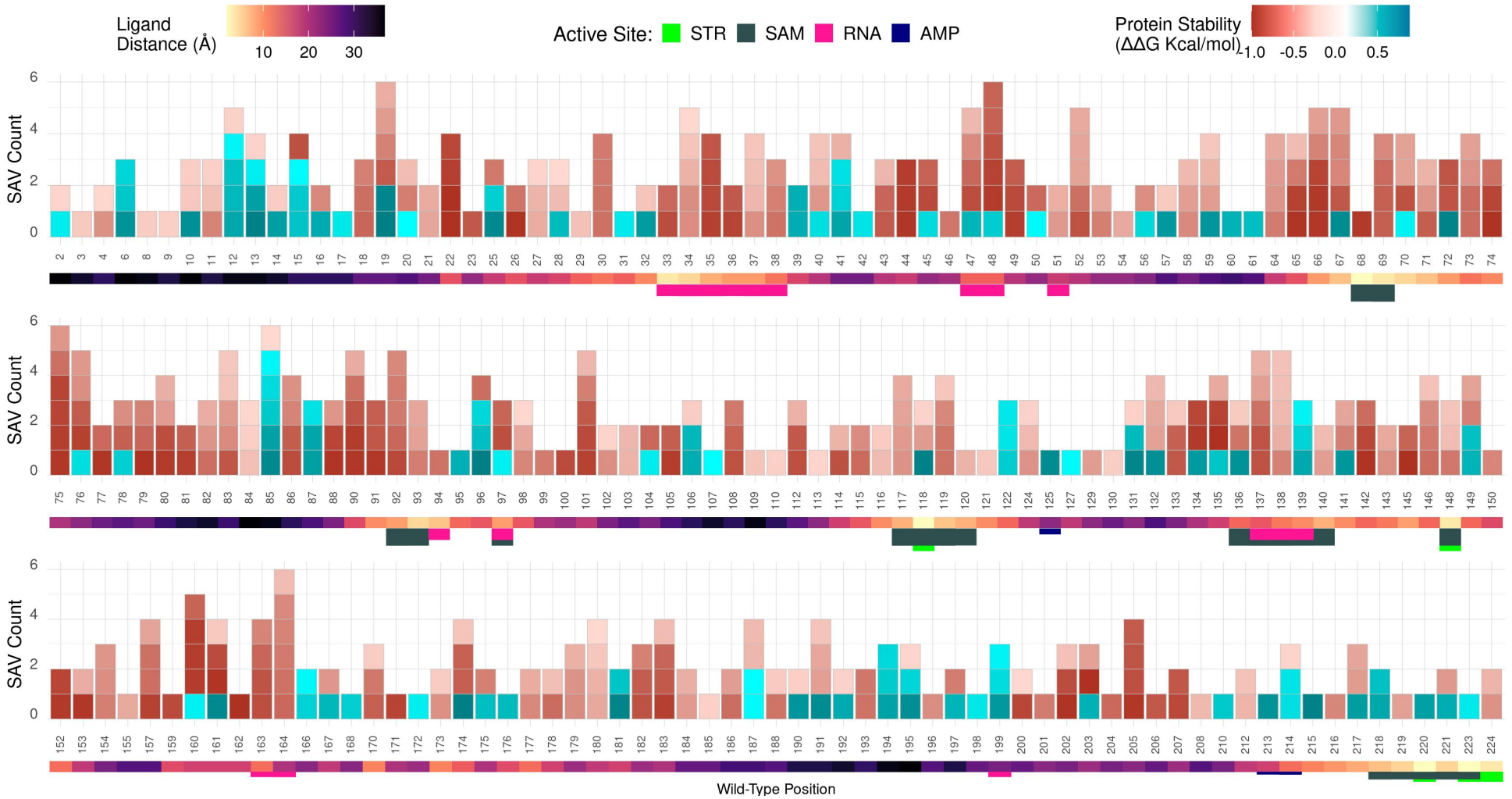


Figure 7: Average protein stability effect for individual SAVs occurring in *M. tuberculosis gidB*

Barplot showing the number of number of single amino acid variation (SAV) mutations at each position in GidB coloured by the average protein stability effect, where the horizontal axis shows the wild-type positions associated with SAVs, and the vertical axis shows the number of SAVs at that position. For a given position, each SAV is coloured by the average protein stability effect calculated across estimates (ΔΔG Kcal/mol) from mCSM-DUET, FoldX, DeepDDG, and Dynamut2. The structural positions associated with SAVs in GidB are indicated on the horizontal axis. The heat bar underneath the positions indicates the distance of that position from STR according to the magma colour gradient where light yellow indicates sites closer to STR (ligand distance in Angstroms). The positions are further annotated to reflect active site residues involved in interactions with binding partners: STR (green), RNA (deep pink), SAM (dark slate grey), and AMP (navy blue). The figure is generated using R statistical software version 4.0.2, ggplot2 package. The structural figure is rendered using UCSF Chimera version 1.14. Abbreviations used: ΔΔG: change in Gibbs free energy in Kcal/mol, SAV: single amino acid variation, AMP: adenosine monophosphate, SAM: S-adenosylmethionine, STR: streptomycin.

5.2.3 Mutational consequences on affinity changes and prominent mutational effects

Mutations decrease binding affinity of STR while increasing affinity for RNA

Only 10% (n=51) of SAVs inducing changes in ligand affinity were within 10Å of STR. These mutations occurred at 21 distinct sites, with most sites (n=17) showing up to three mutations. All mutations were predicted to result in a destabilising effect on STR binding affinity measured by both mCSM-lig and mmCSM-lig **(Figure 8A** top panel, Appendix Table 5.A.1**)**. The average effect on binding affinity at 21 mutational sites were shown to have mildly destabilising mutational consequences **(Figure 8A** bottom panel**)**. Analysing the sites close to RNA highlighted 43% (n=226) of mutations, corresponding to 76 distinct sites that were located within 10Å of the RNA measured by mCSM-NA. Among these, 61% (n=139) of mutations resulted in destabilising effects with triple SAVs being the most frequent (n=22) **(Figure 8B** top panel, Appendix Table 5.B.1**)**. Interestingly, sites close to the RNA fragment predominantly exhibited mutations with mild to moderate stabilising effects, with destabilising mutations located farther away **(Figure 8B** bottom**)**.

Of the total 201 unique sites in GidB displaying SAVs, about 50% of sites exhibited up to two SAVs, with single and double mutations occurring at 49 sites each. Triple mutations occurred most frequently presenting at 52 sites, followed by 32 sites displaying 4 mutations, 14 sites displaying 5 mutations, with 5 sites showing a maximum of 6 mutations **(Figure 8C** top panel**)**.

The most prominent effects on STR interactions were from reduced (destabilising) affinity on STR contributed by mutations occurring at 17 surrounding sites **(Figure 8C**, yellow text boxes, and bottom panel**)**. This was followed by sites close to the RNA fragment where mutational impact increased RNA binding affinity from 9 mutational sites, with only 2 sites contributing to destabilising effects, located farther away from the RNA fragment **(Figure 8C**, brown text boxes, and bottom panel**)**. Though all other sites were affected largely (n=132) by destabilising mutations, stabilising mutations sites were near GidB binding/interacting partners **(Figure 8C** blue and red text boxes, and bottom panel**)**. This suggests that the most prominent mutational effects result in reduced binding affinity for STR while increasing the binding affinity for RNA.

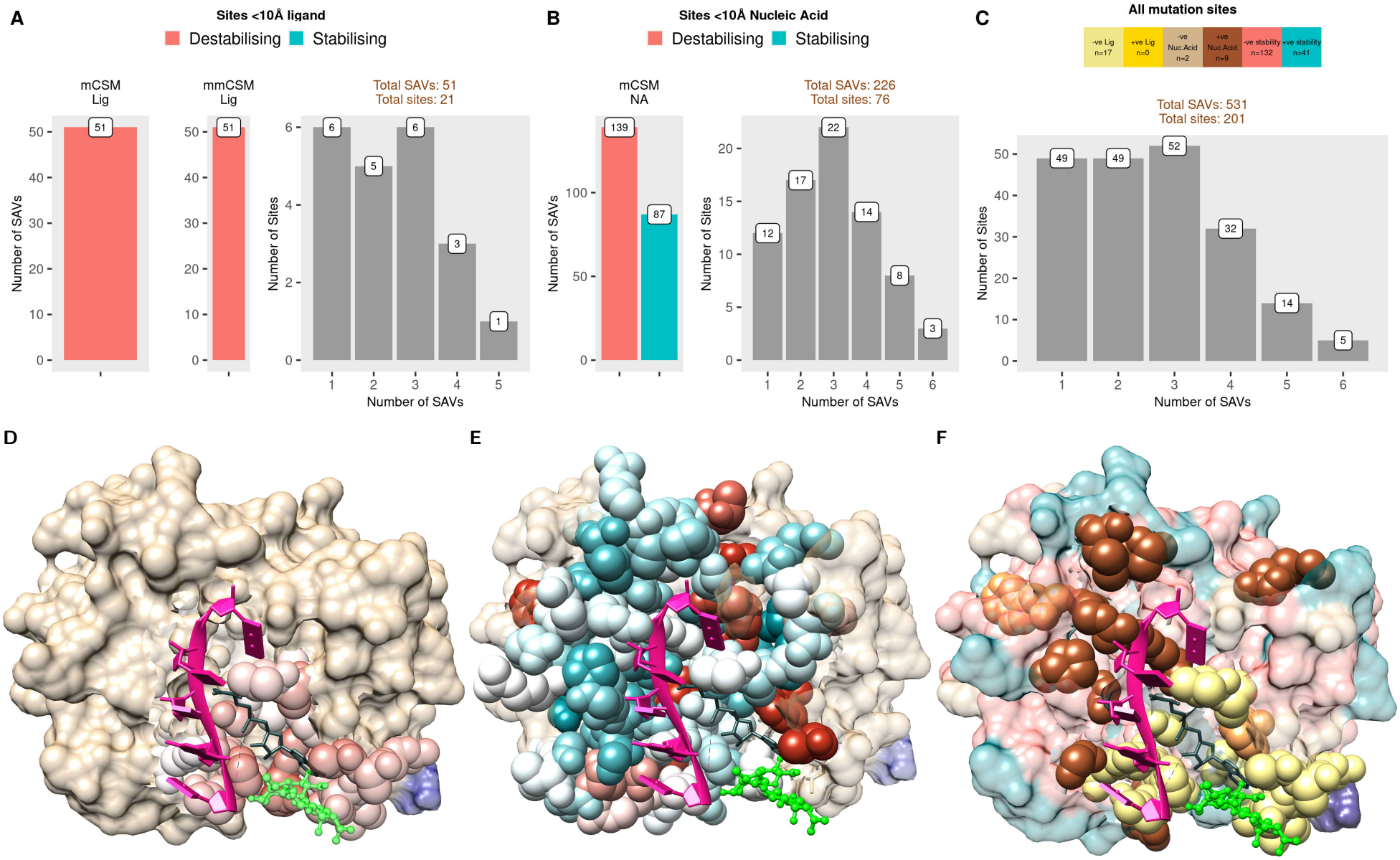

A
Sites <10Å ligand
Destabilising
Stabilising
mCSM Lig
mmCSM Lig
Total SAVs: 51
Total sites: 21
Number of SAVs
Number of Sites
B
Sites <10Å Nucleic Acid
mCSM NA
Total SAVs: 226
Total sites: 76
C
All mutation sites
-ve Lig n=17
+ve Lig n=0
-ve Nuc.Acid n=2
+ve Nuc.Acid n=9
-ve stability n=132
+ve stability n=41
Total SAVs: 531
Total sites: 201
D
E
F

Figure 8: Mutational impact on STR binding affinity, protein-protein interaction on GidB and sites with the most prominent mutational effects within *M. tuberculosis* GidB
The top panel displays barplots showing the mutational outcome of affinity changes and their corresponding site frequency, while the bottom panel shows the corresponding mutational impact mapped onto GidB. STR is shown in green as ball-and-stick. Other binding partners are indicated: RNA fragment (G526-530) in deep pink, co-factor SAM appears as dark slate grey sticks, while the surface of AMP is indicated in navy blue. **A)** Mutational impact on STR binding affinity (log fold change) upon mutation estimated from mCSM-lig and mmCSM-lig mutations for 51 mutations corresponding to 21 sites within 10Å of STR, **B)** Mutational impact on RNA binding affinity (ΔΔG) for 226 mutations, corresponding to 76 sites within 10Å of the RNA fragment. For both parts A) and B), red denotes destabilising mutational sites while blue denotes stabilising mutational sites, and the colour intensity reflects the extent of the effect ranging from -1 (most destabilising) to +1 (most stabilising), **C)** Most prominent mutational effect for all 531 SAVs (corresponding to 201 sites) prioritised in order of increasing effect size: mCSM/mmCSM-lig, mCSM-NA, protomer stability changes. Mutational effects are coloured according to the effect type with brighter colours representing stabilising mutational effects. Sites marked in yellow indicate changes due to ligand (STR) binding affinity with light yellow indicating destabilising effect, brown areas indicate changes in nucleic acid (NA) i.e. RNA binding affinity with light brown indicating destabilising and dark brown denoting stabilising effects. Protomer stability changes are coloured with blue indicating stabilising and red indicating destabilising mutational consequences. The corresponding number of mutation sites contributing to the different effect types are indicated in the text box at the top, and coloured accordingly. The barplot figures are generated using R statistical software version 4.0.4, ggplot2 package. The structure figures are generated using Chimera version 1.14. Abbreviations used: Å: angstroms, ΔΔG: change in Gibbs free energy in kcal/mol, SAV: single amino acid variation, AMP: adenosine monophosphate, SAM: S-adenosylmethionine, STR: streptomycin.

5.2.4 Mutational association with STR resistance and flexibility

RNA sites are more conserved than SAM, and all GidB interacting sites are associated with moderate to high flexibility

Mutational association with resistance according to aggregate DST data showed only a minority (7%, n=38) of mutations as resistant. Mutational sites on GidB were mapped onto the 3D structure to highlight sites with exclusively resistant (red), sensitive (blue) and sites displaying both resistant and sensitive mutations (purple). For GidB, there were 5 sites with exclusively resistant mutations, 29 sites with both resistant and sensitive mutations, while 167 sites with exclusively sensitive mutations **(Figure 9A)**.

ConSurf scores are calculated for each site on the protein, and range from 1 (rapidly evolving, variable sites) to 9 (slowly evolving, conserved sites). Exclusively resistant mutation sites did not appear to occur in the conserved regions of GidB **(Figure 9B** left panel, **Figure 9A)**, though most mutations (n=147) occurred in the highly conserved regions of GidB (ConSurf score 9) **(Figure 9B** right panel**)**. Also, sites surrounding the RNA were particularly more conserved compared with regions surrounding co-factor SAM **(Figure 9B** left panel, **Figure 9A)**. Residues 222-224 (interacting with SAM) appear in yellow due to inconclusive results from ConSurf **(Figure 9B** left panel**)**, similar to residues 1 and 2 which appear at the surface and away from STR, RNA and SAM **(Figure 9B** left panel**)**.

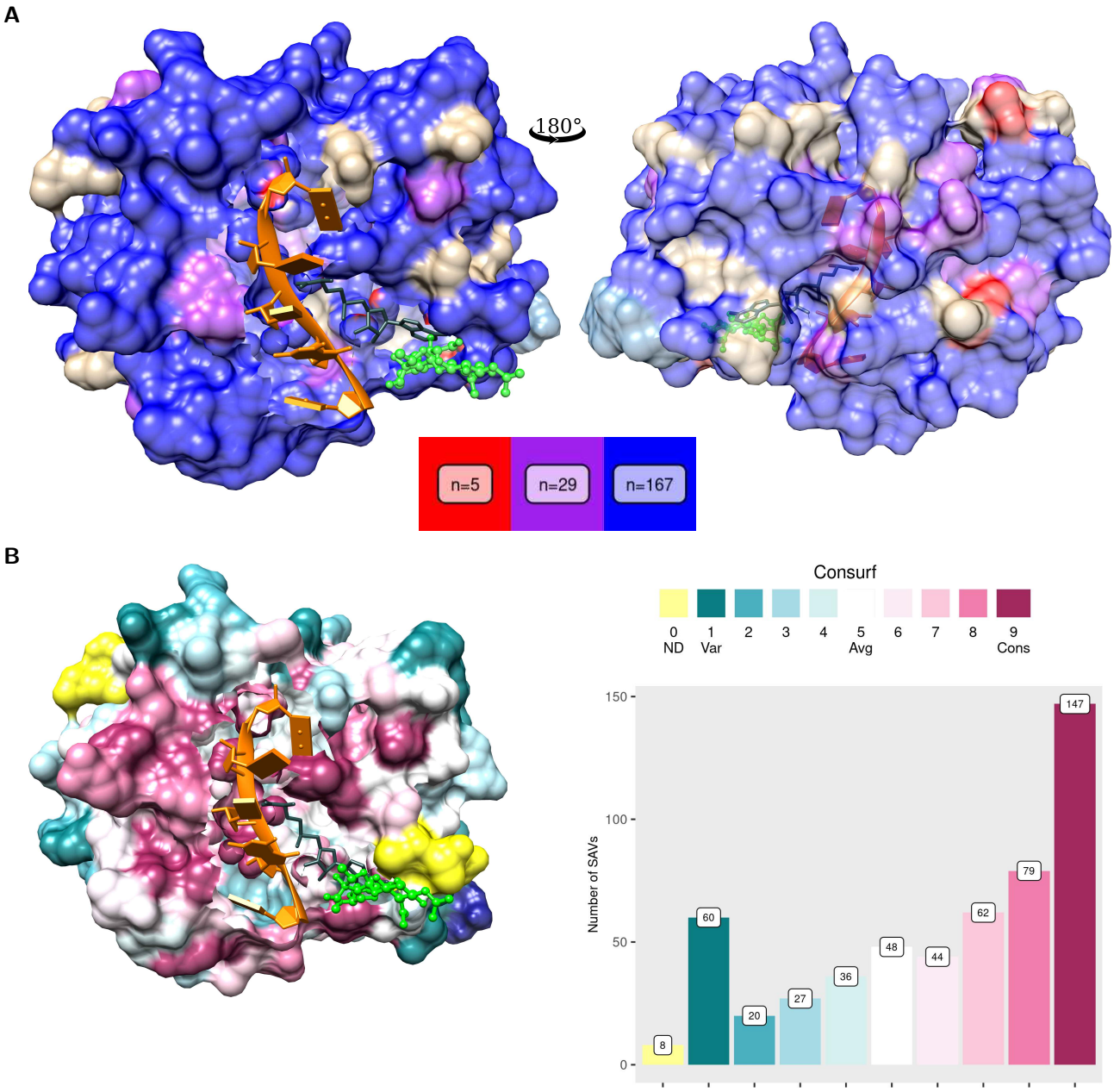


Figure 9: Mutational association with streptomycin resistance and evolutionary conservation in *M. tuberculosis* GidB

Mutational landscape of *M. tuberculosis* GidB according to different measures with all sites associated with SAVs on GidB. STR appears in green as ball-and-stick, co-factor SAM appears in dark slate grey. The RNA fragment appears in orange, while the AMP surface is shown in steel blue in panes A and C, and navy blue in panel B to aid visibility. **A)** The left panel shows all mutational sites associated with resistant (red, n=5 sites), sensitive (blue, n=167 sites), while common sites with both resistant and sensitive mutations appear in purple (n=29). The corresponding right panel depicts the structure rotated by 180°, **B)** Left panel shows GidB coloured according to ConSurf Scores where maroon indicates conserved sites and teal indicates variable sites. Yellow areas reflect sites with uncertainty due to insufficient data for ConSurf score calculation. The barplot on the right panel shows the number of mutations associated with ConSurf values that range from 1 (variable) in teal to 9 (conserved) in maroon, where 0 denotes insufficient data/not defined (ND). The barplot figures are generated using R statistical software version 4.0.4, ggplot2 package. All structure figures were generated using UCSF Chimera version 1.14. Abbreviations used: SAV: single amino acid variation, AMP: adenosine monophosphate, SAM: S-adenosylmethionine, STR: streptomycin.

The local flexibility in GidB in relation to STR resistance was also analysed with thickness of the ribbon/tube (thinthick) indicating the extent of flexibility. Normal mode analysis (from Dynamut2) of the protein component of GidB-complex highlighted that regions surrounding GidB interacting partners were associated with moderate to high flexibility **(Figure 10** left panel**)**. SAM interacting residue S220 showed the highest flexibility **(Figure 10** left panel**)**, followed by residues A140, R118 and A119 **(Figure 10** right panel**)**. Sites with exclusively resistant mutations were not located in areas of high flexibility **(Figure 10** right panel**)**. All GidB binding partner residues associated with moderate-to-high flexibility were sites with sensitive mutations **(Figure 10** left panel**)**.

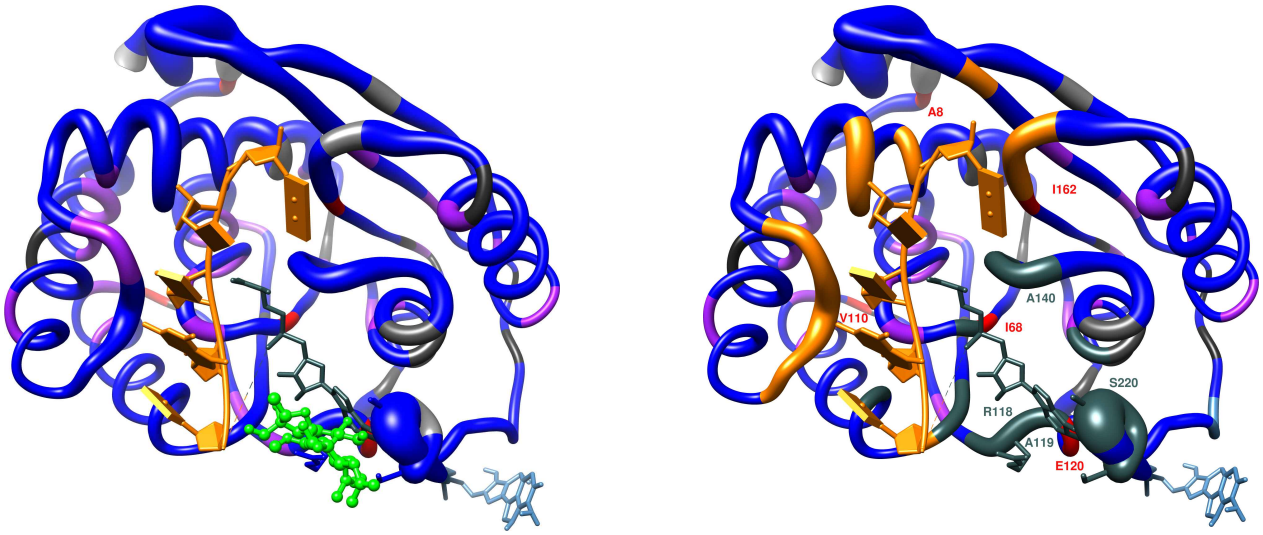


Figure 10: Mutational association with streptomycin resistance and local protein flexibility of *M. tuberculosis* GidB
Mutational landscape of *M. tuberculosis* GidB according to flexibility in GidB according to normal mode analysis (NMA) measuring atomic deformation according to protein dynamics to denote flexibility associated at sites in GidB. The magnitude of flexibility is represented from thin (low flexibility) to thick (high flexibility) tubes. Left panel: The tubes are further coloured to show mutational association with STR resistance, red: resistant sites, blue: sensitive sites, purple: shared sites, black: sites with no SAVs. Right panel: indicates RNA and its interacting residues in orange, SAM interacting residues in dark slate grey, and AMP interacting residues in steel blue. The three resistant sites are labelled with the wild-type residues using the standard one-letter code. Other residues marked are those associated with moderate-to-high flexibility as related to GidB interacting partners. The drug (STR) is hidden here to help highlight the labelled residues. All structure figures were generated using UCSF Chimera version 1.14. Abbreviations used: SAV: single amino acid variation, AMP: adenosine monophosphate, SAM: S-adenosylmethionine, STR: streptomycin.

5.2.5 Relating mutational frequency and biophysical and evolutionary conservational changes

Correlation analysis was performed to understand the relationship between frequently occurring mutations as assessed by MAF and their association with stability (mCSM-DUET, FoldX, DeepDDG, Dynamut2), conservation (ConSurf, SNAP2, PROVEAN) and affinity changes (mCSM-lig/mmCSM-lig, and mCSM-NA), and distance to ligand (Lig-Dist) and nucleic acid (NA-Dist). A combined analysis with all mutations, as well as separately for resistant (R) and sensitive (S) mutations was un-

dertaken **(Figures 11** and **12)**. Analyses focused on determining the strength of association without regard for the direction of the association due to dissimilarity of threshold criteria used by the various estimators.

Frequently occurring sensitive mutations were weakly related to protomer stability changes and distance from the RNA

Frequently occurring mutations were weakly related to protomer stability changes: DeepDDG (ρ_{R+S}=-0.24, P<0.001), FoldX (ρ_{R+S}=0.21,P<0.001), Dynamut2 (ρ_{R+S}=-0.14, P<0.01), and mCSM-DUET (ρ_{R+S}=-0.12, P<0.01) **(Figure 11)** with sensitive mutations driving this association (P<0.05 for sensitive mutations, P>0.05 for resistant mutations) suggesting that frequently occurring sensitive mutations did not introduce strong changes in protomer stability **(Figure 11)**. Frequently occurring mutations were overall weakly associated with distance from STR (ρ_{R+S}=-0.10, P<0.05), and RNA (ρ=-0.18, P<0.001) with sensitive mutations driving the association (P<0.05 for sensitive mutations, P>0.05 for resistant mutations) **(Figure 11)**.

The different computational tools showed good consensus (moderate to strong associations) amongst their predicted estimates, both overall as well as for resistant and sensitive mutation groups individually (0.4≤ ρ_{R+S}<0.8, P<0.001). As expected, mCSM-DUET and Dynamut2 were strongly correlated as these tools share common methodology (ρ_{R+S}=0.74, P<0.001) **(Figure 11)**. Of note, the negative sign associated with FoldX correlations with other predictors is due to the inverse classification criteria used by these tools (Chapter 2: Methods).

Frequently occurring resistant mutations were moderately associated with evolutionary conservation changes

Frequently occurring mutations were moderately related to evolutionary conservation estimates overall (ρ_{R+S}≥0.3, P<0.001) **(Figure 12** left panel**)**. Frequently occurring resistant mutations were moderately associated with rate of evolution according to ConSurf (ρ_{R}=0.40, P<0.05), and conservation of protein function according to SNAP2 (ρ_{R}=0.44, P<0.001), and PROVEAN (ρ_{R}=-0.30, P>0.05). Also, frequently occurring sensitive mutations were moderately related to evolutionary conservation changes concordantly (ρ_{R}≥0.3, P<0.001). There was good agreement (moderate to strong association) between estimates across the three conservation estimators both overall (ρ_{R+S}≥0.6, P<0.001) and in the mutation groups ($\rho_{R/S}$≥0.4, P<0.001) **(Figure 12** left panel**)**.

Frequently occurring mutations were not related to affinity changes

Frequently occurring mutations were not related to STR binding affinity (mCSM- and mmCSM-lig) or RNA binding affinity changes (mCSM-NA) either overall or within the mutation groups (ρ_{R+S}<0.1

and $\rho_{R/S}$<0.1, P>0.05) **(Figure 12** right panel**)**. As expected, estimates from mCSM- and mmCSM-lig were highly correlated, both, overall and within the mutation groups (ρ_{R+S}=1, P<0.001) due to shared underlying methodology **(Figure 12** right panel**)**.

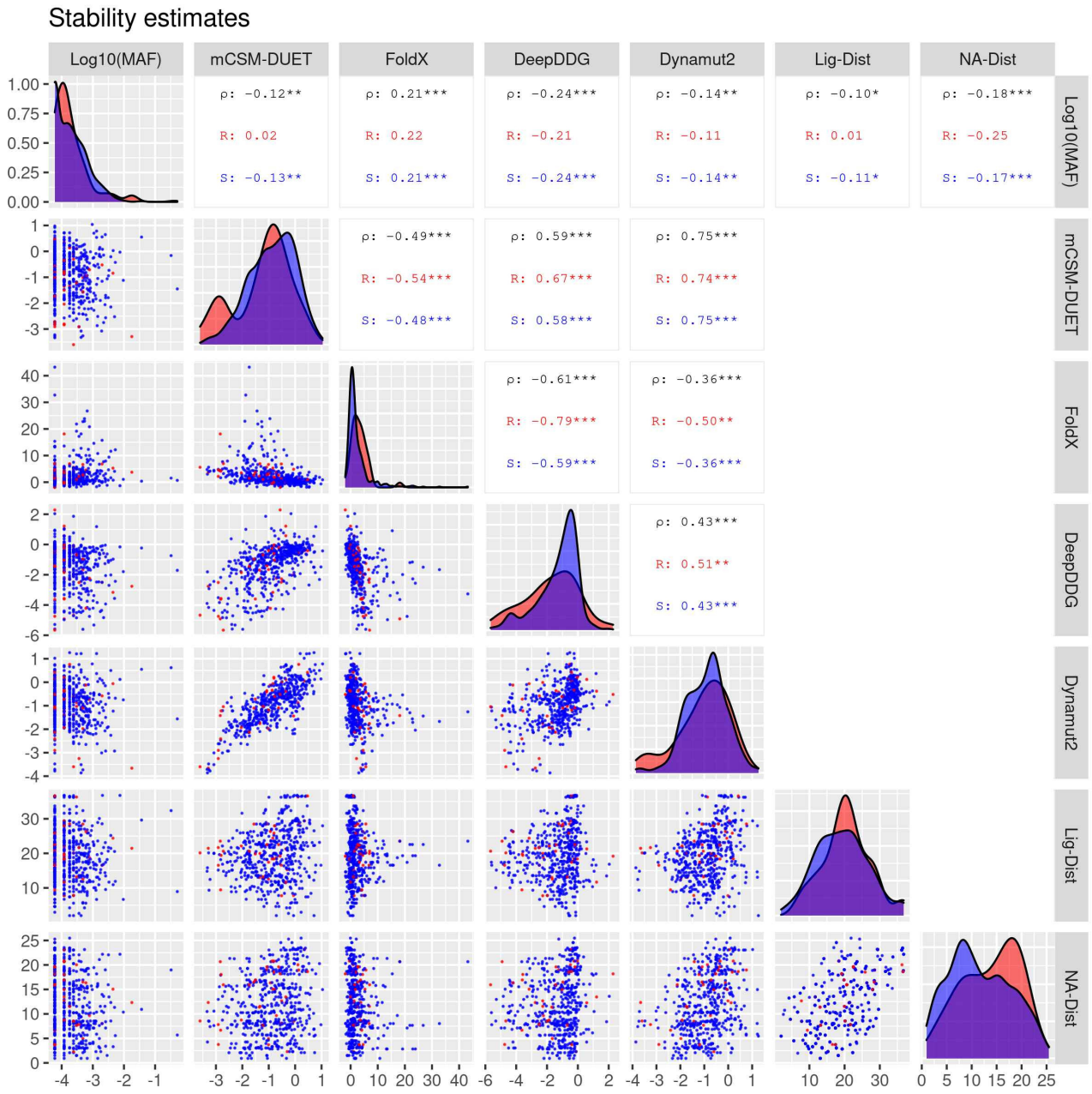


Figure 11: Correlation of protein stability changes and genomics measures
Pairwise correlations between minor allele frequency (MAF), protein stability changes (ΔΔG) estimated using DUET, FoldX, DeepDDG, and Dynamut2, and distances to STR and RNA for 531 SAVs. The upper panel in both plots include the pairwise Spearman (ρ) correlation values along with their statistical significance (.P<0.10, *P<0.05, **P<0.01, ***P<0.001). Three correlation values appear in each plot where black denotes the overall correlation with both resistant (R) and sensitive (S) mutations, while red denotes correlation estimates for resistant mutations, and blue denotes correlation estimates for sensitive mutations individually. The points in the lower panel represent SAVs, where red dots denote resistant mutations and blue represent sensitive mutations. The diagonal in each plot displays the density distribution of the corresponding parameter split by R and S mutation groups. The figure is generated using R statistical software version 4.0.4, ggplot2 package. Abbreviations used: Å: Angstroms, ΔΔG: change in Gibbs free energy in Kcal/mol, SAV: single amino acid variation, Lig-Dist: distance to ligand in Å, NA-Dist: distance to nucleic-acid Å, STR: streptomycin.

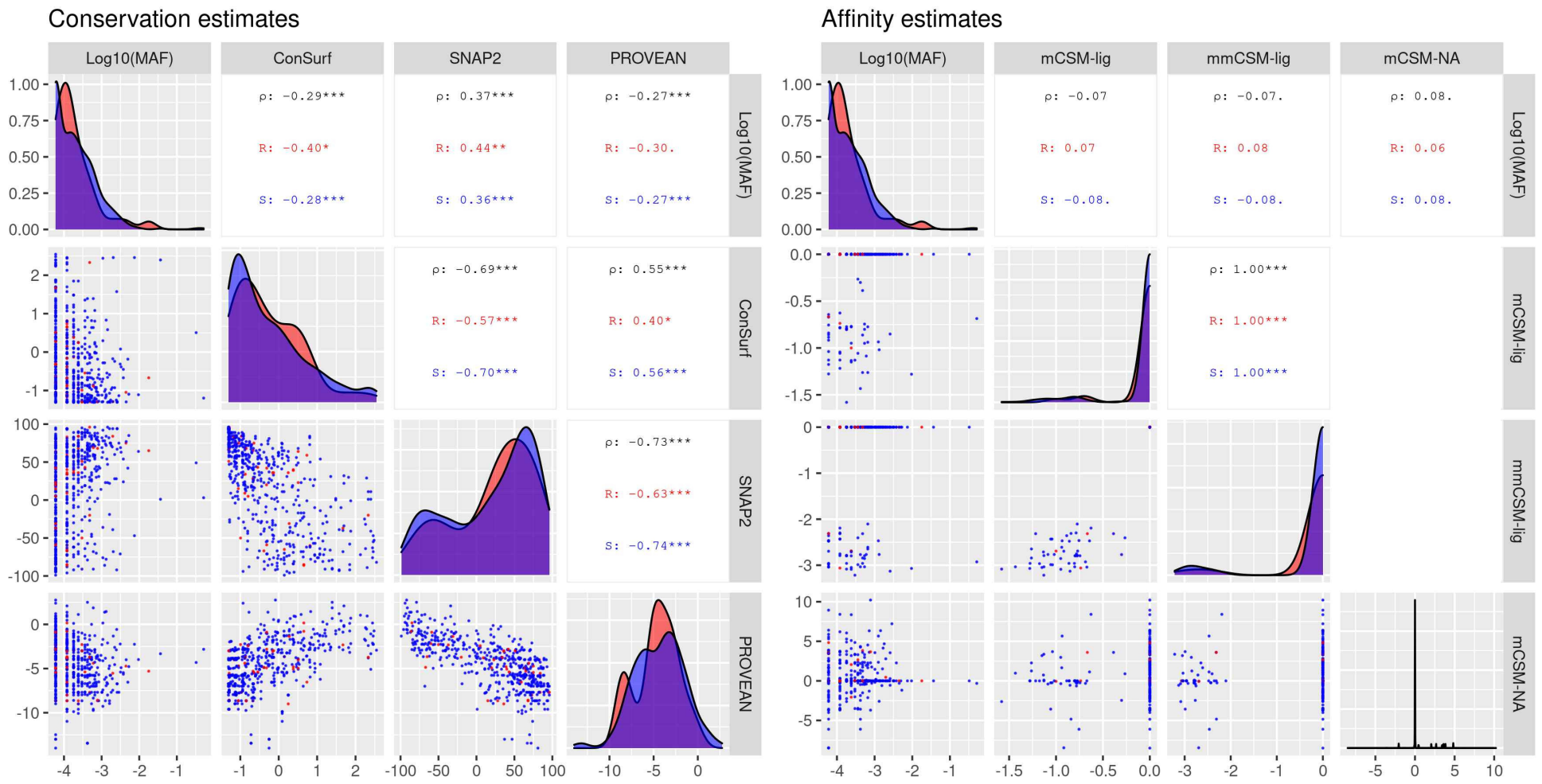


Figure 12: Correlation of evolutionary conservation, affinity changes, and genomics measures

Pairwise correlations of evolutionary conservation, affinity changes, and genomic measure of minor allele frequency (MAF) for 858 SAVs. **Left panel:** Evolutionary conservation predictors: ConSurf, SNAP2, and PROVEAN, **Right panel:** STR binding affinity changes estimated as log fold change (mCSM-lig and mmCSM-lig) of 51 SAVs lying within 10Å of STR, and RNA affinity changes (mCSM-NA) estimated as ΔΔG for 226 SAVs lying within 10Å of the RNA fragment. All corresponding affinity measures for mutations located more than 10Å of STR, and the RNA fragment were given a value of 0 to allow complete SAVs to be used for analysis, while respecting the distance threshold for the respective tools. The upper panel in both plots include the pairwise Spearman (ρ) correlation values along with their statistical significance (.P<0.10, *P<0.05, **P<0.01, ***P<0.001). Three correlation values appear in each plot where black denotes the overall correlation with both resistant (R) and sensitive (S) mutations, while red denotes correlation estimates for resistant mutations, and blue denotes correlation estimates for sensitive mutations individually. The points in the lower panel represent SAVs, where red dots denote resistant mutations and blue represent sensitive mutations. The diagonal in each plot displays the density distribution of the corresponding parameter split by R and S mutation groups. The figure is generated using R statistical software version 4.0.4, ggplot2 package. Abbreviations used: Å: Angstroms, ΔΔG: change in Gibbs free energy in Kcal/mol, SAV: single amino acid variation, Lig-Dist: distance to ligand in Å, NA-Dist: distance to nucleic-acid Å, STR: streptomycin.

5.2.6 Comparing resistant and sensitive mutations

Resistant mutations are marginally destabilising for protomer stability and located further away from RNA interacting sites

Resistant mutations were slightly more destabilising for changes in overall protomer stability compared with sensitive mutations but only according to FoldX ($P<0.01$) **(Figures 13B)**, and not according to mCSM-DUET, DeepDDG, and Dynamut2 ($P>0.05$) **(Figures 13A, 13C, 13D)**. Resistant and sensitive mutations are likely to occur with a similar frequency ($P>0.05$) **(Figure 13E)**. Resistant mutations were not closer to the drug binding site ($P>0.05$) **(Figure 13F)** but occurred further away from the nucleic acid (RNA) interacting sites with sensitive mutations occurring closer to sites interacting with RNA ($P<0.05$) **(Figure 13G)**. While there were no differences in mutational impact resulting from ligand and nucleic acid binding affinity changes ($P<0.05$) these results are inconclusive due to low numbers in the resistant group. There were only 3 resistant mutations (versus 48 sensitive mutations) for mCSM/mmCSM-lig analyses, and 12 resistant mutations (versus 214 sensitive mutations) for mCSM-NA analyses **(Figures 13K, 13L, 13M)**. All measures of evolutionary conservation were also statistically insignificant ($P>0.05$) **(Figures 13H**, **13I** and **13J)**.

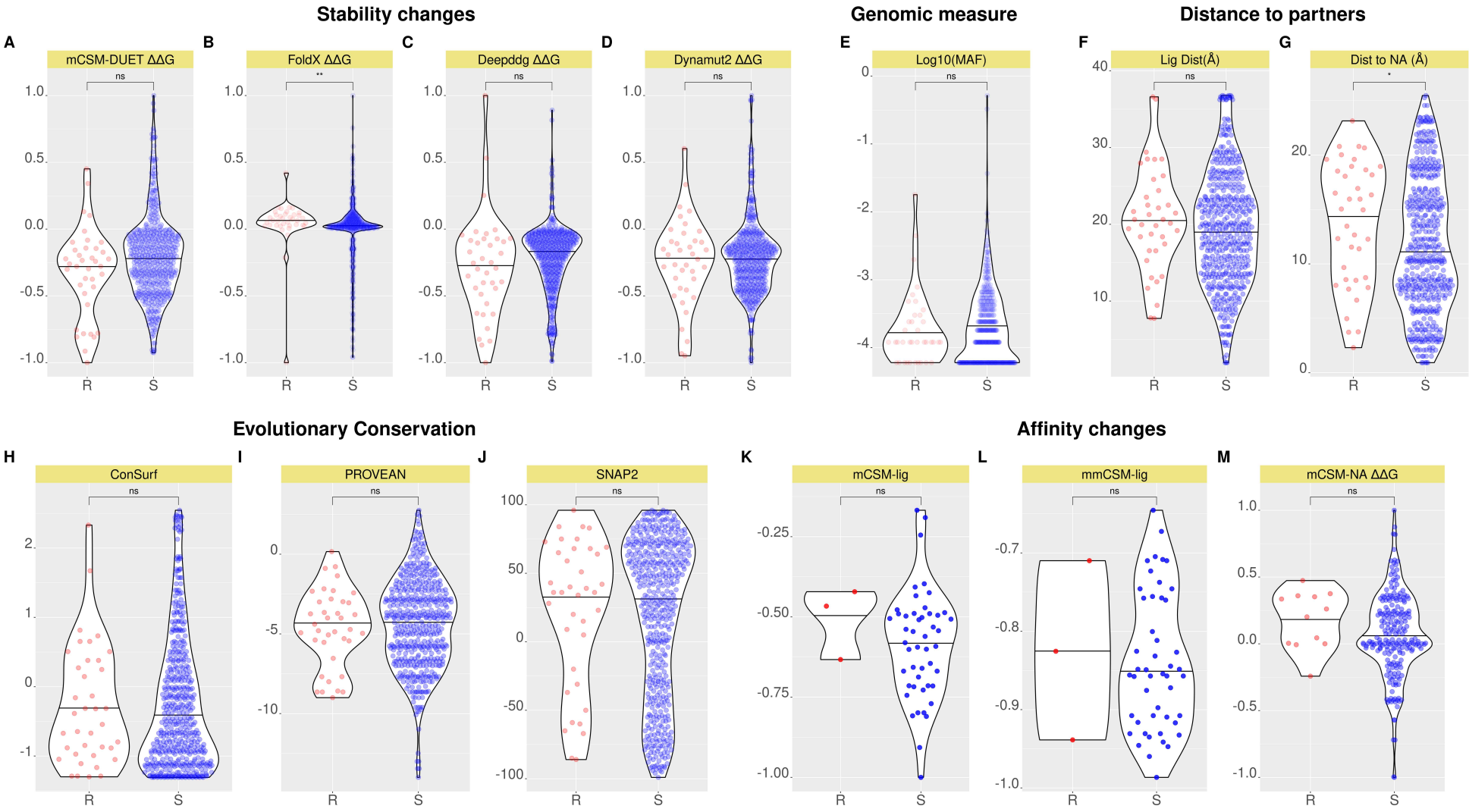

Stability changes
Genomic measure
Distance to partners
A
mCSM-DUET ΔΔG
B
FoldX ΔΔG
C
Deepddg ΔΔG
D
Dynamut2 ΔΔG
E
Log10(MAF)
F
Lig Dist(Å)
G
Dist to NA (Å)
Evolutionary Conservation
Affinity changes
H
ConSurf
I
PROVEAN
J
SNAP2
K
mCSM-lig
L
mmCSM-lig
M
mCSM-NA ΔΔG
ns
R
S

Figure 13: Comparison of resistant (R) and sensitive (S) mutations
Violin plots showing the distribution of features related to structural properties, genomic measure, evolutionary conservation for 531 SAVs. For affinity changes related to ligand (STR) measured by mCSM- and mmCSM-lig, only those mutations within 10Å of STR (n=51) were considered. Similarly, for nucleic acid (NA) affinity changes measured by mCSM-NA, only those mutations within 10Å of the RNA fragment (n=226) were considered. Mutations were grouped as either resistant (R, n=38) or sensitive (S, n=493), and were compared using the Wilcoxon rank-sum (unpaired) test, with statistical significance indicated as: *P $<$ 0.05, **P $<$ 0.01, ***P $<$ 0.001, ****P $<$ 0.0001). Mutations in the resistant group appear as red dots, while those in the sensitive group appear as blue dots, and the horizontal line in the violin plots display the median value. The two mutations groups were compared based on **A-D)** Stability changes ($\Delta\Delta$G) estimated from four computational tools: mCSM-DUET, FoldX, DeepDDG and Dynamut2, **E)** genomic measure of average mutational occurrence (Log10MAF), **F-G)** Distance to ligand (Lig-Dist) and Distance to Nucleic acid (Distance to NA), **H-J)**Evolutionary conservation measured by ConSurf (<0: Conserved, >0: Variable), PROVEAN (>-2.5: Neutral, < -2.5: Deleterious) and SNAP2 (<=0: Neutral, >0: Effect) computational tools, **K-L)** Comparison of STR binding affinity changes from mCSM-lig and mmCSM-lig measured as log fold change for R (n=3) and S (n=48) mutations, **M)** RNA binding affinity changes (mCSM-NA) estimated as $\Delta\Delta$G for R (n=12) and S (n=214) mutations. The figure is generated using R statistical software version 4.0.4. Abbreviations used: Å: Angstroms, $\Delta\Delta$G: change in Gibbs free energy in Kcal/mol, SAV: single amino acid variation, ns: not-significant, STR: streptomycin, MAF: minor allele frequency, Lig-Dist: distance to ligand in Å, NA-Dist: distance to nucleic acid in Å, R: resistant mutations, S: sensitive mutations.

5.2.7 Associating mutations with Odds Ratio and extreme effects

Mutations with high OR are not restricted to GidB binding partners

Based on DST data available for 268 (out of 531) SAVs, mutational association with resistance was further estimated using Odds Ratio (OR), with values above 1 suggesting association with STR resistance. The higher the OR, the greater the likelihood of a given mutation being resistant. This resulted in nearly 50% (n=130/268) of mutations predicted to be associated with STR resistance, much higher than observed in our data (7%, n=38/531).

An overview of mutations in GidB shows that sites with high OR are not restricted to residues interacting with GidB binding partners **(Figure 14)**. The mutations with the highest OR (OR=23.44) were A134E and L44Q **(Figure 14, Table 1).** Of the sites that were interacting with one or more binding partners, sites interacting with RNA displayed mutations with higher OR compared with those interacting with SAM, while sites interacting with AMP were not associated with resistance (OR<1) **(Figure 14)**.

Other sites with prominent association with resistance like A200, G76, G71, P75, T146, A8, D85, C52, A180 were not involved with any GidB binding partners. The only notable exception is G34E that is involved in interaction with RNA.

Mutations with extreme effects primarily affect SAM, and RNA but not AMP

The most frequently occurring mutation, E92D (MAF ∼52%) and the most destabilising mutation for STR affinity, R118L, are involved with SAM interactions. The most destabilising mutation for nucleic acid affinity was W45G, site not involved with any GidB binding partners. The most stabilising mutation for nucleic acid affinity was G34W, which is involved in RNA interactions **(Table 1)**.

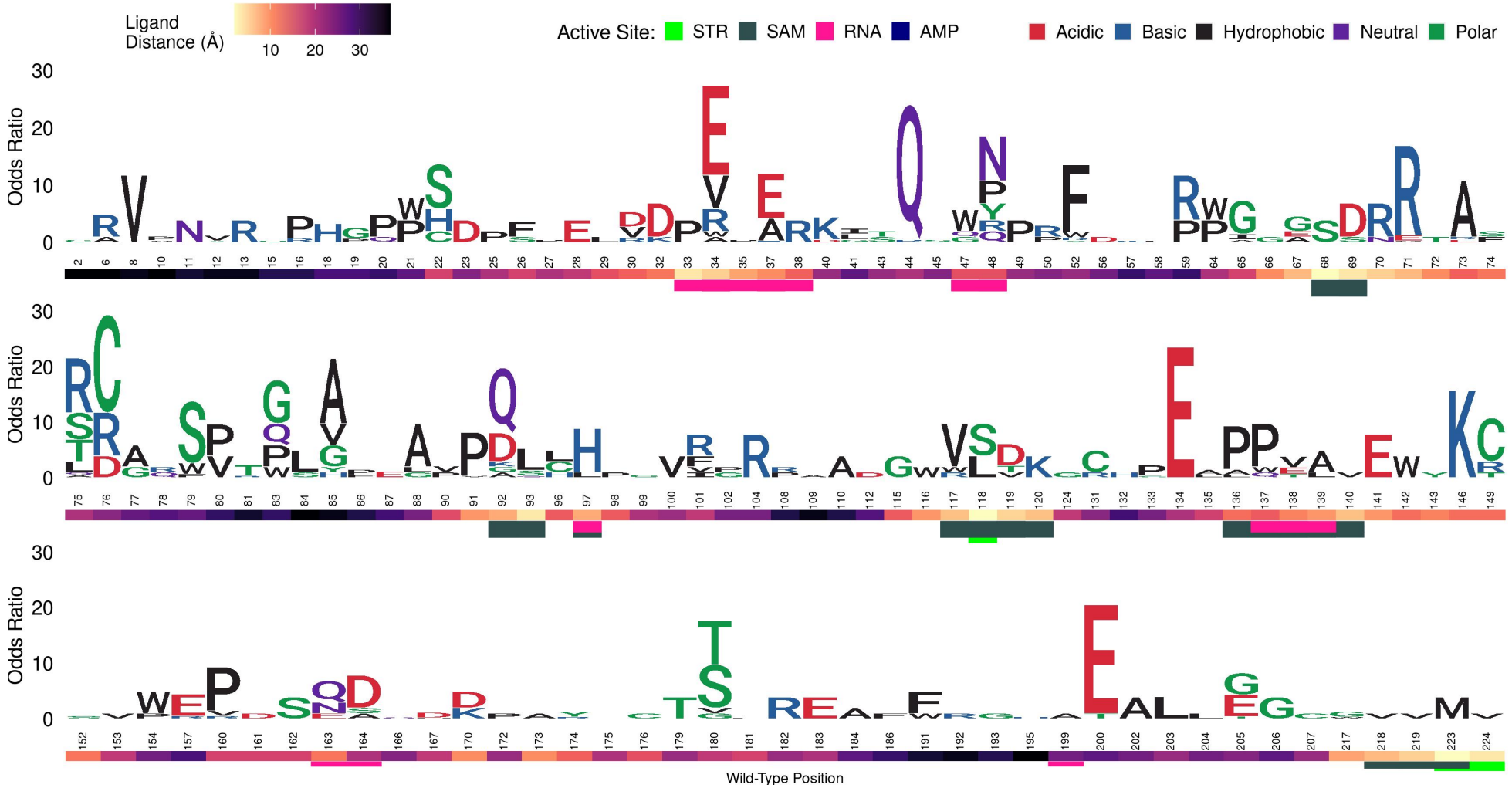


Figure 14: Logo plot showing mutational sites and their association with resistance according to Odds Ratio

Logo plot showing 268 SAVs by mutational site according to their association with STR resistance calculated using Odds Ratio (OR). The vertical axis represents the OR where letters denote mutant residues which are proportional to their corresponding OR, highlighting the most resistant mutation at each site and overall. The mutant residues are coloured according to the amino acid (aa) properties as denoted where red denotes acidic aa, basic aa appear in blue, hydrophobic aa in black, neutral aa in purple, and polar aa in darkgreen. The structural positions associated with SAVs with OR are indicated on the horizontal axis. The heat bar underneath the positions indicate the distance of that position from STR according to the magma colour gradient where light yellow indicates sites closer to STR (ligand distance in Angstroms). The positions are further annotated to reflect residues involved in interactions with binding partners: STR (green), RNA (deep pink), SAM (dark slate grey), and AMP (navy blue). The figure is generated using R statistical software version 4.0.2, ggplot2 package. Abbreviations used: SAV: single amino acid variation, AMP: adenosine monophosphate, SAM: S-adenosylmethionine, STR: streptomycin.

Mutation	Mutational effect	Mutational effect value	Lig-Dist (Å)	NA-Dist (Å)	Interacting partner
A134E	Mutations with the highest OR	OR = 23.44	16.85	13.60	none
L44Q		OR = 23.44	21.67	7.82	none
E92D	Most frequent mutation	MAF (%) = 51.60	8.94	5.69	SAM
Y22S	Most Destabilising for protomer	ΔΔG = -0.66	20.14	8.56	none
R96L	Most Stabilising for protomer	ΔΔG = 0.58	13.08	5.87	none
R118L	Most Destabilising for STR binding affinity	Log fold change = -0.97	1.98	5.46	SAM
W45G	Most Destabilising for RNA affinity	ΔΔG = -8.49	25.25	9.42	none
G34W	Most Stabilising for RNA affinity	ΔΔG = 10.22	11.82	3.95	RNA

Table 1: Mutations with extreme effects
Mutations (SAVs) with extreme effects related to Odds Ratio (OR), mutational frequency,stability and affinity changes. For affinity changes only mutations within 10Å of STR and RNA for their respective binding affinities were considered. The protomer stability changes are the average effect of all four estimates (mCSM-DUET, FoldX, DeepDDG and Dynamut2) combined, and the STR binding affinity changes are the average effect of the two mCSM based tools (mCSM-lig and mmCSM-lig) combined. Changes in RNA binding affinity correspond to estimates from mCSM-NA. The estimated effects were categorised as Destabilising (log fold affinity change/ΔΔG<0) and Stabilising (log fold affinity change/ΔΔG>0). Abbreviations used: Å: Angstroms, ΔΔG: change in Gibbs free energy in Kcal/mol, MAF: minor allele frequency, SAV: single amino acid variation, Lig-Dist: distance to ligand, NA: nucleic acid, NA-Dist: distance to nucleic acid, STR: streptomycin.

5.2.8 Relating lineage and protomer stability

Lineages 1 and 3 have high SAV diversity, with stabilising sensitive mutations overrepresented in Lineage 4

About 50% of samples (n=18,584) consisted of SAVs in the protein coding region of *gidB*, where 18,252 samples contributed to the four main *M. tuberculosis* lineages (Lineages 1-4). Most samples with GidB mutations belonged to lineage 2 (n=9,465), followed by lineage 4 (n=7,372),lineage 1 (n=747) and finally by lineage 3 with the least number of samples (n=668) **(Figure 15A)**. However, Lineages 1 and 3 were nearly equal when assessing SAV diversity (Lineage 1: 26%, n=196; Lineage 3: 25%, n=430). Lineage 4 showed 5% (n=372) while lineage 2 showed only 1% (n=96) SAV diversity despite contributing the highest numbers of samples **(Figure 15B)**. Resistant mutations for all lineages showed prominent peaks around the highly destabilising protomer stability (ΔΔG ∼ -0.8 Kcal/mol). Resistant mutations were prominently bimodal for lineages 4 and 3 with additional peaks around mildly stabilising ΔΔG (-0.25 Kcal/mol). Lineage 1 was multimodal with additional peaks around the moderately stabilising (ΔΔG ∼ -0.4 Kcal/mol) and around marginal stability (ΔΔG ∼ -0.2 Kcal/mol) **(Figure 15C)**.

Sensitive mutations were most pronounced in lineages 4 and 2. While lineage 4 showed a prominent peak around the moderately stabilising ΔΔG values (∼0.35 Kcal/mol), lineage 2 showed a similar

peak but towards the moderately destabilising ΔΔG values (∼ -0.6 Kcal/mol). Lineage 1 peaked around the mildly destabilising (ΔΔG ∼ -0.25), but spanned a wider range of protomer stability estimates. Lineage 3 was bimodal with peaks around mildly destabilising (ΔΔG ∼ -0.25 Kcal/mol) and mildly stabilising (ΔΔG ∼0.25 Kcal/mol) protomer stability values **(Figure 15C)**. Overall lineage distributions were significantly different between all lineages (adjusted P<0.0001), as well as in a given lineage between resistant and sensitive mutation distributions (adjusted P<0.0001) except for lineage 1 (adjusted P>0.05) (Appendix Table 5.C.1).

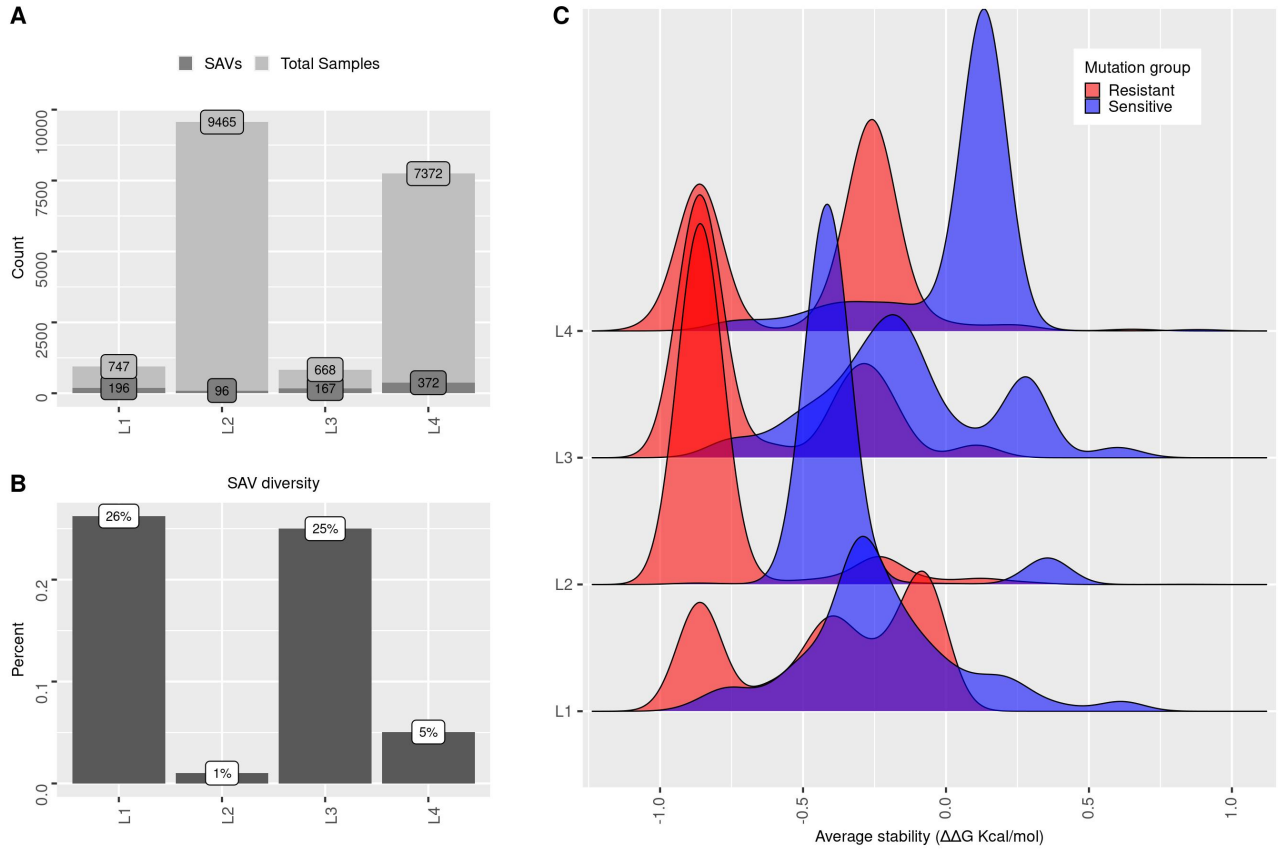


Figure 15: Lineage and protomer stability distribution
Total number of samples (n=18,252) along with the number of mutations associated with STR resistance in the four *M. tuberculosis* lineages (L1-L4). **A)** The dark grey bars show the number of mutations (SAVs), while the light grey bar show the total number of samples in each lineage, **B)** Mutational diversity in each lineage, **C)** Density distribution of lineages according to average protein stability changes (ΔΔG).Estimates from four different computational tools: mCSM-DUET, FoldX, DeepDDG,and Dynamut2 were combined to calculate the average mutational stability impact for each SAV. The horizontal axis shows the average stability values (-1: highly destabilising and +1:highly stabilising) further coloured by mutational association with STR resistance: Red denotes resistant mutations (n=38, from 510 samples) and blue indicates sensitive mutations (n=493, from 17742 samples). The figure is generated using R statistical software version 4.0.4. Abbreviations used: ΔΔG: change in Gibbs free energy, SAV: single amino acid variation, STR: streptomycin.

5.3 Chapter summary

The resistance profile for GidB is evolving, with triple SAVs being the most frequent, and extending to sites beyond those involved in GidB binding partners. Most mutations occur in conserved regions

with sites around RNA, SAM and STR with multiple SAVs displaying moderate-to-high flexibility. The overall mutational effect on sites around the RNA are stabilising without affecting RNA binding affinity. Resistance in *gidB* is underestimated from DST (7%), with GWAS inference predicting nearly 50% of mutations as resistant. As such, resistance hotspots can be located away from GidB binding partners. This mutational promiscuity comes without a large fitness penalty as STR does not directly bind to GidB, and mutations are unlikely to affect protein function.

References

[1] T. M. Daniel. "Selman Abraham Waksman and the Discovery of Streptomycin". In: *The International Journal of Tuberculosis and Lung Disease: The Official Journal of the International Union Against Tuberculosis and Lung Disease* 9.2 (Feb. 2005), pp. 120–122. ISSN: 1027-3719.

[2] Rolf Zetterström. "Selman A. Waksman (1888-1973) Nobel Prize in 1952 for the Discovery of Streptomycin, the First Antibiotic Effective against Tuberculosis". In: *Acta Paediatrica (Oslo, Norway: 1992)* 96.2 (Feb. 2007), pp. 317–319. ISSN: 0803-5253. DOI: 10.1111/j.1651-2227.2007.00182.x.

[3] D. Moazed and H. F. Noller. "Interaction of Antibiotics with Functional Sites in 16S Ribosomal RNA". In: *Nature* 327.6121 (June 4, 1987–10), pp. 389–394. ISSN: 0028-0836. DOI: 10.1038/327389a0.

[4] M. Finken et al. "Molecular Basis of Streptomycin Resistance in Mycobacterium Tuberculosis: Alterations of the Ribosomal Protein S12 Gene and Point Mutations within a Functional 16S Ribosomal RNA Pseudoknot". In: *Molecular Microbiology* 9.6 (Sept. 1993), pp. 1239–1246. ISSN: 0950-382X. DOI: 10.1111/j.1365-2958.1993.tb01253.x.

[5] J. S. Verma et al. "Evaluation of gidB Alterations Responsible for Streptomycin Resistance in Mycobacterium Tuberculosis". In: *Journal of Antimicrobial Chemotherapy* 69.11 (Nov. 1, 2014), pp. 2935–2941. ISSN: 0305-7453, 1460-2091. DOI: 10.1093/jac/dku273.

[6] Susumu Okamoto et al. "Loss of a Conserved 7-Methylguanosine Modification in 16S rRNA Confers Low-Level Streptomycin Resistance in Bacteria". In: *Molecular Microbiology* 63.4 (Feb. 2007), pp. 1096–1106. ISSN: 0950-382X. DOI: 10.1111/j.1365-2958.2006.05585.x.

[7] J. Crofton. "The MRC Randomized Trial of Streptomycin and Its Legacy: A View from the Clinical Front Line". In: *Journal of the Royal Society of Medicine* 99.10 (Oct. 2006), pp. 531–534. ISSN: 0141-0768. DOI: 10.1258/jrsm.99.10.531.

[8] J Crofton and DA Mitchison. "Streptomycin Resistance in Pulmonary Tuberculosis". In: *British medical journal* 2.4588 (Dec. 1948), pp. 1009–1015. ISSN: 0007-1447. DOI: 10.1136/bmj.2.4588.1009.

[9] J. Nair et al. "The rpsL Gene and Streptomycin Resistance in Single and Multiple Drug-Resistant Strains of Mycobacterium Tuberculosis". In: *Molecular Microbiology* 10.3 (Nov. 1993), pp. 521–527. ISSN: 0950-382X. DOI: 10.1111/j.1365-2958.1993.tb00924.x.

[10] R C Cooksey et al. "Characterization of Streptomycin Resistance Mechanisms among Mycobacterium Tuberculosis Isolates from Patients in New York City". In: *Antimicrobial Agents and Chemotherapy* 40.5 (May 1996), pp. 1186–1188. ISSN: 0066-4804, 1098-6596. DOI: 10.1128/AAC.40.5.1186.

[11] Sharon Y. Wong et al. "Mutations in gidB Confer Low-Level Streptomycin Resistance in Mycobacterium Tuberculosis". In: *Antimicrobial Agents and Chemotherapy* 55.6 (June 2011), pp. 2515–2522. ISSN: 1098-6596. DOI: 10.1128/AAC.01814-10.

[12] Henry D. Isenberg. "Antimicrobial Susceptibility Testing: A Critical Evaluation". In: *Journal of Antimicrobial Chemotherapy* 22 (Supplement_A July 1988), pp. 73–86. ISSN: 1460-2091, 0305-7453. DOI: 10.1093/jac/22.Supplement_A.73.

[13] Nat Smittipat et al. “Mutations in Rrs, rpsL and gidB in Streptomycin-Resistant Mycobacterium Tuberculosis Isolates from Thailand”. In: *Journal of Global Antimicrobial Resistance* 4 (Mar. 2016), pp. 5–10. ISSN: 2213-7173. DOI: 10.1016/j.jgar.2015.11.009.

[14] Tomasz Jagielski. “Screening for Streptomycin Resistance-Conferring Mutations in Mycobacterium Tuberculosis Clinical Isolates from Poland.” In: (June 17, 2014).

[15] FS Spies et al. “Streptomycin Resistance and Lineage-Specific Polymorphisms in Mycobacterium Tuberculosis gidB Gene”. In: *Journal of Clinical Microbiology* 49.7 (2011), pp. 2625–2630. ISSN: 0095-1137. DOI: 10.1128/JCM.00168-11.

[16] Bashir A. Sheikh et al. “Development of New Therapeutics to Meet the Current Challenge of Drug Resistant Tuberculosis”. In: *Current Pharmaceutical Biotechnology* 22.4 (Mar. 2021), pp. 480–500. ISSN: 13892010. DOI: 10.2174/1389201021666200628021702.

[17] Steven T. Gregory et al. “Structural and Functional Studies of the Thermus Thermophilus 16S rRNA Methyltransferase RsmG”. In: *RNA* 15.9 (Jan. 9, 2009), pp. 1693–1704. ISSN: 1355-8382, 1469-9001. DOI: 10.1261/rna.1652709.

[18] Hasan Demirci et al. “A Structural Basis for Streptomycin-Induced Misreading of the Genetic Code”. In: *Nature Communications* 4.1 (1 Jan. 15, 2013), p. 1355. ISSN: 2041-1723. DOI: 10.1038/ncomms2346.

Appendix for Chapter 5

5.A Mutations close to streptomycin

Mutation	Lig-Dist (Å)	mCSM-lig affinity	mCSM-lig outcome	mmCSM-lig affinity	mmCSM-lig outcome	MAF (%)	Odds Ratio	P-value	Adjusted P-value	Adjusted P-value significance
E92Q	8.94	-0.94	Destabilising	-2.61	Destabilising	0.07	11.7	0.04	0.47	ns
G117V	7.81	-1	Destabilising	-2.69	Destabilising	0.02	7.8	0.11	0.65	ns
R97H	8.88	-0.27	Destabilising	-2.41	Destabilising	0.04	7.8	0.11	0.65	ns
R118S	1.98	-1.43	Destabilising	-3.04	Destabilising	0.04	5.85	0.12	0.65	ns
G69D	9.54	-0.85	Destabilising	-3.13	Destabilising	0.26	5.85	0.12	0.65	ns
E92D	8.94	-0.69	Destabilising	-2.93	Destabilising	51.6	5.09	<0.0001	<0.0001	****
E120K	7.75	-0.74	Destabilising	-3.06	Destabilising	0.01	3.9	0.34	0.8	ns
I68S	9.45	-0.67	Destabilising	-2.32	Destabilising	0.01	3.9	0.34	0.8	ns
R118L	1.98	-1.58	Destabilising	-3.08	Destabilising	0.02	3.9	0.34	0.8	ns
V139A	9.85	-0.78	Destabilising	-2.8	Destabilising	0.02	3.9	0.34	0.8	ns
T223M	2.62	-1.06	Destabilising	-2.99	Destabilising	0.04	3.9	0.27	0.8	ns
P93L	5.58	-1.09	Destabilising	-2.77	Destabilising	0.06	3.9	0.34	0.8	ns
A119D	6.66	-1.02	Destabilising	-2.77	Destabilising	0.13	3.9	0.34	0.8	ns
A119T	6.66	-1.28	Destabilising	-3.07	Destabilising	0.95	1.19	0.7	>1	ns
G117W	7.81	-0.63	Destabilising	-2.19	Destabilising	0.01	0.97	>1	>1	ns
A140V	8.18	-1.13	Destabilising	-2.79	Destabilising	0.01	0.97	>1	>1	ns
R217G	8.43	-1.22	Destabilising	-2.32	Destabilising	0.01	0.97	>1	>1	ns
A119V	6.66	-1.28	Destabilising	-2.96	Destabilising	0.01	0.97	>1	>1	ns
E92A	8.94	-0.77	Destabilising	-2.79	Destabilising	0.02	0.97	>1	>1	ns
E92G	8.94	-0.77	Destabilising	-2.69	Destabilising	0.02	0.97	>1	>1	ns
P93S	5.58	-1.15	Destabilising	-3.05	Destabilising	0.02	0.97	>1	>1	ns
M218V	7.4	-0.79	Destabilising	-2.8	Destabilising	0.02	0.97	>1	>1	ns

A224V	4.23	-0.8	Destabilising	-2.96	Destabilising	0.02	0.97	>1	>1	ns
G69S	9.54	-1.13	Destabilising	-3.22	Destabilising	0.03	0.97	>1	>1	ns
R97L	8.88	-0.3	Destabilising	-2.77	Destabilising	0.04	0.97	>1	>1	ns
E92K	8.94	-0.95	Destabilising	-2.86	Destabilising	0.05	0.97	>1	>1	ns
V139L	9.85	-0.75	Destabilising	-2.8	Destabilising	0.05	0.97	>1	>1	ns
A219V	5.75	-0.86	Destabilising	-2.96	Destabilising	0.05	0.97	>1	>1	ns
G117R	7.81	-1.26	Destabilising	-2.65	Destabilising	0.07	0.97	>1	>1	ns
P93Q	5.58	-1.26	Destabilising	-2.84	Destabilising	0.05	0.32	0.56	0.99	ns
R217W	8.43	-0.77	Destabilising	-2.11	Destabilising	0.07	0.32	0.56	0.99	ns
G69I	9.54	-1.04	Destabilising	-2.43	Destabilising	0.01	NA	NA	NA	ns
G69V	9.54	-1.06	Destabilising	-2.43	Destabilising	0.01	NA	NA	NA	ns
L94P	6.71	-0.78	Destabilising	-2.75	Destabilising	0.01	NA	NA	NA	ns
R97C	8.88	-0.64	Destabilising	-2.47	Destabilising	0.01	NA	NA	NA	ns
A119G	6.66	-1.13	Destabilising	-3.04	Destabilising	0.01	NA	NA	NA	ns
V139M	9.85	-0.79	Destabilising	-2.3	Destabilising	0.01	NA	NA	NA	ns
A140S	8.18	-1.04	Destabilising	-2.99	Destabilising	0.01	NA	NA	NA	ns
W148C	4.61	-0.84	Destabilising	-2.31	Destabilising	0.01	NA	NA	NA	ns
W148L	4.61	-0.92	Destabilising	-2.7	Destabilising	0.01	NA	NA	NA	ns
W148R	4.61	-0.76	Destabilising	-2.46	Destabilising	0.01	NA	NA	NA	ns
R217L	8.43	-1.18	Destabilising	-2.97	Destabilising	0.01	NA	NA	NA	ns
S220R	2.96	-0.67	Destabilising	-3.04	Destabilising	0.01	NA	NA	NA	ns
G221R	5.71	-0.86	Destabilising	-2.48	Destabilising	0.01	NA	NA	NA	ns
G221V	5.71	-0.89	Destabilising	-2.36	Destabilising	0.01	NA	NA	NA	ns
G117E	7.81	-0.95	Destabilising	-3.09	Destabilising	0.02	NA	NA	NA	ns
R118H	1.98	-1.13	Destabilising	-2.71	Destabilising	0.02	NA	NA	NA	ns
M218I	7.4	-0.8	Destabilising	-2.8	Destabilising	0.02	NA	NA	NA	ns

L145S	9.33	-0.39	Destabilising	-2.32	Destabilising	0.05	NA	NA	NA	ns
A224G	4.23	-0.8	Destabilising	-3.04	Destabilising	0.05	NA	NA	NA	ns
L145F	9.33	-0.94	Destabilising	-2.47	Destabilising	0.13	NA	NA	NA	ns

Table 5.A.1: Mutations close to STR
Fifty one single amino acid variation (SAV) mutations lying within 10Å of STR and their corresponding ligand affinity changes (log fold change) measured by mCSM-Lig and mmCSM-lig. The estimated effect are categorised as Destabilising (log fold affinity change<0) and Stabilising ($\Delta\Delta G>0$). The genomic measures of minor allele frequency (MAF), Odds Ratio (OR), OR related P-values, and FDR adjusted P-values are shown. Statistical significance is indicated as: *P < 0.05, **P < 0.01, ***P < 0.001, ****P < 0.0001, ns:>0.05. The table is arranged by OR to show mutation with the highest OR at the top for mutations close to STR. Columns with NA indicate insufficient data to calculate OR. Abbreviations used: FDR: false discovery rate, ns: not significant, STR: streptomycin.

5.B Mutations close to the nucleic acid

Mutation	NA-Dist (Å)	mCSM-NA (ΔΔG)	mCSM-NA outcome	MAF (%)	Odds Ratio	P-value	Adjusted P-value	Adjusted P-value significance
L44Q	7.82	2.66	Increased affinity	0.04	23.44	<0.001	0.04	*
A200E	4.26	-0.18	Reduced affinity	0.1	19.52	<0.05	0.11	ns
G76C	9.67	3.67	Increased affinity	0.08	17.59	<0.001	0.02	*
G71R	4.97	3.84	Increased affinity	0.05	15.61	0.01	0.25	ns
G34E	3.95	1.46	Increased affinity	0.1	15.61	0.01	0.25	ns
C52F	7.66	2.03	Increased affinity	0.05	11.7	0.04	0.47	ns
E92Q	5.69	1.66	Increased affinity	0.07	11.7	0.04	0.47	ns
P75R	7.12	2.29	Increased affinity	0.08	9.76	0.02	0.32	ns
G73A	8.03	0.46	Increased affinity	0.2	9.76	0.02	0.32	ns
Y22S	8.56	-2.06	Reduced affinity	0.02	7.8	0.11	0.65	ns
G37E	2.3	2.06	Increased affinity	0.02	7.8	0.11	0.65	ns
G117V	8.49	-0.06	Reduced affinity	0.02	7.8	0.11	0.65	ns
G76R	9.67	2.38	Increased affinity	0.04	7.8	0.11	0.65	ns
R97H	3.93	0.9	Increased affinity	0.04	7.8	0.11	0.65	ns
A141E	6.12	-0.44	Reduced affinity	0.05	7.8	0.11	0.65	ns
H48N	3.04	0.1	Increased affinity	0.06	7.8	0.11	0.65	ns
R137P	0.93	-2.13	Reduced affinity	0.07	7.8	0.05	0.54	ns
S136P	6.84	-3.7	Reduced affinity	0.08	7.8	0.11	0.65	ns
L91P	8.54	0.01	Increased affinity	0.11	7.8	0.11	0.65	ns
S70R	5.57	-1.18	Reduced affinity	0.53	5.86	0.02	0.34	ns
R118S	5.46	0.97	Increased affinity	0.04	5.85	0.12	0.65	ns
G164D	2.84	-0.45	Reduced affinity	0.05	5.85	0.12	0.65	ns
G34V	3.95	1.75	Increased affinity	0.08	5.85	0.12	0.65	ns

E32D	7.12	0.18	Increased affinity	0.11	5.85	0.12	0.65	ns
G69D	3.39	-0.56	Reduced affinity	0.26	5.85	0.12	0.65	ns
E92D	5.69	-0.32	Reduced affinity	51.6	5.09	<0.0001	<0.0001	****
P75S	7.12	3.55	Increased affinity	0.26	4.88	0.05	0.54	ns
V202A	8.74	0	Increased affinity	0.01	3.9	0.34	0.8	ns
P38R	3.72	4.84	Increased affinity	0.01	3.9	0.34	0.8	ns
I68S	6.67	3.6	Increased affinity	0.01	3.9	0.34	0.8	ns
I162S	3.76	3.42	Increased affinity	0.01	3.9	0.34	0.8	ns
Y22H	8.56	-2.25	Reduced affinity	0.02	3.9	0.34	0.8	ns
R33P	2.08	-1.37	Reduced affinity	0.02	3.9	0.34	0.8	ns
G34R	3.95	4.08	Increased affinity	0.02	3.9	0.34	0.8	ns
H48P	3.04	-1.54	Reduced affinity	0.02	3.9	0.34	0.8	ns
R118L	5.46	-2.62	Reduced affinity	0.02	3.9	0.34	0.8	ns
V139A	2.88	-0.27	Reduced affinity	0.02	3.9	0.34	0.8	ns
L142W	7.41	8.36	Increased affinity	0.02	3.9	0.34	0.8	ns
K163Q	3.12	-1.51	Reduced affinity	0.02	3.9	0.34	0.8	ns
C191F	6.92	2.26	Increased affinity	0.02	3.9	0.34	0.8	ns
E40K	8.11	3.32	Increased affinity	0.03	3.9	0.34	0.8	ns
L49P	8.05	0.02	Increased affinity	0.04	3.9	0.34	0.8	ns
G76D	9.67	-0.26	Reduced affinity	0.04	3.9	0.34	0.8	ns
T223M	9.9	-3.57	Reduced affinity	0.04	3.9	0.27	0.8	ns
G37A	2.3	2.34	Increased affinity	0.05	3.9	0.27	0.8	ns
P93L	4.51	-0.4	Reduced affinity	0.06	3.9	0.34	0.8	ns
P75T	7.12	3.56	Increased affinity	0.1	3.9	0.19	0.8	ns
R47W	1.51	7.21	Increased affinity	0.29	3.74	<0.001	0.01	*
E170D	7.68	0.04	Increased affinity	0.03	2.92	0.34	0.8	ns

H48Y	3.04	4.14	Increased affinity	0.17	2.92	0.34	0.8	ns
L26F	7.62	6.33	Increased affinity	0.07	2.6	0.24	0.8	ns
G30D	8.02	0.99	Increased affinity	0.12	2.6	0.24	0.8	ns
D67G	8.41	0.28	Increased affinity	0.26	2.6	0.24	0.8	ns
G30V	8.02	1.29	Increased affinity	0.02	1.95	>1	>1	ns
H48Q	3.04	0.12	Increased affinity	0.02	1.95	>1	>1	ns
R116W	7.91	6.06	Increased affinity	0.02	1.95	>1	>1	ns
Y22C	8.56	-2.02	Reduced affinity	0.03	1.95	>1	>1	ns
H48R	3.04	0.78	Increased affinity	0.04	1.95	>1	>1	ns
R96C	5.87	1.01	Increased affinity	0.04	1.95	>1	>1	ns
K163N	3.12	-1.51	Reduced affinity	0.05	1.95	>1	>1	ns
E170K	7.68	3.33	Increased affinity	0.05	1.95	>1	>1	ns
P75L	7.12	-0.02	Reduced affinity	0.06	1.95	>1	>1	ns
R96L	5.87	-2.59	Reduced affinity	0.13	1.95	>1	>1	ns
L50R	8.56	2.27	Increased affinity	0.38	1.95	>1	>1	ns
A138V	3.31	-0.3	Reduced affinity	0.45	1.95	0.46	0.99	ns
A138E	3.31	-0.6	Reduced affinity	0.16	1.17	>1	>1	ns
P29L	8.95	0.05	Increased affinity	0.01	0.97	>1	>1	ns
G117W	8.49	8.41	Increased affinity	0.01	0.97	>1	>1	ns
A140V	4.63	0.1	Increased affinity	0.01	0.97	>1	>1	ns
G192R	7.45	2.65	Increased affinity	0.01	0.97	>1	>1	ns
A200T	4.26	3.74	Increased affinity	0.01	0.97	>1	>1	ns
E32K	7.12	3.47	Increased affinity	0.01	0.97	>1	>1	ns
R43S	6.66	1.64	Increased affinity	0.01	0.97	>1	>1	ns
R43T	6.66	1.65	Increased affinity	0.01	0.97	>1	>1	ns
A72T	5.67	5.31	Increased affinity	0.01	0.97	>1	>1	ns

L74F	4.27	6.41	Increased affinity	0.01	0.97	>1	>1	ns
S136A	6.84	-3.73	Reduced affinity	0.01	0.97	>1	>1	ns
A161D	7.79	-0.35	Reduced affinity	0.01	0.97	>1	>1	ns
G164A	2.84	-0.15	Reduced affinity	0.01	0.97	>1	>1	ns
G164S	2.84	3.47	Increased affinity	0.01	0.97	>1	>1	ns
D67A	8.41	0.29	Increased affinity	0.02	0.97	>1	>1	ns
D67E	8.41	0.03	Increased affinity	0.02	0.97	>1	>1	ns
E92A	5.69	-0.04	Reduced affinity	0.02	0.97	>1	>1	ns
E92G	5.69	-0.05	Reduced affinity	0.02	0.97	>1	>1	ns
P93S	4.51	3.19	Increased affinity	0.02	0.97	>1	>1	ns
C191W	6.92	5.09	Increased affinity	0.02	0.97	>1	>1	ns
A193G	5.87	0.18	Increased affinity	0.02	0.97	>1	>1	ns
R96H	5.87	0.76	Increased affinity	0.02	0.97	>1	>1	ns
R47G	1.51	-1.3	Reduced affinity	0.03	0.97	>1	>1	ns
R47Q	1.51	0.39	Increased affinity	0.03	0.97	>1	>1	ns
C52W	7.66	4.85	Increased affinity	0.03	0.97	>1	>1	ns
G69S	3.39	3.36	Increased affinity	0.03	0.97	>1	>1	ns
G71E	4.97	1.21	Increased affinity	0.03	0.97	>1	>1	ns
H174Y	8.05	2.32	Increased affinity	0.03	0.97	>1	>1	ns
L26S	7.62	4.26	Increased affinity	0.04	0.97	>1	>1	ns
S70N	5.57	-1.86	Reduced affinity	0.04	0.97	>1	>1	ns
L74S	4.27	4.34	Increased affinity	0.04	0.97	>1	>1	ns
R97L	3.93	-2.46	Reduced affinity	0.04	0.97	>1	>1	ns
T98P	8.4	-3.58	Reduced affinity	0.04	0.97	>1	>1	ns
R137Q	0.93	-0.47	Reduced affinity	0.04	0.97	>1	>1	ns
K163E	3.12	-3.47	Reduced affinity	0.04	0.97	>1	>1	ns

G34A	3.95	1.74	Increased affinity	0.05	0.97	>1	>1	ns
E92K	5.69	2.97	Increased affinity	0.05	0.97	>1	>1	ns
V139L	2.88	-0.24	Reduced affinity	0.05	0.97	>1	>1	ns
P199A	2.1	0.48	Increased affinity	0.05	0.97	>1	>1	ns
G117R	8.49	2.27	Increased affinity	0.07	0.97	>1	>1	ns
G34W	3.95	10.22	Increased affinity	0.08	0.97	>1	>1	ns
L50P	8.56	-0.07	Reduced affinity	0.08	0.97	>1	>1	ns
A167D	7.62	-0.3	Reduced affinity	0.08	0.97	>1	>1	ns
R137W	0.93	6.35	Increased affinity	0.14	0.97	>1	>1	ns
G30R	8.02	3.62	Increased affinity	0.07	0.78	>1	>1	ns
A138T	3.31	3.31	Increased affinity	0.04	0.65	>1	>1	ns
E40D	8.11	0.03	Increased affinity	0.01	0.49	0.55	0.99	ns
P75Q	7.12	1.62	Increased affinity	0.01	0.49	0.55	0.99	ns
S136L	6.84	-3.69	Reduced affinity	0.01	0.49	0.55	0.99	ns
L44R	7.82	3.34	Increased affinity	0.02	0.49	0.55	0.99	ns
L35P	3.04	2.2	Increased affinity	0.04	0.49	0.55	0.99	ns
W45S	9.42	-4.85	Reduced affinity	0.04	0.49	0.55	0.99	ns
C52R	7.66	-1.3	Reduced affinity	0.07	0.49	0.55	0.99	ns
P75A	7.12	-0.07	Reduced affinity	0.07	0.49	0.55	0.99	ns
G73E	8.03	0.17	Increased affinity	0.11	0.49	0.55	0.99	ns
G73R	8.03	2.79	Increased affinity	0.2	0.49	0.55	0.99	ns
H174R	8.05	-1.04	Reduced affinity	0.03	0.32	0.56	0.99	ns
P93Q	4.51	1.25	Increased affinity	0.05	0.32	0.56	0.99	ns
C52Y	7.66	2.02	Increased affinity	0.07	0.32	0.56	0.99	ns
L91V	8.54	0.01	Increased affinity	0.12	0.32	0.56	0.99	ns
G71V	4.97	1.52	Increased affinity	0.16	0.28	0.28	0.8	ns

G37R	2.3	4.68	Increased affinity	0.18	0.28	0.28	0.8	ns
R166Q	6.74	-0.61	Reduced affinity	0.02	0.24	0.31	0.8	ns
A167P	7.62	0.01	Increased affinity	0.16	0.16	0.1	0.65	ns
Y22N	8.56	-4	Reduced affinity	0.01	NA	NA	NA	ns
V31G	8.85	0.02	Increased affinity	0.01	NA	NA	NA	ns
R33G	2.08	-1.41	Reduced affinity	0.01	NA	NA	NA	ns
L35M	3.04	2.2	Increased affinity	0.01	NA	NA	NA	ns
L35Q	3.04	3.86	Increased affinity	0.01	NA	NA	NA	ns
V36A	2.08	2.53	Increased affinity	0.01	NA	NA	NA	ns
P38H	3.72	5.88	Increased affinity	0.01	NA	NA	NA	ns
P38L	3.72	2.52	Increased affinity	0.01	NA	NA	NA	ns
R39C	7.7	1.33	Increased affinity	0.01	NA	NA	NA	ns
R39P	7.7	-2.27	Reduced affinity	0.01	NA	NA	NA	ns
E40G	8.11	0.3	Increased affinity	0.01	NA	NA	NA	ns
R43G	6.66	-2	Reduced affinity	0.01	NA	NA	NA	ns
L44P	7.82	1.01	Increased affinity	0.01	NA	NA	NA	ns
W45G	9.42	-8.49	Reduced affinity	0.01	NA	NA	NA	ns
D46H	6.79	3.64	Increased affinity	0.01	NA	NA	NA	ns
R47L	1.51	-1.25	Reduced affinity	0.01	NA	NA	NA	ns
R47P	1.51	-1.26	Reduced affinity	0.01	NA	NA	NA	ns
H48D	3.04	-1.86	Reduced affinity	0.01	NA	NA	NA	ns
L49V	8.05	0.02	Increased affinity	0.01	NA	NA	NA	ns
N51T	3.72	2.58	Increased affinity	0.01	NA	NA	NA	ns
C52S	7.66	-0.03	Reduced affinity	0.01	NA	NA	NA	ns
V54A	8.75	0.02	Increased affinity	0.01	NA	NA	NA	ns
D67H	8.41	3.69	Increased affinity	0.01	NA	NA	NA	ns

G69I	3.39	-0.23	Reduced affinity	0.01	NA	NA	NA	ns
G69V	3.39	-0.24	Reduced affinity	0.01	NA	NA	NA	ns
S70G	5.57	-3.54	Reduced affinity	0.01	NA	NA	NA	ns
A72V	5.67	1.72	Increased affinity	0.01	NA	NA	NA	ns
G73W	8.03	8.93	Increased affinity	0.01	NA	NA	NA	ns
L74V	4.27	0.76	Increased affinity	0.01	NA	NA	NA	ns
G76S	9.67	3.65	Increased affinity	0.01	NA	NA	NA	ns
L94P	1.98	-0.64	Reduced affinity	0.01	NA	NA	NA	ns
L95V	7.12	0	Reduced affinity	0.01	NA	NA	NA	ns
R96P	5.87	-2.6	Reduced affinity	0.01	NA	NA	NA	ns
R97C	3.93	1.13	Increased affinity	0.01	NA	NA	NA	ns
T98A	8.4	-3.61	Reduced affinity	0.01	NA	NA	NA	ns
T98I	8.4	-3.57	Reduced affinity	0.01	NA	NA	NA	ns
R116G	7.91	-2.46	Reduced affinity	0.01	NA	NA	NA	ns
R137G	0.93	-2.16	Reduced affinity	0.01	NA	NA	NA	ns
R137L	0.93	-2.12	Reduced affinity	0.01	NA	NA	NA	ns
A138P	3.31	-0.3	Reduced affinity	0.01	NA	NA	NA	ns
A138S	3.31	3.3	Increased affinity	0.01	NA	NA	NA	ns
V139M	2.88	-0.25	Reduced affinity	0.01	NA	NA	NA	ns
A140S	4.63	3.69	Increased affinity	0.01	NA	NA	NA	ns
A141P	6.12	-0.12	Reduced affinity	0.01	NA	NA	NA	ns
A141T	6.12	3.48	Increased affinity	0.01	NA	NA	NA	ns
L142F	7.41	5.54	Increased affinity	0.01	NA	NA	NA	ns
L142S	7.41	3.48	Increased affinity	0.01	NA	NA	NA	ns
W148C	8.9	-4.84	Reduced affinity	0.01	NA	NA	NA	ns
W148L	8.9	-8.46	Reduced affinity	0.01	NA	NA	NA	ns

W148R	8.9	-6.12	Reduced affinity	0.01	NA	NA	NA	ns
A161P	7.79	-0.01	Reduced affinity	0.01	NA	NA	NA	ns
A161V	7.79	-0.01	Reduced affinity	0.01	NA	NA	NA	ns
K163R	3.12	-0.83	Reduced affinity	0.01	NA	NA	NA	ns
G164C	2.84	3.47	Increased affinity	0.01	NA	NA	NA	ns
R166W	6.74	6.2	Increased affinity	0.01	NA	NA	NA	ns
E170G	7.68	0.32	Increased affinity	0.01	NA	NA	NA	ns
H174D	8.05	-3.66	Reduced affinity	0.01	NA	NA	NA	ns
H174N	8.05	-1.7	Reduced affinity	0.01	NA	NA	NA	ns
T190I	8.62	-3.46	Reduced affinity	0.01	NA	NA	NA	ns
T190S	8.62	0.14	Increased affinity	0.01	NA	NA	NA	ns
C191R	6.92	-1.04	Reduced affinity	0.01	NA	NA	NA	ns
C191Y	6.92	2.26	Increased affinity	0.01	NA	NA	NA	ns
G192V	7.45	0.32	Increased affinity	0.01	NA	NA	NA	ns
A193E	5.87	-0.09	Reduced affinity	0.01	NA	NA	NA	ns
N194D	9.09	-1.9	Reduced affinity	0.01	NA	NA	NA	ns
N194S	9.09	2	Increased affinity	0.01	NA	NA	NA	ns
R197H	7.26	1.09	Increased affinity	0.01	NA	NA	NA	ns
P198R	3.97	2.48	Increased affinity	0.01	NA	NA	NA	ns
P199H	2.1	3.88	Increased affinity	0.01	NA	NA	NA	ns
P199R	2.1	2.84	Increased affinity	0.01	NA	NA	NA	ns
T201P	5.69	-3.04	Reduced affinity	0.01	NA	NA	NA	ns
V202G	8.74	-0.02	Reduced affinity	0.01	NA	NA	NA	ns
V202L	8.74	0.06	Increased affinity	0.01	NA	NA	NA	ns
G30C	8.02	4.9	Increased affinity	0.02	NA	NA	NA	ns
V36G	2.08	2.51	Increased affinity	0.02	NA	NA	NA	ns

E40A	8.11	0.31	Increased affinity	0.02	NA	NA	NA	ns
L49R	8.05	2.34	Increased affinity	0.02	NA	NA	NA	ns
A72S	5.67	5.29	Increased affinity	0.02	NA	NA	NA	ns
G76V	9.67	0.08	Increased affinity	0.02	NA	NA	NA	ns
G117E	8.49	-0.36	Reduced affinity	0.02	NA	NA	NA	ns
R118H	5.46	0.74	Increased affinity	0.02	NA	NA	NA	ns
A161G	7.79	-0.06	Reduced affinity	0.02	NA	NA	NA	ns
L196S	6.03	3.75	Increased affinity	0.02	NA	NA	NA	ns
S70I	5.57	-3.5	Reduced affinity	0.03	NA	NA	NA	ns
G164R	2.84	2.19	Increased affinity	0.03	NA	NA	NA	ns
W45R	9.42	-6.11	Reduced affinity	0.04	NA	NA	NA	ns
D67N	8.41	1.98	Increased affinity	0.04	NA	NA	NA	ns
N194T	9.09	2.01	Increased affinity	0.04	NA	NA	NA	ns
R197C	7.26	1.34	Increased affinity	0.04	NA	NA	NA	ns
L91R	8.54	2.35	Increased affinity	0.05	NA	NA	NA	ns
L145S	8.51	3.6	Increased affinity	0.05	NA	NA	NA	ns
G164V	2.84	-0.14	Reduced affinity	0.05	NA	NA	NA	ns
N51K	3.72	1.95	Increased affinity	0.06	NA	NA	NA	ns
L35R	3.04	4.53	Increased affinity	0.07	NA	NA	NA	ns
G37V	2.3	2.35	Increased affinity	0.13	NA	NA	NA	ns
L145F	8.51	5.66	Increased affinity	0.13	NA	NA	NA	ns

Table 5.B.1: Mutations close to nucleic acid in GidB

Two hundred and twenty six single amino acid variation (SAV) mutations lying within 10Å of the nucleic acid (NA) and their corresponding NA affinity changes ($\Delta\Delta$G) measured by mCSM-NA. The estimated effect are categorised as Destabilising ($\Delta\Delta$G<0) and Stabilising ($\Delta\Delta$G>0). The genomic measures of minor allele frequency (MAF), Odds Ratio (OR), OR related P-values, and FDR adjusted P-values are shown. Statistical significance is indicated as: *P < 0.05, **P < 0.01, ***P < 0.001, ****P < 0.0001, ns: >0.05. The table is arranged by OR to to show mutation with the highest OR at the top for mutations close to the nucleic acid. Columns with NA indicate insufficient data to calculate OR. Abbreviations used: $\Delta\Delta$G: change in Gibbs free energy in Kcal/mol, FDR: false discovery rate, ns: not significant.

5.C Average stability comparisons for lineages

Lineage (L) comparisons	**Samples (n)**	**Adjusted P-values**	***Adjusted P-values* Significance**
L1 vs L2	L1 (747), L2 (9465)	<2.2e-16	****
L1 vs L3	L1 (747), L3 (668)	3.3e-016	****
L1 vs L4	L1 (747), L4 (7372)	<2.2e-16	****
L2 vs L3	L2 (9465), L3 (668)	<2.2e-16	****
L2 vs L4	L2 (9465), L4 (7372)	<2.2e-16	****
L3 vs L4	L3 (668), L4 (7372)	<2.2e-16	****
	Within Lineage comparisons		
L1: R vs S	R (n=19), S (n=728)	0.05	ns
L2: R vs S	R (n=218), S (n=9247)	<2.2e-16	****
L3: R vs S	R (n=33), S (n=635)	0	****

Table 5.C.1: Lineage comparisons for GidB mutations
Kolmogorov-Smirnoff (KS) test reporting the statistical differences in distributions between *M. tuberculosis* lineages when assessed based on average stability changes ($\Delta\Delta$G) measured by mCSM-DUET, FoldX, DeepDDG, and Dynamut2. Lineage comparisons were performed for samples containing mutations associated with sensitivity (R: Resistant, S: Sensitive). These comparisons were performed for R and S samples between and in lineages. Statistical significance thresholds used are *P<0.05, **P<0.01, ***P<0.001, ****P<0.0001, ns >0.05. Abbreviations used: $\Delta\Delta$G: change in Gibbs free energy in Kcal/mol, Adj. P-values: Bonferroni adjusted P-values, n: number of samples, ns: not significant.

Chapter 6

KatG-isoniazid results

6.1 Background

6.1.1 Mechanism of action of isoniazid

Isoniazid, also known as isonicotinic acid hydrazide (INH) is an antibiotic used in the treatment of active and latent TB. For active TB, it is used in combination with other drugs like rifampicin, pyrazinamide, streptomycin or ethambutol. INH is a pro-drug that is activated by the enzyme catalase peroxidase encoded by the *katG* gene.[1] The primary mechanism of action of INH is inhibition of mycolic acid synthesis by binding to the NADH-dependent enoyl-acyl carrier reductase protein, encoded by the *inhA* gene.[2,3] Activation of INH by *katG* produces a range of radicals including nitric oxide capable of attacking multiple targets in *M. tuberculosis*.[4] KatG catalyses the formation of isonicotinic acyl radical which together with NADH forms the nicotinoyl-NAD adduct. This complex then binds to inhA (enoyl-acyl carrier protein reductase protein) to block the natural enoyl-AcpM substrate leading to inhibition of the biosynthesis of mycolic acids - an essential component of mycobacterial cell walls **(Figure 1)**.

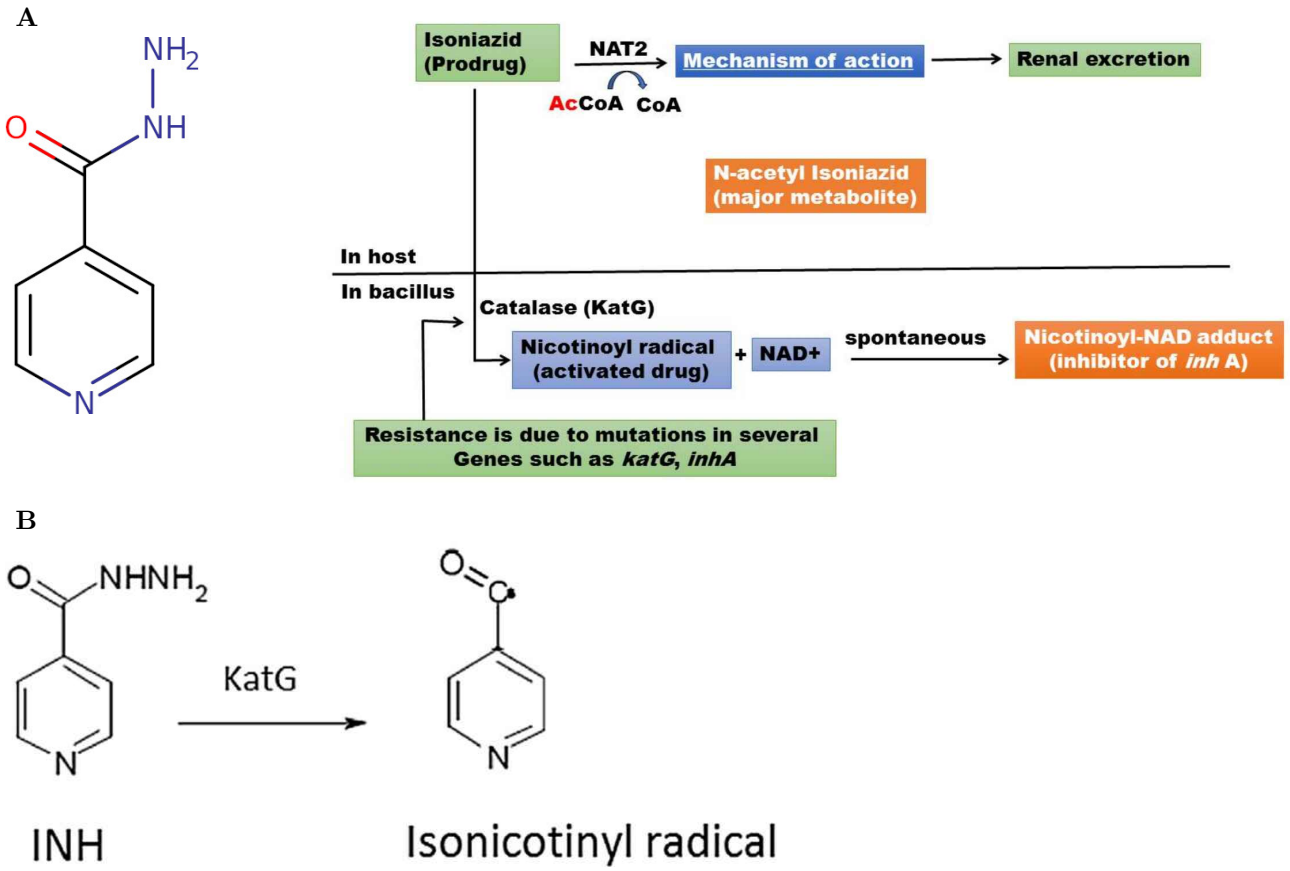


Figure 1: Chemical structure and mechanism of action and resistance for isoniazid
A) The chemical structure of isoniazid (INH) is shown at the top left and is sourced from DrugBank (ID:DB00951). An overview of the mechanism of action and resistance for isoniazid (INH) is shown, with mutations in *katG* and *inhA* playing an important role in INH resistance. Figure adapted from Sheikh *et. al.*,[5] **B)**(enoyl-acyl carrier protein reductase protein). Figure adapted from Unissa *et. al.*.[6]

6.1.2 Isoniazid resistance in *M. tuberculosis*

TB treatment with INH has been compromised due to drug resistance, with mutations in the NADH-dependent enoyl-acyl-carrier-protein reductase (*inhA*) and Catalase-peroxidase-peroxynitritase T (*katG*) genes, with *katG* being the major mechanism of INH resistance.[6–8] The most prevalent mutation identified is S315T in *katG* which results in the formation of an INH derivative unable to form the INH-NAD adduct required for its antibacterial activity.[8–12] This mutation is commonly observed in MDR (resistant to both isoniazid and rifampicin) strains of *M. tuberculosis*.[13] Additional mutations at this site (S315I, S315R, S315N, S315G) in clinical strains have been shown to interfere with flexibility and stability of the INH binding site causing rigidity, leading to reduced- or in- activity of the enzyme.[8] Among other mutations commonly associated with INH resistance are R104L, H108Q, N138S, D142A, L148R, H270Q, T275P, W321G, D381G, L587M, A350T, R463L, R463G,[1,14,15] as well as high confidence mutations associated with resistance: A139P, S140N, S140R, D142A, G279D, G285D, G316D, S457I, and G593D.[16]

6.1.3 Active site description and INH resistance

An experimentally determined atomic structure of KatG in *M. tuberculosis* is available as PDB entry 1SJ2[17] as a homo-dimeric enzyme. The KatG protein is a homo-dimeric bifunctional heme-dependent enzyme. It exhibits catalytic and broad spectrum peroxidatic activity, comparable to monofunctional catalases. As such, INH binding is mediated by the co-factor heme bound to the KatG protein.[17,18]

Interactions in KatG

Molecular interactions with residues between KatG, INH, and the heme co-factor were identified using LigPlus, PLIP and Arpeggio tools, resulting in a total of thirty-four interacting residues:

- Ten residues at sites 104, 107, 108, 136, 137, 228, 229, 230, 232, 315 were identified as interacting with INH.

- Twenty-nine residues at sites 94, 100, 101, 103, 104, 107, 230, 231, 232, 248, 252 265, 266, 269, 270, 272, 273, 274, 275, 276, 314, 315, 317, 321, 378, 380, 381, 408, 412 were identified as interacting with co-factor heme.

An overview of the KatG structural complex with all interactions identified is shown in **Figure 2**.

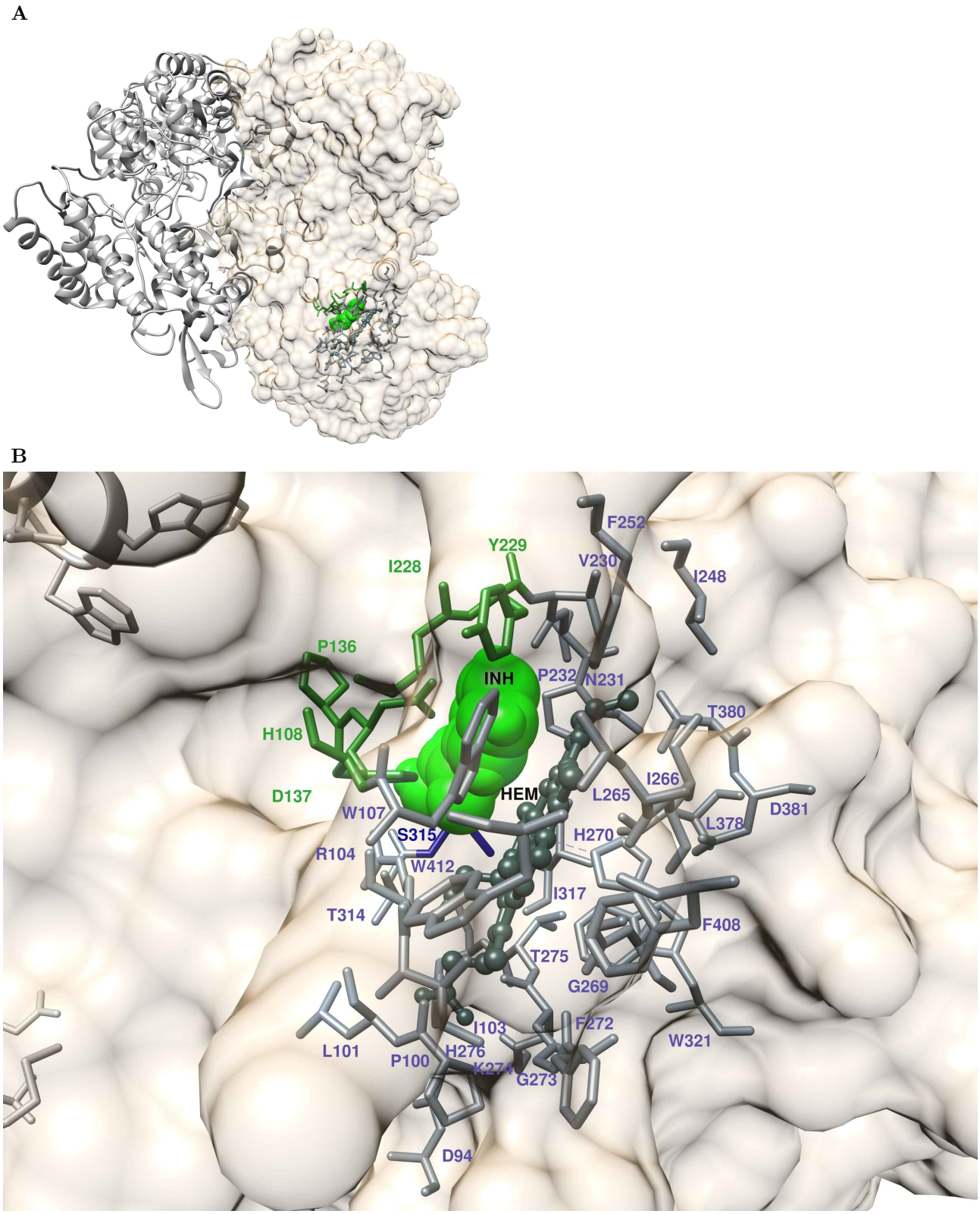


Figure 2: Description of *M. tuberculosis* KatG protein complex with INH and co-factor heme Overall description of KatG-INH complex. INH appears as green spheres while co-factor heme is shown in dark slate grey as ball-and-stick. **A)** homo-dimer KatG in complex with INH with chain A indicated as surface representation in tan colour, while chain B is shown as grey ribbons, **B)** Close-up view of all interacting residues coloured green for INH and dark slate grey for co-factor heme, and labelled accordingly. The key active site residue S315 is highlighted in blue. Abbreviation used: INH: isoniazid.

6.2 Structural and genomic insights into isoniazid resistance

6.2.1 Mutational landscape of KatG

Multiple SAVs are spread along KatG including the active site and beyond

A total of 817 SAVs were found in the protein coding region of *katG* (Rv1908c: 2153889-2156111). The mutational landscape of KatG appears distributed across the protein **(Figure 3)**, with mutations present in 460 unique positions with a maximum of nine SAVs at a single site **(Figure 4)**.

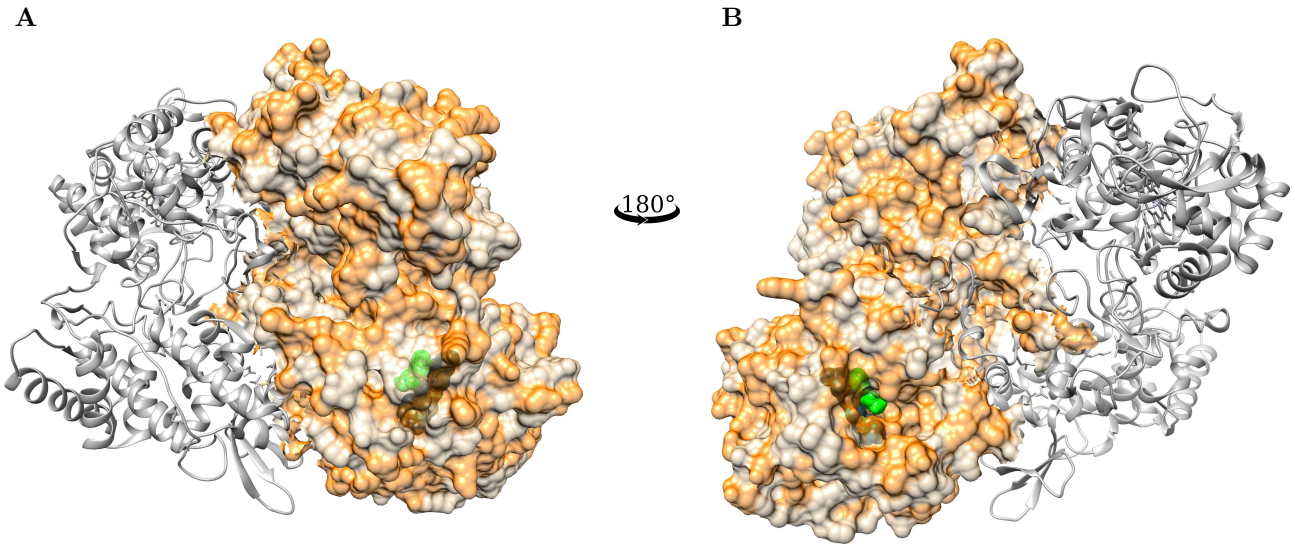


Figure 3: Mutational landscape of *M. tuberculosis* KatG
An overview of all mutational sites on *M. tuberculosis* KatG appearing as surface representation in tan colour with chain B is shown as grey ribbons. Sites associated with SAVs are coloured orange. Panels **A)** and **B)** are opposing representations (rotated 180°) of INH, with INH shown as green spheres in the binding pocket while co-factor heme appears as dark slate grey sphere. The figure is generated using UCSF Chimera version 1.14. Abbreviations used: SAV: single amino acid variation, INH: isoniazid.

Half (50%, n=5) of residues interacting with INH **(Figure 2)** had SAVs, as did 72% (n=21) of residues interacting with heme **(Figure 2)**. Mapping mutations by positions in KatG highlighted **(Figure 4)** the following:

Sites with INH interactions associated with a maximum of 5 SAVs (sites marked in dark green)

- Single mutations: P136, I228
- Budding resistant hotspots: R104, shared interaction with heme.
- Hotspots with four mutations: P232, shared interaction with heme.
- Hotspots with five mutations: S315, shared interaction with heme.

Sites with heme interactions associated with a maximum of 5 SAVs (sites marked in dark slate grey)

- Single mutations: D94N, P100T, N231K, I248T, F252L, H276Q
- Budding resistant hotspots: L101, R104, G269, G273, T275, and L378 where R104 shared interaction with INH.
- Hotspots with three mutations: I103, F272, T314, and I317 showed triple mutations.
- Hotspots with four mutations: P232 and F408 showed 4 mutations each where P232 shared interaction with INH.
- Hotspots with five mutations: S315 and T380 showed 5 mutations each where key residue S315 shared interaction with INH.

Sites in KatG which did not involve residues interacting with INH or heme, and were located more than 10Å away from INH had many mutations at a single site to a maximum of 9. These were H116 (9 SAV mutations) and G124 (8 SAV mutations). The majority (61%, n=504) of the mutational effects resulted in electrostatic changes **(Figure 4)**.

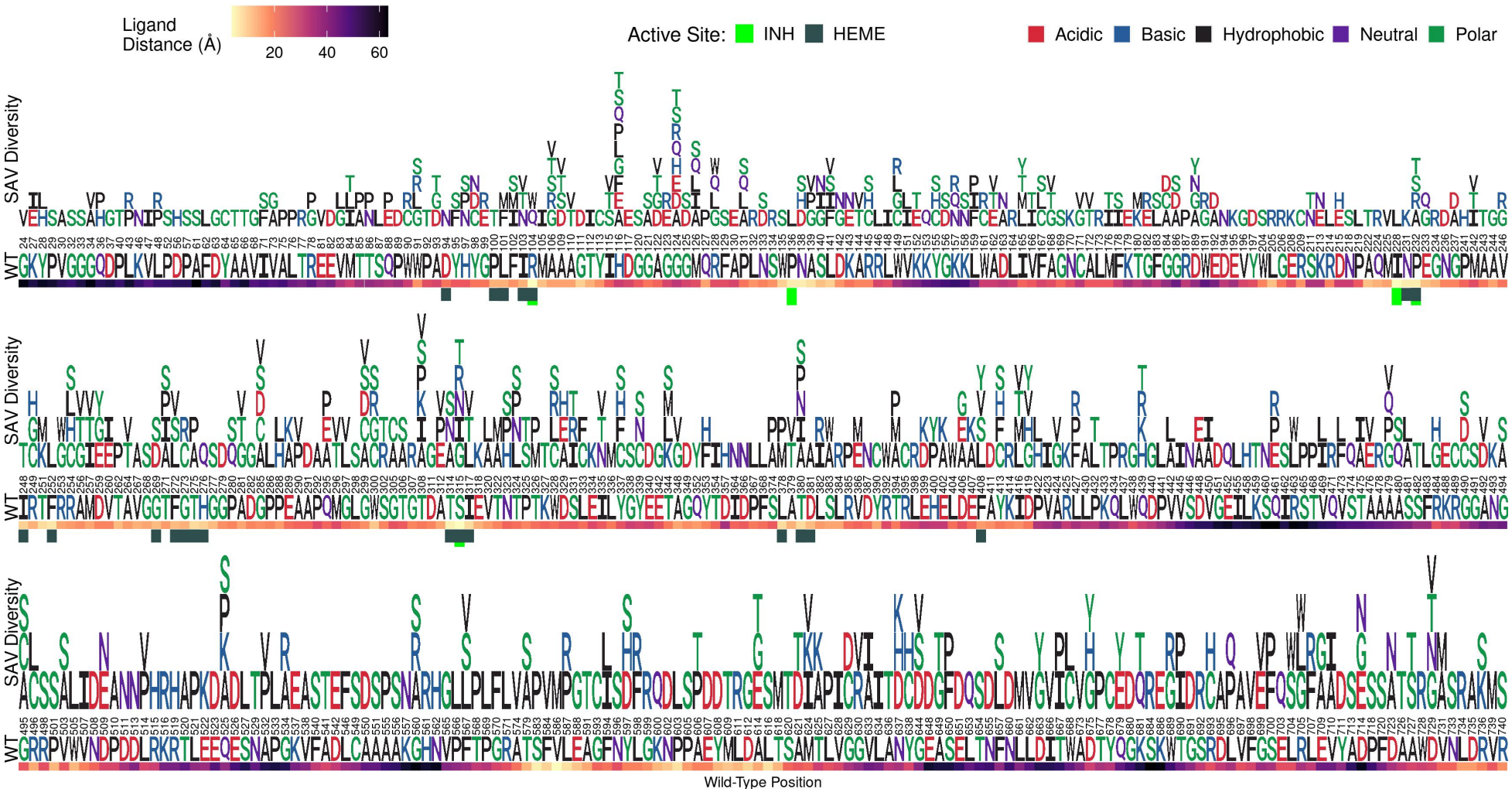


Figure 4: Sites associated with SAVs in *M. tuberculosis* KatG

Logo plot showing 460 unique sites/positions associated with 817 SAVs in *M. tuberculosis* KatG. The horizontal axis shows the wild-type positions associated with SAVs in KatG and the vertical axis shows all the mutant residues observed in our data highlighting SAV diversity at a given site. Residues are coloured according to the amino acid (aa) property where acidic aa appear in red, basic aa appear in blue, hydrophobic aa in black, neutral aa in purple, and polar aa in darkgreen. The structural positions associated with SAVs in KatG are indicated on the horizontal axis. The wild-type (WT) residues also coloured according to aa property appear under the respective position markings. The heat bar underneath the WT residues indicate the distance of that position from INH according to the magma colour gradient where light yellow indicates sites closer to INH (ligand distance in Angstroms). The positions are further annotated to reflect active site residues involved in interactions with INH (green), and co-factor heme in dark slate grey. The figure is generated using R statistical software version 4.0.2, ggplot2 package. Abbreviations used: SAV: single amino acid variation, INH: isoniazid.

6.2.2 Mutational outcome from protomer stability changes and evolutionary conservation

Most mutational consequences are destabilising for protomer stability and have deleterious impact on protein function

Most mutations have a destabilising effect on the overall protomer stability when measured by the different computational tools **(Figure 5A-4D)**, with DeepDDG estimating 90% (n=740) as destabilising, followed by ∼84% of mutations estimated by Dynamut2 (n=687) and mCSM (n=685) as destabilising, and then by FoldX predicting ∼81% (n=658) mutations as destabilising. From an evolutionary conservation perspective, most mutations were predicted to result in a deleterious impact (effect) on protein function indicated by PROVEAN and SNAP2 scores. PROVEAN estimated around 71% (n=580) **Figure 5E)** and SNAP2 estimated nearly 68% (n=553) SAVs resulting in deleterious effects **(Figure 5F)**.

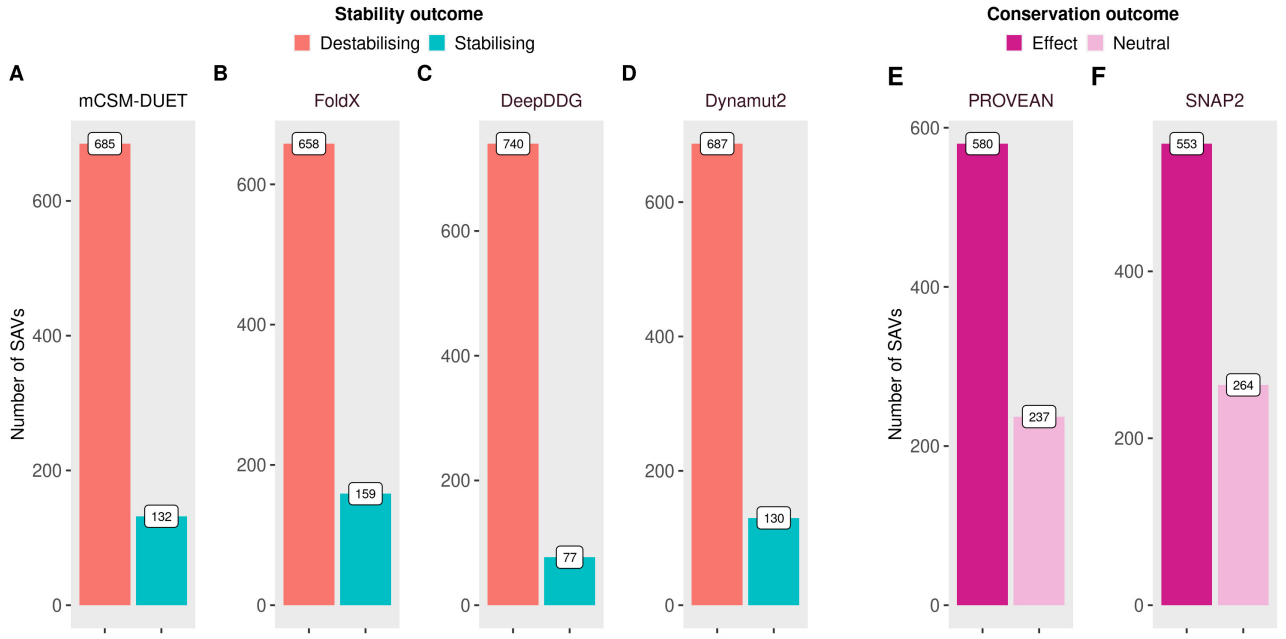


Figure 5: Protein stability outcome of SAVs in *M. tuberculosis* KatG
Mutational impact on overall protein stability and evolutionary conservation changes for 817 SAVs, **A-D)** Barplots showing number of SAVs categorised as destabilising (red) or stabilising (blue) according to protein stability changes (ΔΔG Kcal/mol) measured by four computational tools: mCSM-DUET, FoldX, DeepDDG, and Dynamut2, **E-F)** Number of SAVs categorised as Effect/Deleterious (magenta) or Neutral (pink) according to evolutionary conservation changes estimated by computational tools: PROVEAN, and SNAP2. The figure is generated using R statistical software version 4.0.2, ggplot2 package. Abbreviations used: ΔΔG: change in Gibbs free energy, SAV: single amino acid variation.

Evolutionary and structure-based predictors provide different insights into understanding mutational impact. Mutational impact in this context is considered to be its effect on protein stability, drug binding affinity, other binding affinities such as PPI or nucleic acid, and functional effects arising from protein sequence variations. The first three mutational consequences are assessed by structure

based predictors relying on the 3D structure of a protein, while the last is assessed by sequence based predictors relying mainly on evolutionary conservation trends across many proteins using multiple sequence alignments. The sequence based predictors are aimed at predicting pathogenicity or change of molecular function, structure based tools rely on estimating variant effects in relation to structure damage, corresponding to stability changes, as protein stability is considered the basic characteristic affecting function, activity, and regulation. Predictors such as ConSurf are able to use both structural and sequence information to identify important functional regions conserved in proteins. A variant classified as 'deleterious' to protein conservation may display gain-of-function in the presence of a drug through optimised protein stability. Thus, different methodological strategies benefit from complementary information when assessing specific proteins.

6.2.3 Mutational consequences on affinity changes and prominent mutational effects

Mutations decrease binding affinity for INH and the dimer interface

Around 9% (n=74) of SAVs inducing changes in INH binding affinity were within 10Å of INH. These mutations occurred at 36 distinct sites, with most sites (n=18) having single mutations. Over 91% of mutations (n=68) were predicted to result in a destabilising effect on INH binding affinity measured by mCSM-lig, and all (n=74) mutations were destabilising according to mmCSM-lig **(Figure 6A** top panel, Appendix Table 6.A.1**)**. The average binding affinity effect of the 36 mutational sites showed mildly destabilising mutational consequences **(Figure 6A** bottom panel **)**. Analysing the sites close to the dimer interface highlighted about 32% (n=260) mutations, corresponding to 144 distinct sites, to be within 10Å of the dimer interface as measured by mCSM-PPI2, where 80% (n=210) of mutations resulted in destabilising effects **(Figure 6B** top panel, Appendix Table 6.B.1**)**. Sites around the PPI showed mixed stability effects with stabilising mutations appearing closer to the dimer interface on visual inspection **(Figure 6B** bottom panel**)**.

Of the total 460 unique sites in KatG displaying SAVs, approximately 52% (n=239) of sites showed single mutations, followed by ∼30% (n=136) sites as budding resistant hotspots **(Figure 6C** top panel**)**. The most prominent effects on INH binding were from reduced (destabilising) binding affinity to INH contributed by mutations from 27 surrounding sites **(Figure 6C**, yellow text boxes and bottom panel**)**. Sites close to the dimer interface were mostly affected by destabilising mutations from 43 surrounding sites, while 12 sites contributing to stabilising mutational impacts **(Figure 6C**, pink text boxes, and bottom panel**)**. All other sites were largely (n=332) affected by destabilising mutations **(Figure 6C**, blue and red text boxes, and bottom panel**)** impacting protomer stability.

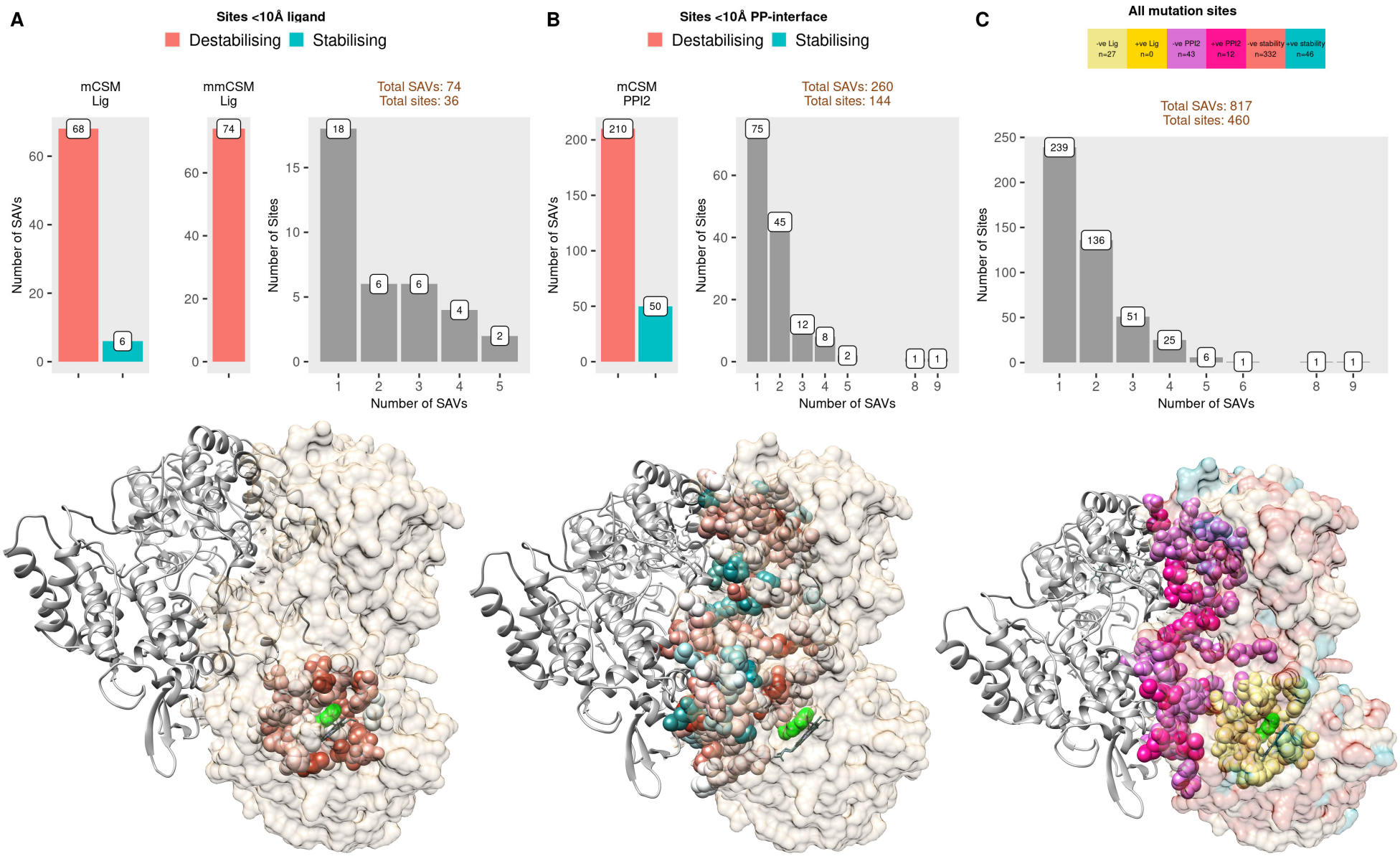

A
Sites <10Å ligand
Destabilising
Stabilising
mCSM Lig
mmCSM Lig
Total SAVs: 74
Total sites: 36
Number of SAVs
Number of Sites
B
Sites <10Å PP-interface
mCSM PPI2
Total SAVs: 260
Total sites: 144
C
All mutation sites
-ve Lig n=27
+ve Lig n=0
-ve PPI2 n=43
+ve PPI2 n=12
-ve stability n=332
+ve stability n=46
Total SAVs: 817
Total sites: 460

Figure 6: Mutational impact on INH binding affinity, protein-protein interaction on KatG, and sites with the most prominent mutational effects within *M. tuberculosis* KatG
The top panel displays barplots showing the mutational outcome of affinity changes and their corresponding site frequency, while the bottom panel shows the corresponding mutational impact mapped onto the KatG (chain A) appearing in tan colour, while chain B is shown as grey ribbons. INH is shown as green spheres in the binding site, while heme appears as dark slate grey sticks. **A)** Mutational impact on INH binding affinity (log fold change) from mCSM-lig and mmCSM-lig with 74 mutations occurring at 36 sites within 10Å of INH, **B)** Mutational impact on protein-protein (PP) binding affinity ($\Delta\Delta$G) for 260 mutations at 144 sites within 10Å of the PPI. For both parts A) and B), red denotes destabilising mutational sites while blue denotes stabilising mutational sites, and the colour intensity reflects the extent of the effect ranging from -1 (most destabilising) to +1 (most stabilising), **C)** Most prominent mutational effect for all 817 SAVs present at 460 sites, prioritised in order of increasing effect size: mCSM/mmCSM-lig, mCSM-PPI2, followed by overall stability changes where brighter colours indicate stabilising effects. Sites marked in yellow indicate changes due to ligand (INH) binding affinity with light yellow denoting destabilising changes, pink areas indicate changes due to PPI affinity with bright pink highlighting stabilising and light pink areas indicating destabilising mutational effects. All other sites are coloured by protomer stability changes with blue showing stabilising and red indicating destabilising effects. The corresponding number of mutation sites contributing to these changes are indicated in the text box at the top, and coloured accordingly. The barplot figures are generated using R statistical software version 4.0.4, ggplot2 package. The structure figures are generated using Chimera version 1.14. Abbreviations used: Å: Angstroms, $\Delta\Delta$G: change in Gibbs free energy in kcal/mol, SAV: single amino acid variation, INH: isoniazid.

6.2.4 Mutational association with INH resistance and flexibility

Most mutations lie in conserved areas and are associated with mild-to-moderate flexibility

Mutational association with resistance according to aggregate DST data showed approximately equal number of resistant (45%, n=369) and sensitive mutations (54%, n=448). Mutational sites were mapped onto KatG to highlight the location of sites with exclusively resistant (red) and sensitive (blue) mutations while sites displaying both resistant and sensitive mutations were coloured purple. There were 140 sites with resistant mutations, 119 sites with both resistant and sensitive mutations, while 201 sites with sensitive mutations **(Figure 7A)**.

ConSurf scores are calculated for each site on the protein, and range from 1 (rapidly evolving, variable sites) to 9 (slowly evolving, conserved sites). While there were some resistant mutations close to INH and the PPI, such mutations were not restricted in these areas, with resistant mutations occurring away from INH, heme co-factor, and the dimer interface **(Figure 7B** left panel**)**. Most mutations (n=270) occurred in the highly conserved regions of katG ConSurf score 9)**(Figure 7B** right panel**)**.

Further, the local flexibility in KatG in relation to INH resistance was also analysed, with thickness of the ribbon/tube (thinthick) corresponding to the extent of flexibility. Normal mode analysis of KatG highlighted that regions associated with SAVs in KatG were located in regions of moderate flexibility **(Figure 8** left panel**)** though the key active site residue S315 was in a region of low flexibility **(Figure 8** right panel**)**. A budding resistant hotspot site E233 was associated with high flexibility **(Figure 8** right panel**)**, and the heme interacting residue site N231 exhibited mild-to-moderate flexibility on

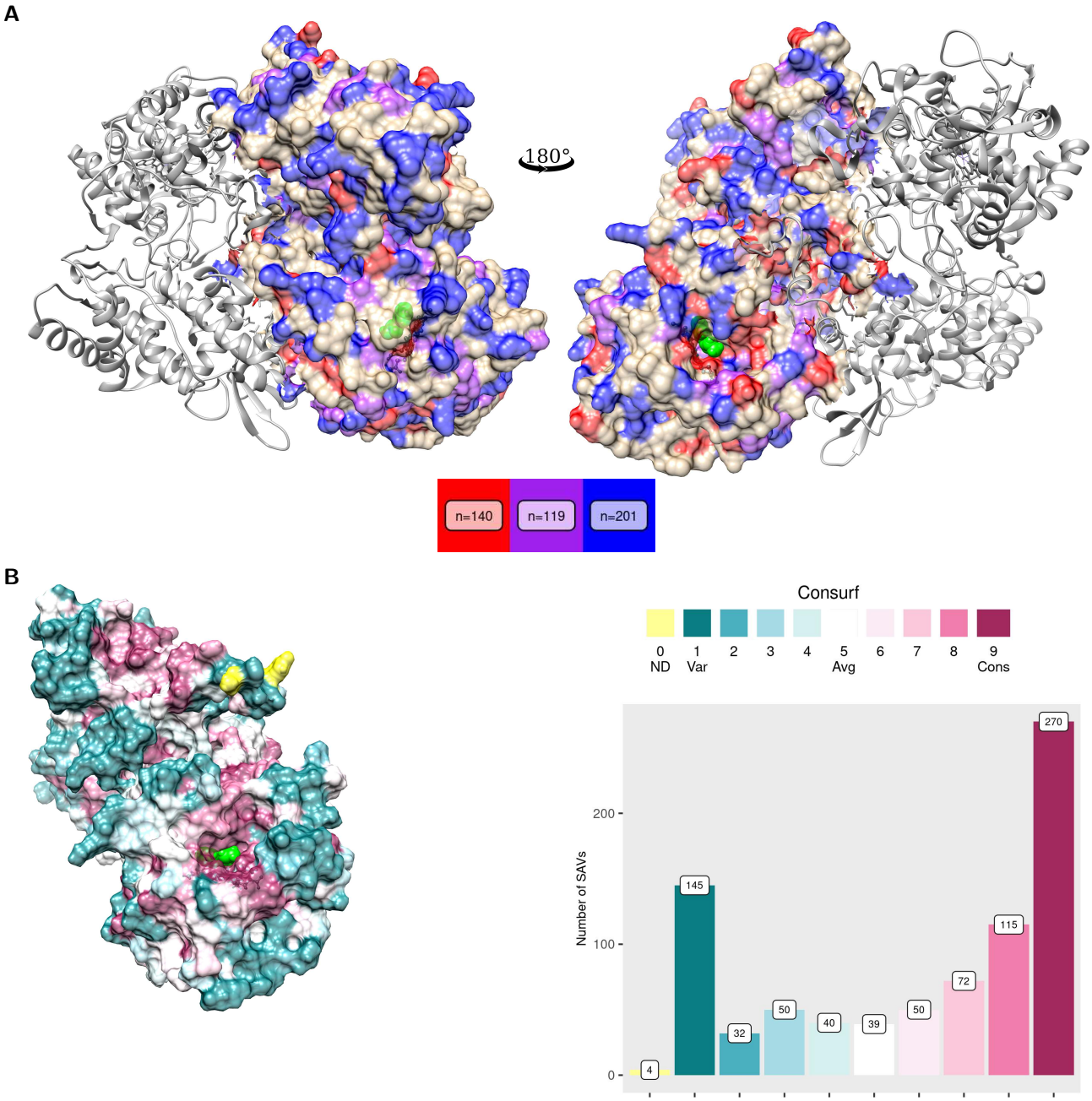


Figure 7: Mutational association with isoniazid resistance and evolutionary conservation in *M. tuberculosis* KatG

Mutational landscape of *M. tuberculosis* KatG according to different measures where **A)** All sites associated with SAVs on *M. tuberculosis* KatG with INH shown as green spheres, **A)** The left panel shows all mutational sites associated with resistant (red, n=140 sites), sensitive (blue, n=201 sites), while common sites with both resistant and sensitive mutations appear in purple (n=119). The corresponding right panel depicts the structure rotated by 180°, **B)** Left panel shows KatG coloured according to ConSurf scores where maroon indicates conserved sites and teal indicates variable sites. Yellow areas reflect sites with uncertainty due to insufficient data for ConSurf score calculation. The barplot on the right panel shows the the number of mutations associated with ConSurf values that range from 1 (variable) in teal to 9 (conserved) in maroon, where 0 denotes insufficient data/not defined (ND). The barplot figures are generated using R statistical software version 4.0.4, ggplot2 package. All structure figures were generated using UCSF Chimera version 1.14. Abbreviations used: SAV: single amino acid variation, INH: isoniazid.

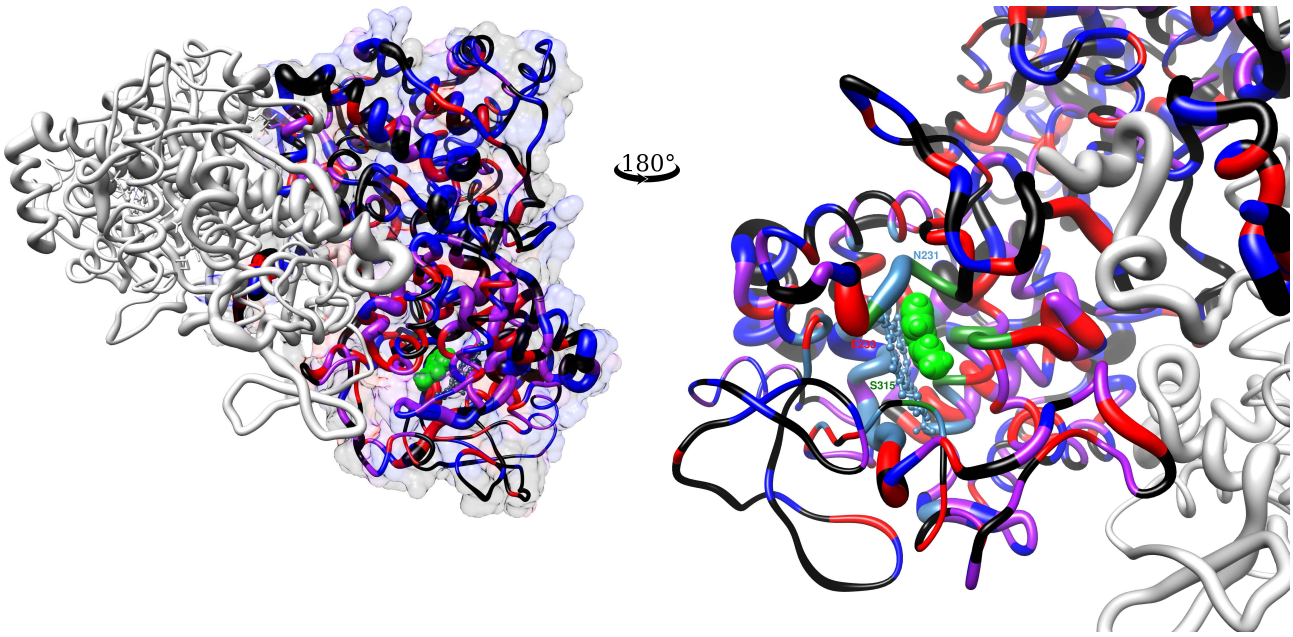


Figure 8: Mutational association with isoniazid resistance and local protein flexibility of *M. tuberculosis* KatG
Mutational landscape of *M. tuberculosis* KatG according to flexibility in KatG according to normal mode analysis (NMA), measuring atomic deformation according to protein dynamics to denote flexibility associated at sites in KatG. The magnitude of flexibility is represented from thin (low flexibility) to thick (high flexibility) tubes. Left panel: The tubes are further coloured to show mutational association with INH resistance, red: resistant sites, blue: sensitive sites, purple: shared sites, black: sites with no SAVs. Right panel: further coloured to indicate INH and heme interacting residues in green and steel blue respectively. Wild-type residues marked using the standard one-letter amino acid code denote key active site residue S315 associated with low flexibility, heme interacting residue N231 associated with marginal flexibility, and residue E223 as a resistant site associated with high flexibility. All structure figures were generated using UCSF Chimera version 1.14. Abbreviations used: SAV: single amino acid variation, INH: isoniazid.

visual inspection. Regions with the highest flexibility (black thick tubes) did not present with SAVs **(Figure 8** left panel**)**.

6.2.5 Relating mutational frequency and biophysical and evolutionary conservational changes

Correlation analysis was performed to understand the relationship between frequently occurring mutations as assessed by MAF and their association between stability (mCSM-DUET, FoldX, DeepDDG, Dynamut2), conservation (ConSurf, SNAP2, PROVEAN) and affinity changes (mCSM-lig/mmCSM-lig, and mCSM-PPI2), distance to ligand (Lig-Dist), and protein-protein interface (PPI-Dist). A combined analysis with all mutations, as well as separately for resistant (R) and sensitive (S) mutations was undertaken **(Figures 9** and **10)**. Analyses focused on determining the strength of association without regard for the direction of the association due to dissimilarity of threshold criteria used by the various estimators.

Frequently occurring sensitive mutations were weakly related to protomer stability changes and distance from INH

Frequently occurring mutations were not related to protomer stability changes (ρ_{R+S}<0.1) **(Figure 9)**, though weak association was observed for frequently occurring sensitive mutations according to DeepDDG (ρ_S=0.21, P<0.001), and FoldX (ρ_S=-0.15, P<0.01). Frequently occurring mutations were overall weakly associated with distance from the drug (ρ_{R+S}=-0.17, P<0.001), but moderate association was observed for the resistant mutation group (ρ_R∼0.3, P<0.001). Mutational frequency was not associated with distance to the dimer interface (ρ_{R+S}<0.1 and $\rho_{R/S<}$0.1, P>0.05). The different computational tools showed good consensus (moderate to strong associations) amongst their predicted estimates, both overall (0.4<ρ_{R+S}<0.8, P<0.001), as well as for resistant and sensitive mutation groups individually (0.3≤$\rho_{R/S}$<0.8, P<0.001). As expected, mCSM-DUET and Dynamut2 were strongly correlated as these tools share common methodology (ρ=0.83, P<0.001) **(Figure 9)**.

Frequently occurring sensitive mutations were weakly associated with evolutionary conservation changes

Overall, there was no association with mutational frequency and rate of evolution according to ConSurf (ρ_{R+S}<0.1, P>0.05), with only weak associations with changes in protein function according to PROVEAN (ρ_{R+S}=0.12, P<0.001) but not SNAP2 (ρ_{R+S}∼0.1, P<0.05). In particular, frequently occurring sensitive mutations were driving the moderate association according to SNAP2 (ρ_S∼0.3, P<0.001) and PROVEAN (ρ_S∼0.3, P<0.001). There was good agreement (moderate to strong association) between estimates across the three conservation predictors both overall (ρ_{R+S}>0.6, P<0.001) and in the mutation groups ($\rho_{R/S}$>0.5, P<0.001) **(Figure 10** left panel**)**.

Frequently occurring resistant mutations were weakly related to INH affinity changes

Frequently occurring mutations were weakly related to INH affinity changes (mCSM-lig: ρ_{R+S}=-0.11, P<0.01, mmCSM-lig: ρ_{R+S}=-0.18, P<0.001) with only resistant mutations driving this weak association for mCSM-lig ρ_R=-0.18, and moderate association for mmCSM-lig ρ_R∼0.3, P<0.001 **(Figure 10** right panel**)**. There was no association between mutational frequency and changes in dimer interface affinity (ρ_{R+S}<0.1, $\rho_{R/S}$<0.1, P>0.05) **Figure 10** right panel**)**. As expected, mCSM- and mmCSM-lig values were strongly correlated overall and in the mutation groups (ρ_{R+S}>0.80, $\rho_{R/S}$>0.80, P<0.001).

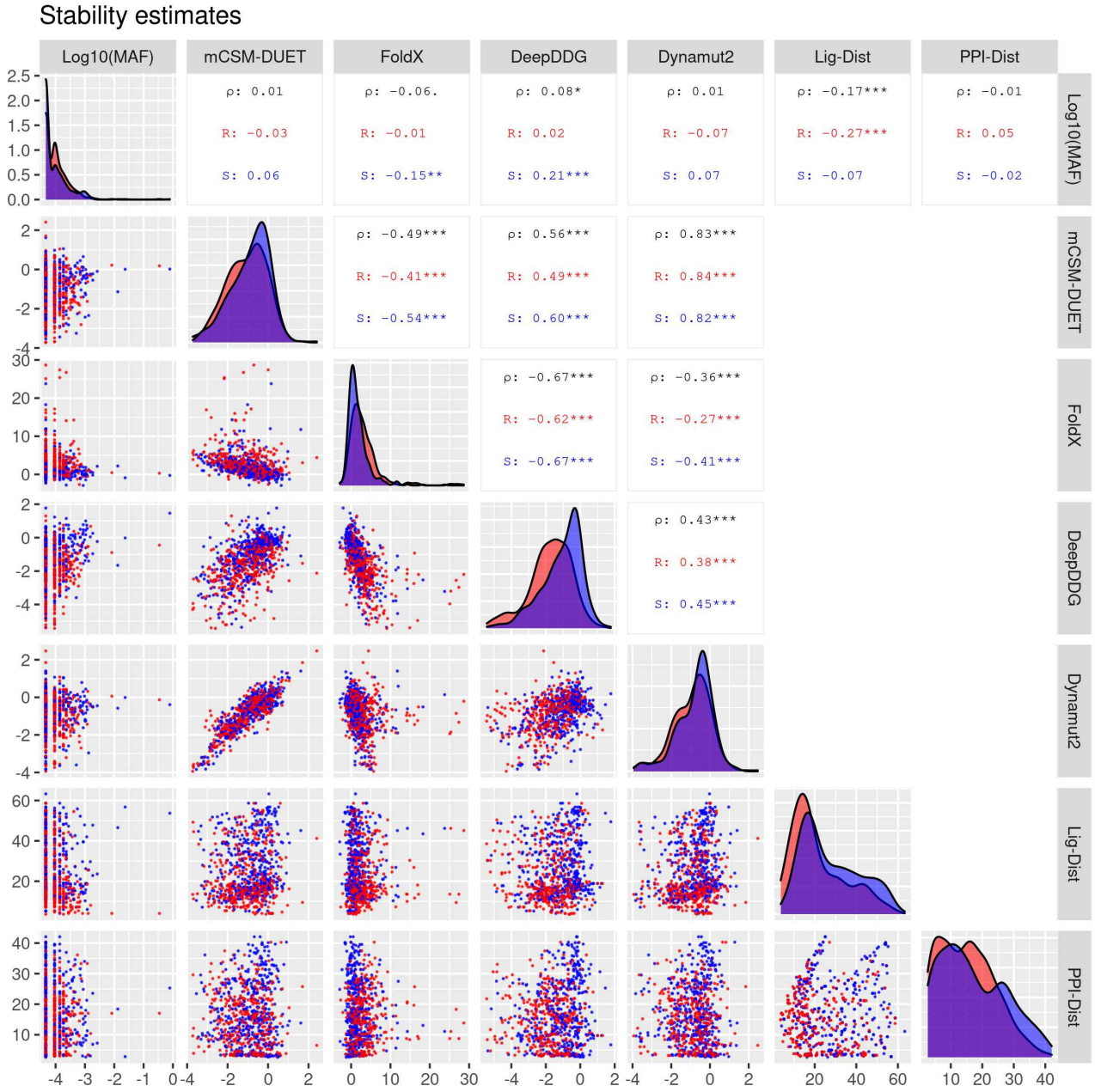


Figure 9: Correlation of protein stability changes and genomics measures
Pairwise correlations between minor allele frequency (MAF), protein stability changes (ΔΔG) estimated using DUET, FoldX, DeepDDG, and Dynamut2, and distance to INH, and the dimer interface for 817 SAVs. The upper panel in both plots include the pairwise Spearman (ρ) correlation values along with their statistical significance (.P<0.10, *P<0.05, **P<0.01, ***P<0.001). Three correlation values appear in each plot where black denotes the overall correlation with both resistant (R) and sensitive (S) mutations, while red denotes correlation estimates for resistant mutations, and blue denotes correlation estimates for sensitive mutations. The points in the lower panel represent SAVs, where red dots denote resistant mutations and blue represent sensitive mutations individually. The diagonal in each plot displays the density distribution of the corresponding parameter split by the two mutation groups. The figure is generated using R statistical software version 4.0.4, ggplot2 package. Abbreviations used: Å: Angstroms, ΔΔG: change in Gibbs free energy in Kcal/mol, SAV: single amino acid variation, Lig-Dist: distance to ligand in Å, PPI-Dist: distance to protein-protein interface in Å, INH: isoniazid.

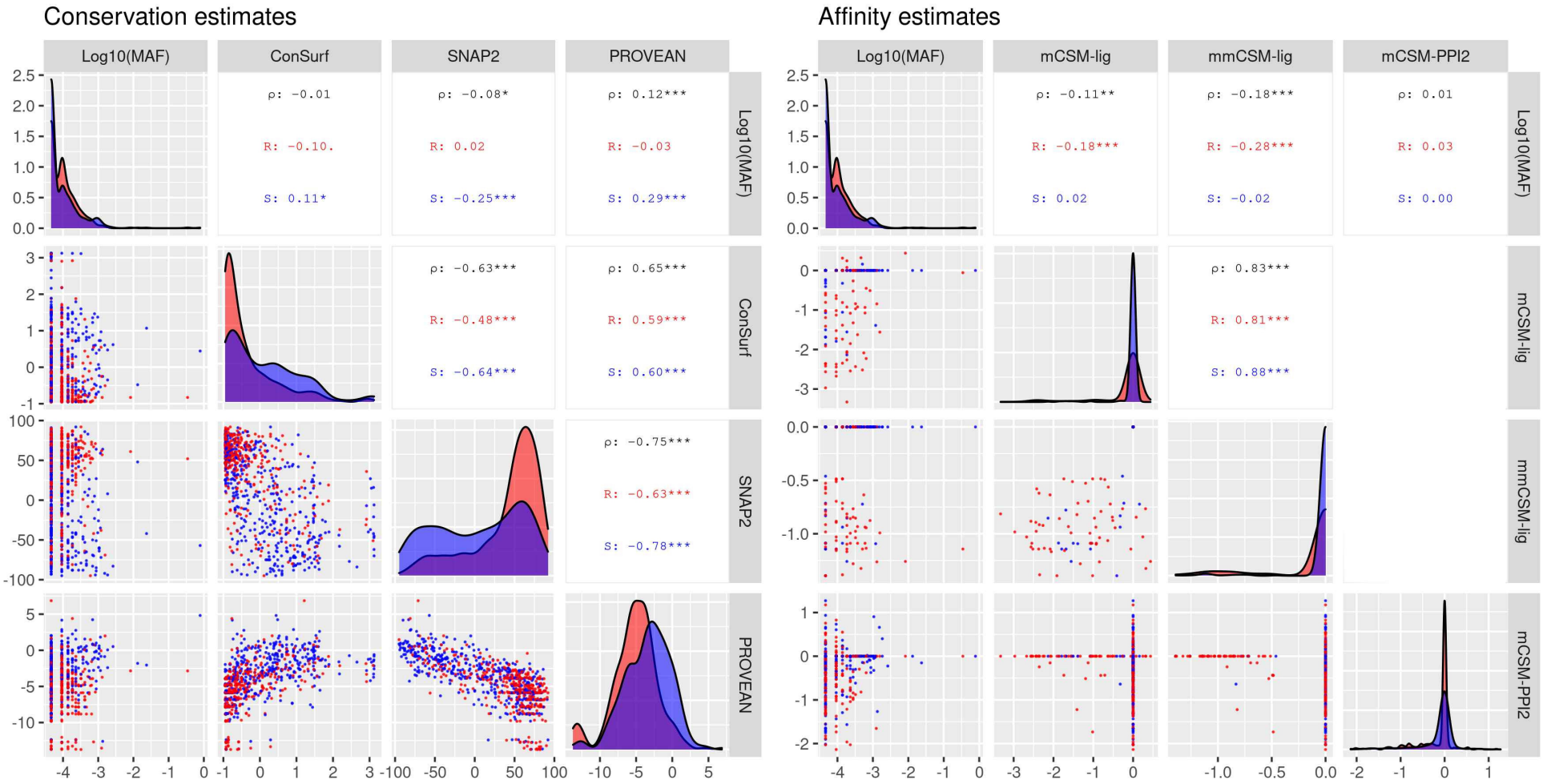


Figure 10: Correlation of evolutionary conservation, affinity changes, and genomics measures
Pairwise correlations of evolutionary conservation, affinity changes, and genomic measure of minor allele frequency (MAF) for 817 SAVs. **Left panel:** Evolutionary conservation predictors: ConSurf, SNAP2, and PROVEAN, **Right panel:** INH binding affinity changes estimated as log fold change (mCSM-lig and mmCSM-lig of 74 SAVs lying within 10Å of INH, protein-protein affinity changes ($\Delta\Delta$G) measured using mCSM-PPI2 of 260 SAVs lying within 10Å of the PPI. All corresponding affinity measures for mutations located more than 10Å of INH, and the PPI were given a value of 0 to allow complete SAVs to be used for analysis, while respecting the distance threshold for the respective tools. The upper panel in both plots include the pairwise Spearman (ρ) correlation values along with their statistical significance (.P<0.10, *P<0.05, **P<0.01, ***P<0.001). Three correlation values appear in each plot where black denotes the overall correlation with both resistant (R) and sensitive (S) mutations, while red denotes correlation estimates for resistant mutations, and blue denotes correlation estimates for sensitive mutations. The points in the lower panel represent SAVs, where red dots denote resistant mutations and blue represent sensitive mutations individually. The diagonal in each plot displays the density distribution of the corresponding parameter split by the two mutation groups. The figure is generated using R statistical software version 4.0.4, ggplot2 package. Abbreviations used: Å: Angstroms, $\Delta\Delta$G: change in Gibbs free energy in Kcal/mol, SAV: single amino acid variation, Lig-Dist: distance to ligand in Å, PPI-Dist: distance to protein-protein interface in Å, INH: isoniazid.

6.2.6 Comparing resistant and sensitive mutations

Resistant mutations occur less frequently and closer to the drug without affecting drug binding affinity, and are also likely to affect protein function despite being destabilising for protomer stability

Resistant mutations were destabilising compared with sensitive mutations for protomer stability changes across all four computational tools **(Figures 11A-D)**, with FoldX and DeepDDG being highly statistically significant (P<0.0001, **Figures 11B** and **11C)**, followed by mCSM-DUET and Dynamut2 (P<0.01, **Figures 11A** and **11D)**. Resistant mutations were also slightly less frequent compared with sensitive mutations (P<0.01, **Figure 11E)**. Similarly, compared with sensitive mutations, resistant mutations were located significantly closer to the drug **(**P<0.0001, **Figure 11F)** without affecting drug binding affinity (P<0.05, **Figures 11K** and **11L)**. Further, resistant mutations were located marginally closer to the dimer interface (P<0.05, **Figure 11G)** resulting in marginal reduction in affinity to the dimer interface (P<0.05, Figure 11M). Resistant mutations were conserved (slower rate of evolution according to ConSurf) (P<0.0001, **Figure 11H)**, and were more likely to result in deleterious impact towards protein function when assessed by both SNAP2 and PROVEAN (P<0.0001, **Figures 11J** and **11H)**.

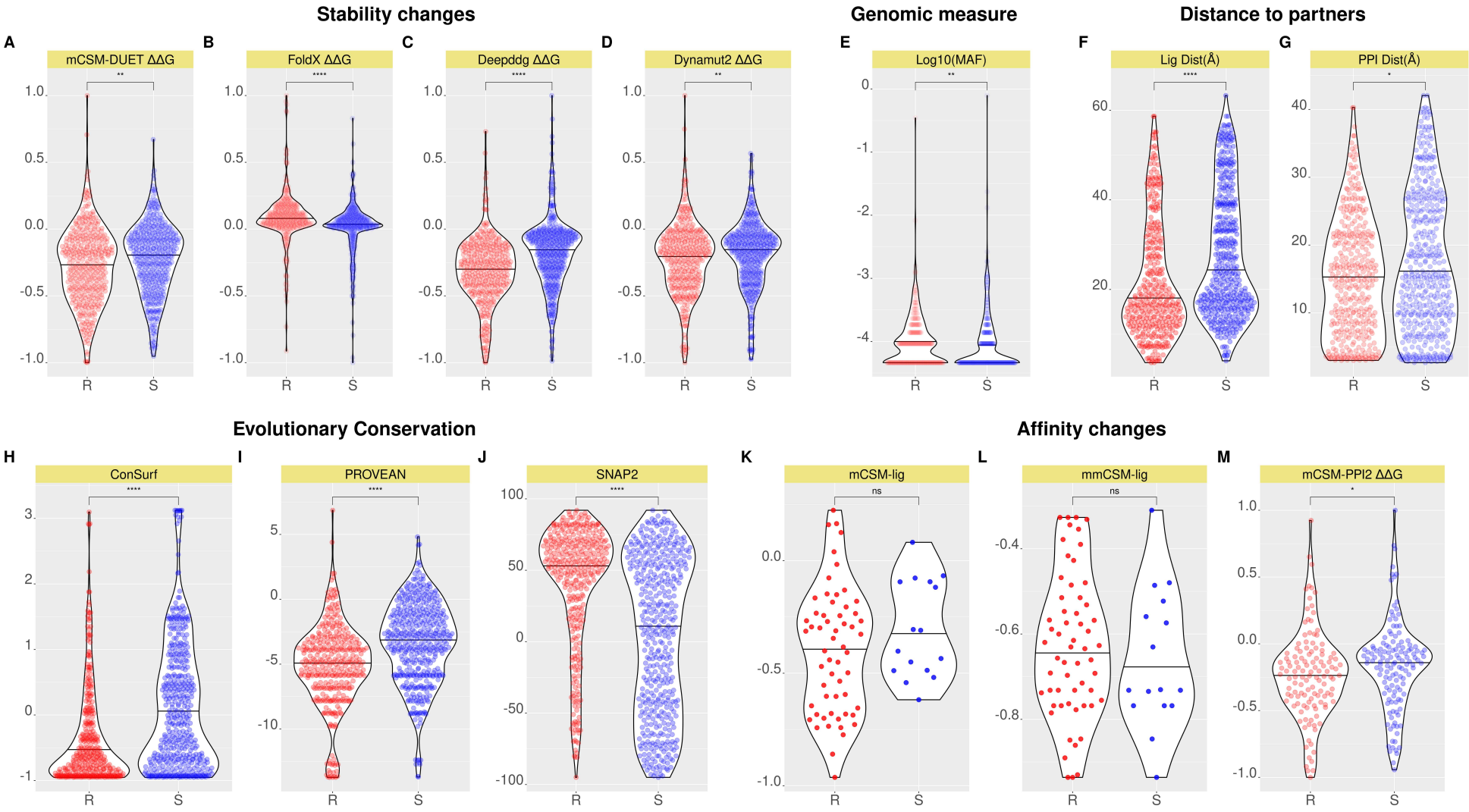

Stability changes
Genomic measure
Distance to partners
A
mCSM-DUET ΔΔG
B
FoldX ΔΔG
C
Deepddg ΔΔG
D
Dynamut2 ΔΔG
E
Log10(MAF)
F
Lig Dist(Å)
G
PPI Dist(Å)
R
S
Evolutionary Conservation
Affinity changes
H
ConSurf
I
PROVEAN
J
SNAP2
K
mCSM-lig
L
mmCSM-lig
M
mCSM-PPI2 ΔΔG
ns

Figure 11: Comparison of resistant (R) and sensitive (S) mutations
Violin plots showing the distribution of features related to structural properties, genomic measure, evolutionary conservation for 817 SAVs. For affinity changes related to the ligand (INH) binding affinity measured by mCSM- and mmCSM-lig, only mutations within 10Å of INH (n=74) were considered. Similarly, for protein-protein (PP) affinity changes measured by mCSM-PPI2, only mutations within 10Å of the PPI (n=260) were analysed. Mutations were grouped as either resistant (R, n=369) or sensitive (S, n=448) and were compared using the Wilcoxon rank-sum (unpaired) test, with statistical significance is indicated as: *P < 0.05, **P < 0.01, ***P < 0.001, ****P < 0.0001, ns>0.05. Mutations in the resistant group appear as red dots, while those in the sensitive group appear as blue dots, and the horizontal line in the violin plots display the median value. The two mutations groups were compared based on **A-D)** Stability changes (ΔΔG) estimated from four computational tools: mCSM-DUET, FoldX, DeepDDG and Dynamut2, **E)** genomic measure of average mutational occurrence (Log10MAF), **F-G)** Distance to ligand (Lig-Dist) and Distance to the PPI (PPI-Dist), **H-J)** Evolutionary conservation measured by ConSurf (<0: Conserved, >0: Variable), PROVEAN (>-2.5: Neutral, < -2.5: Deleterious) and SNAP2 (<=0: Neutral, >0: Effect) computational tools, **K-L)** Comparison of INH binding affinity changes from mCSM-lig and mmCSM-lig measured as log fold change for R (n=58) and S (n=16) mutations, and those for **M)** PP binding affinity changes (mCSM-PPI2) measured as ΔΔG for R (n=125) and S (n=135) mutations. The figure is generated using R statistical software version 4.0.4. Abbreviations used: Å: Angstroms, ΔΔG: change in Gibbs free energy in Kcal/mol, SAV: single amino acid variation, ns: not-significant, INH: isoniazid, MAF: minor allele frequency, Lig-Dist: distance to ligand in Å, PPI-Dist: distance to protein-protein interface in Å, R: resistant mutations, S: sensitive mutations.

6.2.7 Associating mutations with Odds Ratio and extreme effects

Mutations involving the active site S315T and T275P are strongly associated with INH resistance and reduction in INH binding affinity respectively

Based on DST data available for 573 (out of 817) SAVs, mutational association with resistance was further estimated using Odds Ratio (OR), with values above 1 suggesting association with INH resistance. The higher the OR, the greater the likelihood of a given mutation being resistant. This resulted in a majority (70%, n=401/573) of mutations predicted to be associated with INH resistance, much higher than observed in our data (45%, n=369/817).

An overview of mutations in KatG show that mutations involving active site residues showed strong (among top 10 mutations with high OR) association with INH resistance, with the strongest association exhibited by prominent active site residue S315T (OR=806.87), followed by S315N (OR=119.50) which were among the top 5 most frequently occurring mutations. All other SAVs at S315 were also associated with resistance: S315G (OR=36.93), S315I (OR=31.05), S315R showing lowest association with resistance among other mutations at S315 (OR=7.75) **(Figure 12).** Mutations directly following S315T and S315N, linked to INH resistance did not include INH or heme binding residues: I335V (OR=58.28), W328L (OR=50.49), Y98C (OR=38.82), D142G (OR=34.93). Residues interacting with co-factor heme: T380I (OR=34.93), F252L (OR=31.05), and P232 (OR=31.05) followed thereafter **(Figure 12).** Other prominent sites with link to resistance but not involving the active site were G699E, R484H, W161R (OR=23.29) **(Figure 12)**. Additionally, of the 13 commonly known INH resistant mutations mentioned at the start, only 3 were found in this analysis: N138S

(OR=23.27), T275P (OR=11.63), and R463L (1.54). Similarly, of the 9 high confidence INH resistant mutations mentioned in the beginning, only 3 were found in this analysis: A139P (OR=6.05), and S140N (OR=8.11), and G279D (OR=13.54) **(Figure 12)**. This suggests that despite other mutations being linked to INH resistance, the combination of highly frequency and highly resistant mutations, with little fitness cost, explains the widespread prevalence of INH driven MDR-TB prevalence.

Most frequently occurring mutation and other extreme mutational effects occurred at sites away from the active site

The most frequently occurring mutation R463L (MAF ∼52%) was located far away from INH (54Å), as well as the dimer interface (25Å), although the most destabilising mutation for INH binding affinity was co-factor heme binding residue T275P. Mutations with other extreme effects like those affecting protomer stability (Y98C: destabilising, Q679Y: stabilising), dimer affinity (W149G: destabilising, E703Q: stabilising) were not involved with the active site **(Table 1)**.

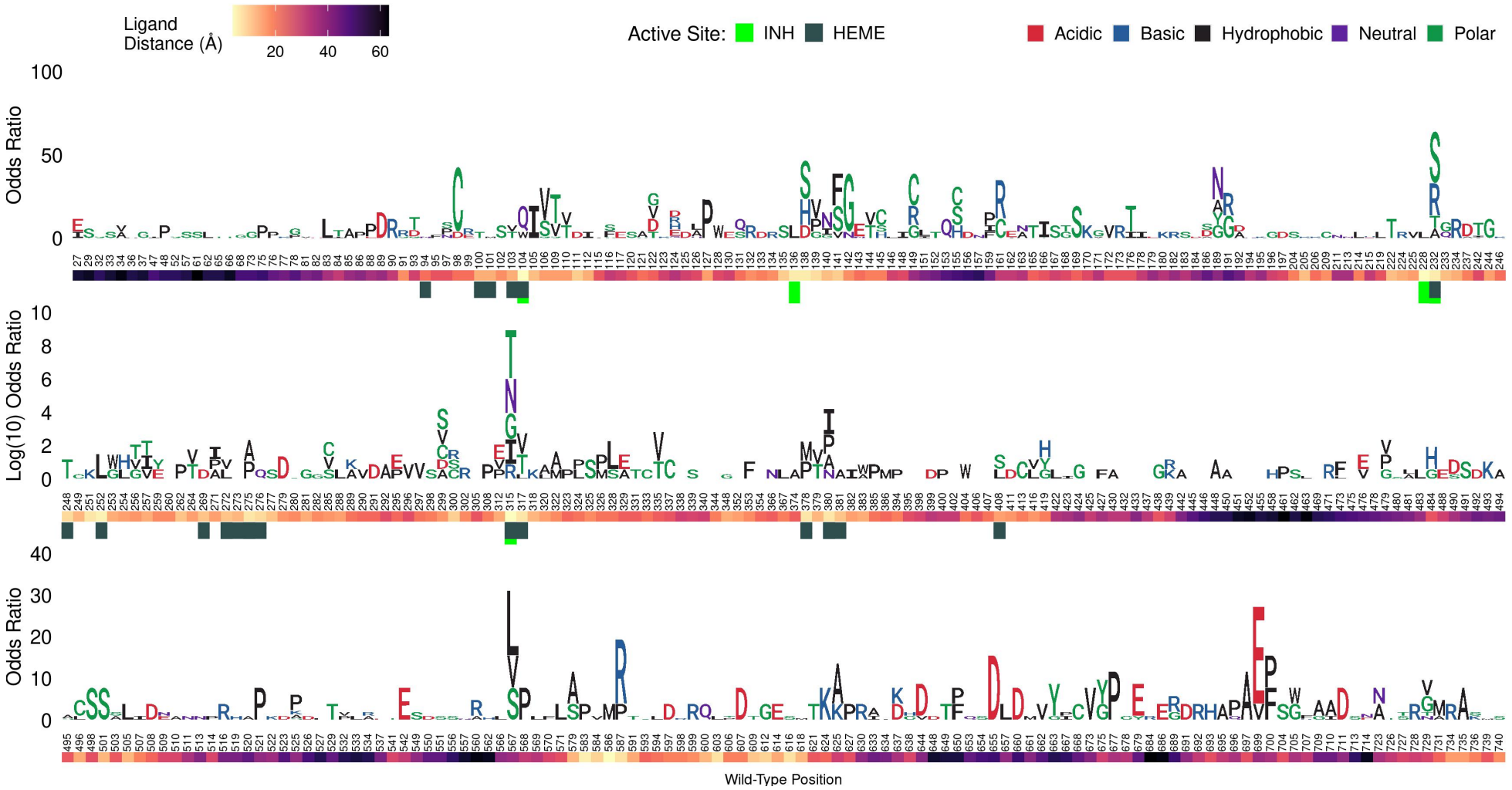


Figure 12: Logo plot showing mutational sites and their association with resistance according to Odds Ratio

Logo plot showing 573 SAVs by mutational site according to their association with INH resistance calculated using Odds Ratio (OR). The vertical axis represents the OR where letters denote mutant residues which are proportional to their corresponding OR, highlighting the most resistant mutation at each site and overall. The mutant residues are coloured according to the amino acid (aa) properties as denoted where red denotes acidic aa, basic aa appear in blue, hydrophobic aa in black, neutral aa in purple, and polar aa in darkgreen. The structural positions associated with SAVs with OR are indicated on the horizontal axis. The heat bar underneath positions indicate the distance of that position from INH according to the magma colour gradient where light yellow indicates sites closer to INH (ligand distance in Angstroms). The positions are further annotated to reflect residues involved in interactions with INH (green), and co-factor heme in dark slate grey. The figure is generated using R statistical software version 4.0.2, ggplot2 package. Abbreviations used: SAV: single amino acid variation, INH: isoniazid.

Mutation	Mutational effect	Mutational effect value	Lig-Dist (Å)	PPI-Dist (Å)	Interacting partner
S315T	Mutation with highest OR	OR = 806.87	4.06	17.02	INH and heme
R463L	Most frequent mutation	MAF (%) = 51.60	53.72	25.28	none
Y98D	Most Destabilising for protomer	ΔΔG = -0.65	13.49	9.62	none
Q679Y	Most Stabilising for protomer	ΔΔG = 0.49	49.47	6.71	none
T275P	Most Destabilising for INH binding affinity	Log fold change = -0.80	7.20	20.45	heme
W149G	Most Destabilising for PPI affinity	ΔΔG = -2.14	20.95	3.48	none
E703Q	Most Stabilising for PPI affinity	ΔΔG = 1.27	37.89	3.05	none

Table 1: Mutations with extreme effects
Mutations (SAVs) with extreme effects related to Odds Ratio (OR), mutational frequency (MAF), stability and affinity changes. For affinity changes only mutations within 10Å of INH for INH binding affinity, and Protein-Protein Interface (PPI) for PPI affinity were considered. The protomer stability changes are the average effect of all four estimates (mCSM-DUET, FoldX, DeepDDG and Dynamut2) combined, and the INH binding affinity changes are the average effect of the two mCSM based tools (mCSM-lig and mmCSM-lig) combined. Changes in PP affinity correspond to estimates from mCSM-PPI. The estimated effects were categorised as Destabilising (log fold affinity change/ΔΔG<0) and Stabilising (log fold affinity change/ΔΔG>0). Abbreviations used: Å: Angstroms, ΔΔG: change in Gibbs free energy in Kcal/mol, MAF: minor allele frequency, SAV: single amino acid variation, Lig-Dist: distance to ligand, PPI-Dist: distance to protein-protein interface, INH: isoniazid.

6.2.8 Relating lineage and protomer stability

A majority of samples contained katG mutations with a fairly homogenous SAV distribution for lineages 1-3. Mutational impact on protomer stability is dominated by high frequency resistant and sensitive mutations

About 80% of samples (n=28,106) consisted of SAVs in the protein coding region of KatG, where 26,439 samples contributed to the four main *M. tuberculosis* lineages (Lineages 1-4). Most samples with KatG mutations belonged to lineage 2 (n=12,809), followed by lineage 4 (n=5,103), lineage 3 (n=4,782) and finally by lineage 1 with the least number of samples (n=3,745) **(Figure 13A)**. All lineages were low (<10%) in their SAV diversity with lineage 4 showing slightly higher (9%, n=436) compared with other lineages: 3% for lineage 2 (n=340) and lineage 3 (n=134), followed by Lineage 1 (2%, n=93) **(Figure 13B)**.

The distributions of average stability for mutations across the four lineages were distinct for resistant and sensitive mutations. Sensitive mutations showed a prominent peak for moderate stability changes ΔΔG ~0.45 Kcal/mol across all four lineages (P<0.0001, Appendix Table 6.C.1). Resistant mutations peaked only around the marginally destabilising (ΔΔG ~ -0.01 Kcal/mol) across all four lineages (P< 0.0001, Appendix Table 6.C.1). Additionally, the distribution of protein stability for sensitive mutations in lineage 4 was multimodal with the highest peak around the moderately stabilising ΔΔG 0.45 Kcal/mol, similar to all other lineages, followed by a second peak around a mildly stabilising

ΔΔG 0.10 Kcal/mol, with a further smaller peak around the mild-moderate destabilising ΔΔG 0.25 Kcal/mol **(Figure 13C)**.

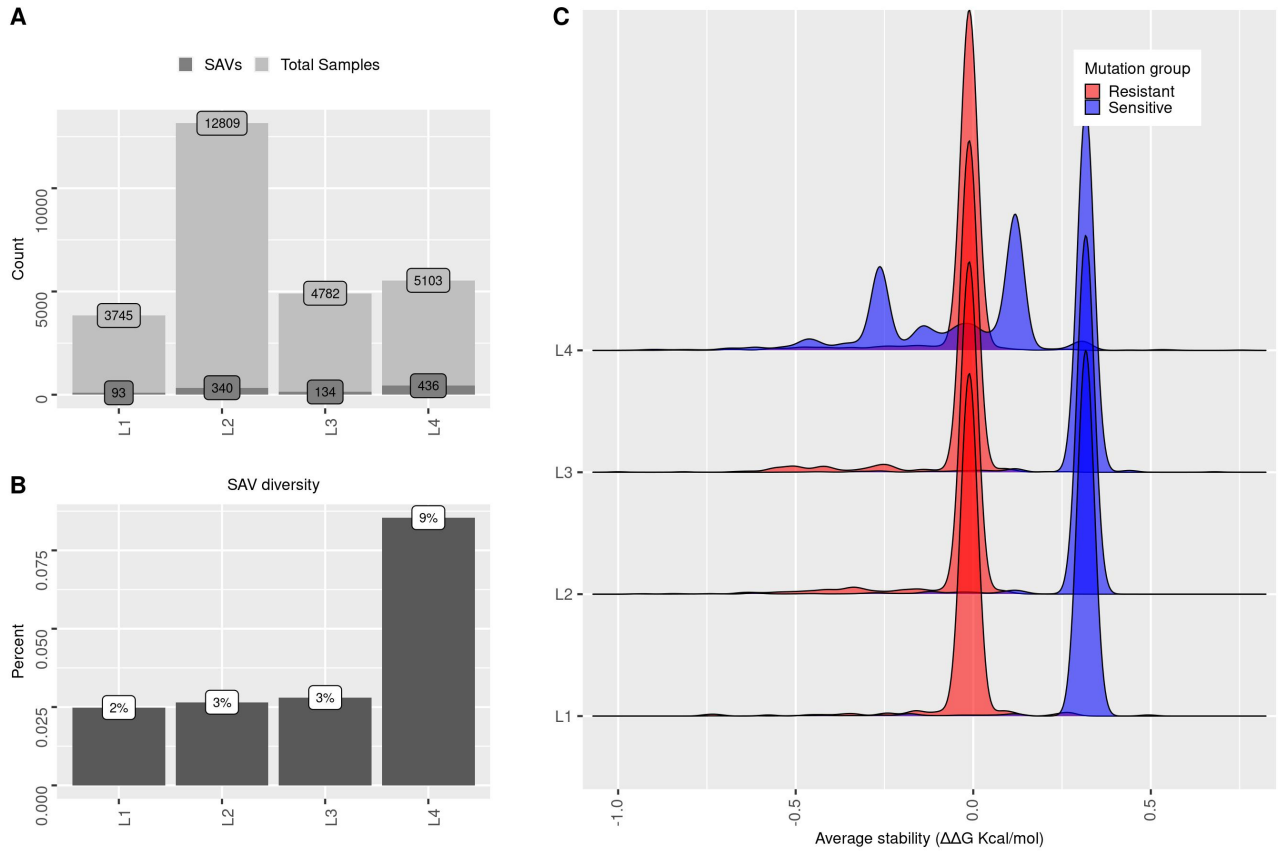


Figure 13: Lineage and protomer stability distribution
Total number of samples (n=26,439) along with the number of mutations associated with INH resistance in the four *M. tuberculosis* lineages (L1-L4). **A)** The dark grey bars show the number of mutations (SAVs), while the light grey bar show the total number of samples in each lineage, **B)** Mutational diversity in each lineage, **C)** Density distribution of lineages according to protein stability changes (ΔΔG). Estimates from four different computational tools: mCSM-DUET, FoldX, DeepDDG,and Dynamut2 were combined to calculate the average mutational stability impact for each SAV. The horizontal axis shows the average stability values (ΔΔG) (-1: highly destabilising and +1: highly stabilising) further coloured by mutational association with INH resistance: Red denotes resistant mutations (n=8,613 samples) and blue indicates sensitive mutations (n=17,826 samples). The figure is generated using R statistical software version 4.0.4. Abbreviations used: ΔΔG: change in Gibbs free energy, SAV: single amino acid variation.

6.3 Chapter summary

Mutations in *katG* are prevalent in *M. tuberculosis*. The active site residue S315 is associated with multiple SAVs, with mutation S315T being the most frequently occurring and strongly linked to INH resistance. The resistance profile of KatG is highly optimised by the presence of highly frequent, low-fitness costs mutations. Despite this, KatG displays mutational promiscuity in and beyond the active site, which comes without a large fitness penalty due to KatG being involved in processing the pro-drug to its active form rather than directly binding to INH. Mutational consequences only marginally impact protomer stability thus conferring fitness advantages towards resistance development.

References

[1] Y. Zhang et al. "The Catalase-Peroxidase Gene and Isoniazid Resistance of Mycobacterium Tuberculosis". In: *Nature* 358.6387 (Aug. 13, 1992), pp. 591–593. ISSN: 0028-0836. DOI: 10.1038/358591a0.

[2] Asesh Banerjee et al. "inhA, a Gene Encoding a Target for Isoniazid and Ethionamide in Mycobacterium Tuberculosis". In: *Science* 263.5144 (Jan. 14, 1994), pp. 227–230. ISSN: 0036-8075. DOI: 10.1126/science.8284673.

[3] Catherine Vilchèze et al. "Altered NADH/NAD+ ratio mediates coresistance to isoniazid and ethionamide in mycobacteria". In: *Antimicrobial agents and chemotherapy* 49.2 (Feb. 2005), pp. 708–720. ISSN: 0066-4804. DOI: 10.1128/AAC.49.2.708-720.2005.

[4] A Telenti. "Genetics and Pulmonary Medicine Bullet 5: Genetics of Drug Resistant Tuberculosis". In: *Thorax* 53.9 (Sept. 1, 1998), pp. 793–797. ISSN: 0040-6376. DOI: 10.1136/thx.53.9.793.

[5] Bashir A. Sheikh et al. "Development of New Therapeutics to Meet the Current Challenge of Drug Resistant Tuberculosis". In: *Current Pharmaceutical Biotechnology* 22.4 (Mar. 2021), pp. 480–500. ISSN: 13892010. DOI: 10.2174/1389201021666200628021702.

[6] Ameeruddin Nusrath Unissa et al. "Overview on Mechanisms of Isoniazid Action and Resistance in Mycobacterium Tuberculosis". In: *Infection, Genetics and Evolution* 45 (Nov. 1, 2016), pp. 474–492. ISSN: 1567-1348. DOI: 10.1016/j.meegid.2016.09.004.

[7] Pedro Eduardo Almeida Da Silva and Juan Carlos Palomino. "Molecular Basis and Mechanisms of Drug Resistance in Mycobacterium Tuberculosis: Classical and New Drugs". In: *Journal of Antimicrobial Chemotherapy* 66.7 (July 1, 2011), pp. 1417–1430. ISSN: 0305-7453. DOI: 10.1093/jac/dkr173.

[8] Ameeruddin Nusrath Unissa et al. "Significance of Catalase-Peroxidase (KatG) Mutations in Mediating Isoniazid Resistance in Clinical Strains of Mycobacterium Tuberculosis". In: *Journal of Global Antimicrobial Resistance* 15 (Dec. 2018), pp. 111–120. ISSN: 2213-7173. DOI: 10.1016/j.jgar.2018.07.001.

[9] N. L. Wengenack et al. "Recombinant Mycobacterium Tuberculosis KatG(S315T) Is a Competent Catalase-Peroxidase with Reduced Activity toward Isoniazid". In: *The Journal of Infectious Diseases* 176.3 (Sept. 1997), pp. 722–727. ISSN: 0022-1899. DOI: 10.1086/514096.

[10] Nancy L. Wengenack et al. "Evidence for Differential Binding of Isoniazid by Mycobacterium Tuberculosis KatG and the Isoniazid-Resistant Mutant KatG(S315T)". In: *ACS Publications* (Oct. 22, 1998). DOI: 10.1021/bi982023k.

[11] Xiangbo Zhao et al. "Hydrogen Peroxide-Mediated Isoniazid Activation Catalyzed by Mycobacterium Tuberculosis CatalasePeroxidase (KatG) and Its S315T Mutant," in: *Biochemistry* 45.13 (Apr. 1, 2006), pp. 4131–4140. ISSN: 0006-2960. DOI: 10.1021/bi051967o.

[12] Catherine Vilchèze and William R. Jacobs. "The Mechanism of Isoniazid Killing: Clarity through the Scope of Genetics". In: *Annual Review of Microbiology* 61 (2007), pp. 35–50. ISSN: 0066-4227. DOI: 10.1146/annurev.micro.61.111606.122346.

[13] Jody Phelan et al. "Mycobacterium Tuberculosis Whole Genome Sequencing and Protein Structure Modelling Provides Insights into Anti-Tuberculosis Drug Resistance". In: *BMC Medicine* 14.1 (Dec. 2016), p. 31. ISSN: 1741-7015. DOI: 10.1186/s12916-016-0575-9.

[14] Beate Heym et al. "Missense Mutations in the Catalase-Peroxidase Gene, katG, Are Associated with Isoniazid Resistance in Mycobacterium Tuberculosis". In: *Molecular Microbiology* 15.2 (1995), pp. 235–245. ISSN: 1365-2958. DOI: 10.1111/j.1365-2958.1995.tb02238.x.

[15] David A. Rouse et al. "Site-Directed Mutagenesis of the katG Gene of Mycobacterium Tuberculosis: Effects on CatalasePeroxidase Activities and Isoniazid Resistance". In: *Molecular Microbiology* 22.3 (1996), pp. 583–592. ISSN: 1365-2958. DOI: 10.1046/j.1365-2958.1996.00133.x.

[16] Victor Barozi et al. "Deciphering Isoniazid Drug Resistance Mechanisms on Dimeric Mycobacterium Tuberculosis KatG via Post-molecular Dynamics Analyses Including Combined Dynamic Residue Network Metrics". In: *ACS Omega* 7.15 (Apr. 7, 2022), pp. 13313–13332. ISSN: 2470-1343. DOI: 10.1021/acsomega.2c01036.

[17] Thomas Bertrand et al. “Crystal Structure of Mycobacterium Tuberculosis Catalase-Peroxidase”. In: *The Journal of Biological Chemistry* 279.37 (Sept. 10, 2004), pp. 38991–38999. ISSN: 0021-9258. DOI: 10.1074/jbc.M402382200.

[18] Pietro Vidossich et al. “Binding of the Antitubercular Pro-Drug Isoniazid in the Heme Access Channel of Catalase-Peroxidase (KatG). A Combined Structural and Metadynamics Investigation”. In: *The Journal of Physical Chemistry. B* 118.11 (Mar. 20, 2014), pp. 2924–2931. ISSN: 1520-5207. DOI: 10.1021/jp4123425.

Appendix for Chapter 6

6.A Mutations close to isoniazid

Mutation	Interacting partner	Lig-Dist (Å)	mCSM-lig affinity	mCSM-lig outcome	mmCSM-lig affinity	mmCSM-lig outcome	MAF (%)	Odds Ratio	P-value	Adjusted P-value	Adjusted P-value significance
S315T	drug and heme	4.06	-0.06	Destabilising	-1.14	Destabilising	34.63	806.87	<0.0001	<0.0001	****
S315N	drug and heme	4.06	0.44	Stabilising	-1.26	Destabilising	0.83	119.5	<0.0001	<0.0001	****
S315G	drug and heme	4.06	-0.85	Destabilising	-1.26	Destabilising	0.13	36.93	<0.0001	<0.0001	****
T380I	heme	7.24	-2.23	Destabilising	-1.14	Destabilising	0.06	34.93	<0.001	0.01	**
F252L	heme	6.25	-1.63	Destabilising	-0.93	Destabilising	0.04	31.05	<0.001	0.01	**
P232S	drug and heme	4.55	0.32	Stabilising	-1.33	Destabilising	0.05	31.05	<0.001	0.01	**
S315I	drug and heme	4.06	-0.51	Destabilising	-0.88	Destabilising	0.08	31.05	<0.001	0.01	**
N138S	none	5.39	-2.43	Destabilising	-0.98	Destabilising	0.03	23.28	<0.001	0.04	*
Q127P	none	9.54	-2.09	Destabilising	-0.64	Destabilising	0.06	23.28	<0.001	0.04	*
M105I	none	8.28	-0.97	Destabilising	-0.85	Destabilising	0.03	19.4	<0.05	0.07	ns
L141F	none	7.13	-0.84	Destabilising	-1.04	Destabilising	0.03	19.4	<0.05	0.07	ns
P232R	drug and heme	4.55	-0.27	Destabilising	-0.77	Destabilising	0.03	19.4	<0.05	0.07	ns
I317V	heme	7.36	-0.92	Destabilising	-1.09	Destabilising	0.04	19.4	<0.05	0.07	ns
T275A	heme	7.2	-2.52	Destabilising	-1.14	Destabilising	0.05	19.4	<0.05	0.07	ns
A109T	none	9.51	-0.45	Destabilising	-0.94	Destabilising	0.06	17.47	0<0.	0.01	*
N138H	none	5.39	-0.94	Destabilising	-0.94	Destabilising	0.06	15.52	<0.05	0.03	*
L378M	heme	7.03	-2.54	Destabilising	-1.09	Destabilising	0.02	15.51	0.01	0.14	ns
R104Q	drug and heme	3.66	-2.08	Destabilising	-1.14	Destabilising	0.02	15.51	0.01	0.14	ns
I248T	heme	8.07	0.08	Stabilising	-0.85	Destabilising	0.02	15.51	0.01	0.14	ns
I317T	heme	7.36	0.24	Stabilising	-1.03	Destabilising	0.02	15.51	0.01	0.14	ns
T380P	heme	7.24	-2.26	Destabilising	-0.92	Destabilising	0.02	15.51	0.01	0.14	ns
L141S	none	7.13	0.16	Stabilising	-0.72	Destabilising	0.05	15.51	0.01	0.14	ns

A139V	none	6.05	-1.39	Destabilising	-1.09	Destabilising	0.12	11.64	0.01	0.11	ns
T275P	heme	7.2	-2.35	Destabilising	-1.38	Destabilising	0.01	11.63	0.04	0.29	ns
A222T	none	7.7	-0.93	Destabilising	-1.1	Destabilising	0.02	11.63	0.04	0.29	ns
G234R	none	9.48	-1.08	Destabilising	-1	Destabilising	0.02	11.63	0.04	0.29	ns
L378P	heme	7.03	-2.06	Destabilising	-1.17	Destabilising	0.02	11.63	0.04	0.29	ns
S140N	none	8.11	-1.56	Destabilising	-0.99	Destabilising	0.16	9.7	0.02	0.19	ns
A109V	none	9.51	-1.85	Destabilising	-0.86	Destabilising	0.02	7.76	0.05	0.33	ns
A139P	none	6.05	-1.39	Destabilising	-0.71	Destabilising	0.01	7.75	0.12	0.38	ns
P232T	drug and heme	4.55	0.31	Stabilising	-0.77	Destabilising	0.01	7.75	0.12	0.38	ns
W135S	none	7.47	-1.02	Destabilising	-0.49	Destabilising	0.01	7.75	0.12	0.38	ns
P136L	drug	4.72	-1.13	Destabilising	-0.7	Destabilising	0.01	7.75	0.12	0.38	ns
N138D	none	5.39	-1.72	Destabilising	-1.16	Destabilising	0.01	7.75	0.12	0.38	ns
T380A	heme	7.24	-2.37	Destabilising	-1.14	Destabilising	0.01	7.75	0.12	0.38	ns
S315R	drug and heme	4.06	-0.52	Destabilising	-0.84	Destabilising	0.03	7.75	0.12	0.38	ns
E233G	none	7.92	-3.33	Destabilising	-0.81	Destabilising	0.02	5.82	0.12	0.38	ns
P232A	drug and heme	4.55	-1.33	Destabilising	-1.28	Destabilising	0.03	5.82	0.12	0.38	ns
I103T	heme	9.18	-0.62	Destabilising	-0.56	Destabilising	0	3.88	0.34	0.5	ns
R104W	drug and heme	3.66	-0.57	Destabilising	-0.72	Destabilising	0	3.88	0.34	0.5	ns
G111D	none	7.31	-1.91	Destabilising	-1.39	Destabilising	0	3.88	0.34	0.5	ns
S140I	none	8.11	-2.42	Destabilising	-0.58	Destabilising	0	3.88	0.34	0.5	ns
L141V	none	7.13	-1.03	Destabilising	-1.09	Destabilising	0	3.88	0.34	0.5	ns
Q224R	none	8.66	-2.36	Destabilising	-0.92	Destabilising	0	3.88	0.34	0.5	ns
M225V	none	9.91	-1.41	Destabilising	-0.82	Destabilising	0	3.88	0.34	0.5	ns
T251K	none	7.38	-2.56	Destabilising	-1.12	Destabilising	0	3.88	0.34	0.5	ns
H276Q	heme	7.82	-0.81	Destabilising	-0.79	Destabilising	0	3.88	0.34	0.5	ns
W300C	none	9.48	-1.19	Destabilising	-0.53	Destabilising	0	3.88	0.34	0.5	ns

T380N	heme	7.24	-1.02	Destabilising	-1.39	Destabilising	0	3.88	0.34	0.5	ns
D381A	heme	9.53	-2.36	Destabilising	-0.62	Destabilising	0	3.88	0.34	0.5	ns
P100T	heme	9.22	-0.64	Destabilising	-1.06	Destabilising	0.01	3.88	0.34	0.5	ns
I103V	heme	9.18	-1.52	Destabilising	-0.51	Destabilising	0.01	3.88	0.34	0.5	ns
T112I	none	9.1	-2.57	Destabilising	-0.88	Destabilising	0.01	3.88	0.34	0.5	ns
I228L	drug	4.07	-1.06	Destabilising	-1.09	Destabilising	0.01	3.88	0.34	0.5	ns
E233Q	none	7.92	-2.44	Destabilising	-0.89	Destabilising	0.01	3.88	0.34	0.5	ns
W300S	none	9.48	-0.71	Destabilising	-0.49	Destabilising	0.01	3.88	0.34	0.5	ns
W300R	none	9.48	-0.81	Destabilising	-0.49	Destabilising	0.01	3.88	0.27	0.5	ns
A139G	none	6.05	-1.56	Destabilising	-1.26	Destabilising	0.02	3.88	0.34	0.5	ns
S140G	none	8.11	-2.67	Destabilising	-1.03	Destabilising	0.01	1.94	>1	>1	ns
I317L	heme	7.36	-1.08	Destabilising	-1.09	Destabilising	0.06	1.94	>1	>1	ns
L205R	none	8.35	-0.32	Destabilising	-0.83	Destabilising	0	0.97	>1	>1	ns
A109S	none	9.51	-0.33	Destabilising	-0.94	Destabilising	0.01	0.97	>1	>1	ns
L101M	heme	9.3	-2.14	Destabilising	-0.85	Destabilising	0.02	0.97	>1	>1	ns
L101F	heme	9.3	-1.69	Destabilising	-0.78	Destabilising	0	NA	NA	NA	ns
I103N	heme	9.18	-0.27	Destabilising	-0.46	Destabilising	0	NA	NA	NA	ns
A109D	none	9.51	-0.23	Destabilising	-0.71	Destabilising	0	NA	NA	NA	ns
L141I	none	7.13	-1.5	Destabilising	-1.09	Destabilising	0	NA	NA	NA	ns
N231K	heme	5.3	-1.87	Destabilising	-1.09	Destabilising	0	NA	NA	NA	ns
T251M	none	7.38	-2.97	Destabilising	-0.98	Destabilising	0	NA	NA	NA	ns
T314N	heme	5.22	-0.41	Destabilising	-1.39	Destabilising	0	NA	NA	NA	ns
T314S	heme	5.22	-1.07	Destabilising	-1.14	Destabilising	0	NA	NA	NA	ns
T380S	heme	7.24	-1.68	Destabilising	-1.14	Destabilising	0	NA	NA	NA	ns
W300G	none	9.48	-1.74	Destabilising	-0.49	Destabilising	0.01	NA	NA	NA	ns
T314A	heme	5.22	-1.79	Destabilising	-1.14	Destabilising	0.01	NA	NA	NA	ns

Table 6.A.1: Mutations close to INH
Seventy-four single amino acid variation (SAV) mutations lying within 10Å of INH and their corresponding ligand affinity changes (log fold change) measured by mCSM-Lig and mmCSM-lig. The estimated effect are categorised as Destabilising (log fold affinity change<0) and Stabilising ($\Delta\Delta$G>0). The genomic measures of minor allele frequency (MAF), Odds Ratio (OR), OR related P-values, and FDR adjusted P-values are shown. Statistical significance indicated as: *P < 0.05, **P < 0.01, ***P < 0.001, ****P < 0.0001, ns: >0.05. The table is arranged by OR to show mutation with the highest OR at the top for mutations close to INH. Columns with NA indicate insufficient data to calculate OR. Abbreviations used: $\Delta\Delta$G: change in Gibbs free energy in Kcal/mol, FDR: false discovery rate, ns: not significant, INH: isoniazid.

6.B Mutations close to the protein-protein interface

Mutation	Interacting partner	PPI-Dist (Å)	mCSM-PPI2 (ΔΔG)	mCSM-PPI2 out-come	MAF (%)	Odds Ratio	P-value	Adjusted P-value	Adjusted P-value significance
Y98C	none	9.62	-0.38	Decreasing	0.06	38.82	<0.0001	<0.01	**
D142G	none	7.23	-0.3	Decreasing	0.04	34.93	<0.0001	<0.01	**
W161R	none	3.2	-0.88	Decreasing	0.03	23.28	<0.001	0.04	*
G699E	none	3.42	-0.99	Decreasing	0.03	23.28	<0.001	0.04	*
Q127P	none	7.53	-0.15	Decreasing	0.06	23.28	<0.001	0.04	*
W149C	none	3.48	-2.03	Decreasing	0.02	19.4	<0.01	0.07	ns
G299S	none	8.75	-0.72	Decreasing	0.02	19.4	<0.01	0.07	ns
D189N	none	7.18	-0.84	Decreasing	0.03	19.4	<0.01	0.07	ns
N655D	none	8.07	-0.01	Decreasing	0.02	15.51	0.01	0.14	ns
P89D	none	6.92	0.02	Increasing	0.07	15.51	0.01	0.14	ns
W191R	none	3.35	-1.65	Decreasing	0.12	13.59	<0.0001	<0.001	***
W191G	none	3.35	-1.27	Decreasing	0.13	13.59	<0.0001	<0.001	***
Y155C	none	4.51	-0.82	Decreasing	0.05	11.64	0.01	0.11	ns
W90R	none	3.3	-1.09	Decreasing	0.01	11.63	0.04	0.29	ns
V697A	none	4.31	-0.79	Decreasing	0.01	11.63	0.04	0.29	ns
W149R	none	3.48	-1.96	Decreasing	0.02	11.63	0.04	0.29	ns
Y155S	none	4.51	-0.79	Decreasing	0.02	11.63	0.04	0.29	ns
W161C	none	3.2	-1.04	Decreasing	0.02	11.63	0.04	0.29	ns
T677P	none	8.47	-0.62	Decreasing	0.03	11.63	0.04	0.29	ns
K153Q	none	2.68	0.9	Increasing	0.11	11.63	0.04	0.29	ns
R145C	none	6.31	-0.36	Decreasing	0.03	9.7	0.02	0.19	ns
G297V	none	3.74	-0.96	Decreasing	0.03	9.7	0.02	0.19	ns

T625A	none	7.49	-0.26	Decreasing	0.09	9.7	0.02	0.19	ns
K27E	none	3.85	-0.93	Decreasing	0.02	7.76	0.05	0.33	ns
D189A	none	7.18	-0.56	Decreasing	0.02	7.76	0.05	0.33	ns
P131Q	none	3.03	-0.77	Decreasing	0.01	7.75	0.12	0.38	ns
L132R	none	5.85	-0.4	Decreasing	0.01	7.75	0.12	0.38	ns
A144T	none	9.13	0.55	Increasing	0.01	7.75	0.12	0.38	ns
Y155H	none	4.51	-0.17	Decreasing	0.01	7.75	0.12	0.38	ns
D189Y	none	7.18	0.08	Increasing	0.01	7.75	0.12	0.38	ns
G299V	none	8.75	-0.81	Decreasing	0.01	7.75	0.12	0.38	ns
N660D	none	7.23	0.21	Increasing	0.01	7.75	0.12	0.38	ns
L48P	none	4.73	-1.61	Decreasing	0.01	7.75	0.12	0.38	ns
Q88P	none	6.49	-0.29	Decreasing	0.01	7.75	0.12	0.38	ns
R128W	none	3.63	-0.34	Decreasing	0.01	7.75	0.12	0.38	ns
W135S	none	6.08	-1.73	Decreasing	0.01	7.75	0.12	0.38	ns
P136L	drug	7.77	-0.15	Decreasing	0.01	7.75	0.12	0.38	ns
K143E	none	6.71	-0.69	Decreasing	0.01	7.75	0.12	0.38	ns
A144V	none	9.13	0.22	Increasing	0.01	7.75	0.12	0.38	ns
W149G	none	3.48	-2.14	Decreasing	0.01	7.75	0.12	0.38	ns
L159P	none	7.27	-0.39	Decreasing	0.01	7.75	0.12	0.38	ns
I165T	none	8.81	-0.17	Decreasing	0.01	7.75	0.12	0.38	ns
A291D	none	4.94	0.33	Increasing	0.01	7.75	0.12	0.38	ns
P292A	none	4.1	-0.23	Decreasing	0.01	7.75	0.12	0.38	ns
Q295E	none	5.2	0.83	Increasing	0.01	7.75	0.12	0.38	ns
M296V	none	3.22	-0.91	Decreasing	0.01	7.75	0.12	0.38	ns
G299C	none	8.75	-0.65	Decreasing	0.01	7.75	0.12	0.38	ns
E607D	none	4.49	-0.29	Decreasing	0.01	7.75	0.12	0.38	ns

M624K	none	8.44	0.09	Increasing	0.01	7.75	0.12	0.38	ns
D663Y	none	5.56	0.02	Increasing	0.01	7.75	0.12	0.38	ns
Q679E	none	6.71	0.43	Increasing	0.01	7.75	0.12	0.38	ns
S700F	none	3.59	-0.53	Decreasing	0.01	7.75	0.12	0.38	ns
S700P	none	3.59	-0.64	Decreasing	0.01	7.75	0.12	0.38	ns
Y711D	none	3.92	-1.37	Decreasing	0.01	7.75	0.12	0.38	ns
R254H	none	9.68	-0.03	Decreasing	0.02	7.75	0.12	0.38	ns
D189G	none	7.18	-0.48	Decreasing	0.03	7.75	0.12	0.38	ns
P29S	none	3.32	-1.21	Decreasing	0.02	5.82	0.12	0.38	ns
D675Y	none	6.99	0	Increasing	0.02	5.82	0.12	0.38	ns
K27I	none	3.85	-0.53	Decreasing	0	3.88	0.34	0.5	ns
G33S	none	4.72	-0.52	Decreasing	0	3.88	0.34	0.5	ns
G34A	none	5.34	-0.07	Decreasing	0	3.88	0.34	0.5	ns
G34V	none	5.34	-0.3	Decreasing	0	3.88	0.34	0.5	ns
D37G	none	6.24	-0.53	Decreasing	0	3.88	0.34	0.5	ns
P57S	none	3.12	-0.93	Decreasing	0	3.88	0.34	0.5	ns
F62L	none	6.76	-0.84	Decreasing	0	3.88	0.34	0.5	ns
Y98D	none	9.62	-0.51	Decreasing	0	3.88	0.34	0.5	ns
H116S	none	6.49	-0.22	Decreasing	0	3.88	0.34	0.5	ns
G124E	none	9.73	0.01	Increasing	0	3.88	0.34	0.5	ns
G124H	none	9.73	-0.07	Decreasing	0	3.88	0.34	0.5	ns
M126A	none	8.23	0.3	Increasing	0	3.88	0.34	0.5	ns
M126L	none	8.23	0.11	Increasing	0	3.88	0.34	0.5	ns
A130E	none	3.33	1.17	Increasing	0	3.88	0.34	0.5	ns
D142N	none	7.23	-0.76	Decreasing	0	3.88	0.34	0.5	ns
R145S	none	6.31	-0.24	Decreasing	0	3.88	0.34	0.5	ns

L148I	none	7.39	-0.63	Decreasing	0	3.88	0.34	0.5	ns
K152T	none	4.4	-1.55	Decreasing	0	3.88	0.34	0.5	ns
G156D	none	3.65	-1.28	Decreasing	0	3.88	0.34	0.5	ns
L159F	none	7.27	0.76	Increasing	0	3.88	0.34	0.5	ns
A162E	none	9.46	0.52	Increasing	0	3.88	0.34	0.5	ns
E192A	none	3.46	-0.56	Decreasing	0	3.88	0.34	0.5	ns
E192D	none	3.46	-0.23	Decreasing	0	3.88	0.34	0.5	ns
V196G	none	3.81	-0.83	Decreasing	0	3.88	0.34	0.5	ns
Y197D	none	3.46	-1.88	Decreasing	0	3.88	0.34	0.5	ns
W204S	none	3.39	-1.24	Decreasing	0	3.88	0.34	0.5	ns
R209C	none	7.53	-0.35	Decreasing	0	3.88	0.34	0.5	ns
P219L	none	3.28	-0.99	Decreasing	0	3.88	0.34	0.5	ns
Q224R	none	8.71	-0.24	Decreasing	0	3.88	0.34	0.5	ns
M225V	none	3.79	-1.22	Decreasing	0	3.88	0.34	0.5	ns
R254L	none	9.68	-0.11	Decreasing	0	3.88	0.34	0.5	ns
P288L	none	5.64	-1.02	Decreasing	0	3.88	0.34	0.5	ns
E289K	none	2.99	-1.31	Decreasing	0	3.88	0.34	0.5	ns
A290V	none	3.95	-0.07	Decreasing	0	3.88	0.34	0.5	ns
Q295P	none	5.2	-0.29	Decreasing	0	3.88	0.34	0.5	ns
L298S	none	3.56	-0.81	Decreasing	0	3.88	0.34	0.5	ns
G299A	none	8.75	-0.49	Decreasing	0	3.88	0.34	0.5	ns
G299D	none	8.75	-0.59	Decreasing	0	3.88	0.34	0.5	ns
W300C	none	8.4	-0.51	Decreasing	0	3.88	0.34	0.5	ns
D612G	none	4.41	-0.43	Decreasing	0	3.88	0.34	0.5	ns
A614E	none	8.39	0.36	Increasing	0	3.88	0.34	0.5	ns
A621T	none	4.92	0.49	Increasing	0	3.88	0.34	0.5	ns

F657L	none	9.29	-0.8	Decreasing	0	3.88	0.34	0.5	ns
L662V	none	3.29	-1.16	Decreasing	0	3.88	0.34	0.5	ns
W668C	none	3.26	-1.82	Decreasing	0	3.88	0.34	0.5	ns
D675G	none	6.99	-0.12	Decreasing	0	3.88	0.34	0.5	ns
W689R	none	9.53	-0.3	Decreasing	0	3.88	0.34	0.5	ns
G691D	none	8.54	-0.37	Decreasing	0	3.88	0.34	0.5	ns
D695A	none	6.52	-0.84	Decreasing	0	3.88	0.34	0.5	ns
L696P	none	3.3	-2.03	Decreasing	0	3.88	0.34	0.5	ns
G699V	none	3.42	-1.29	Decreasing	0	3.88	0.34	0.5	ns
L704S	none	6.06	-0.71	Decreasing	0	3.88	0.34	0.5	ns
R705G	none	3.04	-1.34	Decreasing	0	3.88	0.34	0.5	ns
R705W	none	3.04	-0.38	Decreasing	0	3.88	0.34	0.5	ns
V710A	none	3.29	-1	Decreasing	0	3.88	0.34	0.5	ns
D723A	none	4.24	-0.54	Decreasing	0	3.88	0.34	0.5	ns
D723N	none	4.24	-0.73	Decreasing	0	3.88	0.34	0.5	ns
A61S	none	8.82	-0.09	Decreasing	0.01	3.88	0.34	0.5	ns
D117E	none	6.29	-0.33	Decreasing	0.01	3.88	0.34	0.5	ns
G125D	none	8.84	-0.45	Decreasing	0.01	3.88	0.34	0.5	ns
P131S	none	3.03	-0.92	Decreasing	0.01	3.88	0.34	0.5	ns
N133D	none	3.89	0.23	Increasing	0.01	3.88	0.34	0.5	ns
S134R	none	3.78	0.09	Increasing	0.01	3.88	0.34	0.5	ns
R145H	none	6.31	-0.13	Decreasing	0.01	3.88	0.34	0.5	ns
E289A	none	2.99	-0.97	Decreasing	0.01	3.88	0.34	0.5	ns
W300S	none	8.4	-0.44	Decreasing	0.01	3.88	0.34	0.5	ns
T625K	none	7.49	-0.21	Decreasing	0.01	3.88	0.34	0.5	ns
S692R	none	5.67	-0.34	Decreasing	0.01	3.88	0.34	0.5	ns

R693H	none	3.77	-0.44	Decreasing	0.01	3.88	0.34	0.5	ns
W300R	none	8.4	-0.44	Decreasing	0.01	3.88	0.27	0.5	ns
W91R	none	7.83	-0.58	Decreasing	0.02	3.88	0.27	0.5	ns
H116F	none	6.49	0.17	Increasing	0.02	3.88	0.34	0.5	ns
G124D	none	9.73	-0.28	Decreasing	0.02	3.88	0.27	0.5	ns
E709A	none	3.14	-0.93	Decreasing	0.02	3.88	0.34	0.5	ns
G124R	none	9.73	-0.31	Decreasing	0.03	3.88	0.34	0.5	ns
V151L	none	6.87	-0.14	Decreasing	0.05	3.88	0.34	0.5	ns
L159I	none	7.27	-0.23	Decreasing	0.26	3.88	0.07	0.38	ns
V68G	none	9.95	0.23	Increasing	0.01	1.94	>1	>1	ns
W689G	none	9.53	-0.29	Decreasing	0.02	1.94	>1	>1	ns
K157N	none	3.07	0.66	Increasing	0.15	1.94	0.32	0.5	ns
V151I	none	6.87	0.4	Increasing	0.18	1.16	>1	>1	ns
G32S	none	3.41	-0.2	Decreasing	0	0.97	>1	>1	ns
V47I	none	4.66	0.66	Increasing	0	0.97	>1	>1	ns
W91S	none	7.83	-0.56	Decreasing	0	0.97	>1	>1	ns
K143N	none	6.71	-0.18	Decreasing	0	0.97	>1	>1	ns
R146L	none	3.17	-0.39	Decreasing	0	0.97	>1	>1	ns
L205R	none	6.13	-0.64	Decreasing	0	0.97	>1	>1	ns
G206R	none	6.68	-0.58	Decreasing	0	0.97	>1	>1	ns
P603L	none	3.35	-0.27	Decreasing	0	0.97	>1	>1	ns
L616S	none	9.59	-0.16	Decreasing	0	0.97	>1	>1	ns
L661M	none	4.49	-0.79	Decreasing	0	0.97	>1	>1	ns
T667I	none	5.79	-0.13	Decreasing	0	0.97	>1	>1	ns
Y678C	none	3.86	-1.7	Decreasing	0	0.97	>1	>1	ns
Q679Y	none	6.71	-0.13	Decreasing	0	0.97	>1	>1	ns

L696Q	none	3.3	-1.56	Decreasing	0	0.97	>1	>1	ns
L707F	none	3.42	0.94	Increasing	0	0.97	>1	>1	ns
E709G	none	3.14	-1.08	Decreasing	0	0.97	>1	>1	ns
V710I	none	3.29	-0.74	Decreasing	0	0.97	>1	>1	ns
D714N	none	2.84	-0.31	Decreasing	0	0.97	>1	>1	ns
P52S	none	3.95	0.24	Increasing	0.01	0.97	>1	>1	ns
Q88E	none	6.49	-0.22	Decreasing	0.01	0.97	>1	>1	ns
E195K	none	3.24	-1.02	Decreasing	0.01	0.97	>1	>1	ns
M609T	none	5.07	-0.52	Decreasing	0.01	0.97	>1	>1	ns
D663G	none	5.56	-0.8	Decreasing	0.01	0.97	>1	>1	ns
A713S	none	3.82	0.73	Increasing	0.01	0.97	>1	>1	ns
T667P	none	5.79	-0.5	Decreasing	0.02	0.97	>1	>1	ns
G124Q	none	9.73	-0.32	Decreasing	0.11	0.97	>1	>1	ns
M126Q	none	8.23	-0.05	Decreasing	0.11	0.97	>1	>1	ns
D714E	none	2.84	-0.15	Decreasing	2.36	0.51	<0.01	0.04	*
H116G	none	6.49	0.01	Increasing	0.01	0.48	0.55	0.77	ns
A162V	none	9.46	0.37	Increasing	0.01	0.48	0.55	0.77	ns
A606T	none	7.8	0.17	Increasing	0.01	0.48	0.55	0.77	ns
D194N	none	5.98	-0.49	Decreasing	0.01	0.48	0.55	0.77	ns
Q295A	none	5.2	-0.34	Decreasing	0.01	0.48	0.55	0.77	ns
T618M	none	6.94	-0.06	Decreasing	0.01	0.48	0.55	0.77	ns
A606P	none	7.8	-0.1	Decreasing	0.02	0.48	0.55	0.77	ns
Q36P	none	3.24	-0.77	Decreasing	0.02	0.32	0.56	0.77	ns
L707R	none	3.42	-1.57	Decreasing	0.02	0.24	0.31	0.5	ns
G124A	none	9.73	-0.03	Decreasing	0.12	0.19	0.17	0.5	ns
G24V	none	3.71	-0.02	Decreasing	0	NA	NA	NA	ns

Y28H	none	3.46	-1.66	Decreasing	0	NA	NA	NA	ns
Y28L	none	3.46	-1.76	Decreasing	0	NA	NA	NA	ns
V30A	none	3	-1.11	Decreasing	0	NA	NA	NA	ns
Q36H	none	3.24	-0.16	Decreasing	0	NA	NA	NA	ns
P40T	none	3.02	-0.11	Decreasing	0	NA	NA	NA	ns
L43P	none	3.87	-1.36	Decreasing	0	NA	NA	NA	ns
K46N	none	3.53	0.31	Increasing	0	NA	NA	NA	ns
L48R	none	4.73	-0.93	Decreasing	0	NA	NA	NA	ns
D56H	none	4.03	-0.19	Decreasing	0	NA	NA	NA	ns
Y64C	none	9.76	-0.28	Decreasing	0	NA	NA	NA	ns
I71F	none	9.82	0.28	Increasing	0	NA	NA	NA	ns
I71S	none	9.82	-0.22	Decreasing	0	NA	NA	NA	ns
W90C	none	3.3	-1.31	Decreasing	0	NA	NA	NA	ns
Y98N	none	9.62	-0.43	Decreasing	0	NA	NA	NA	ns
H116E	none	6.49	-0.27	Decreasing	0	NA	NA	NA	ns
H116L	none	6.49	0.03	Increasing	0	NA	NA	NA	ns
H116P	none	6.49	-0.11	Decreasing	0	NA	NA	NA	ns
M126S	none	8.23	0.06	Increasing	0	NA	NA	NA	ns
R128G	none	3.63	-1.31	Decreasing	0	NA	NA	NA	ns
R128L	none	3.63	-0.67	Decreasing	0	NA	NA	NA	ns
R128Q	none	3.63	-1.32	Decreasing	0	NA	NA	NA	ns
F129S	none	3.51	-1.64	Decreasing	0	NA	NA	NA	ns
P131L	none	3.03	-1.04	Decreasing	0	NA	NA	NA	ns
W149L	none	3.48	-2.01	Decreasing	0	NA	NA	NA	ns
K152E	none	4.4	-1.72	Decreasing	0	NA	NA	NA	ns
G156S	none	3.65	-1.09	Decreasing	0	NA	NA	NA	ns

K158N	none	4.96	-0.14	Decreasing	0	NA	NA	NA	ns
A162T	none	9.46	0.53	Increasing	0	NA	NA	NA	ns
L164R	none	8.88	-0.49	Decreasing	0	NA	NA	NA	ns
I165L	none	8.81	-0.33	Decreasing	0	NA	NA	NA	ns
I165Y	none	8.81	0.33	Increasing	0	NA	NA	NA	ns
E208K	none	2.66	-0.38	Decreasing	0	NA	NA	NA	ns
N218S	none	3.7	-0.09	Decreasing	0	NA	NA	NA	ns
R254C	none	9.68	-0.04	Decreasing	0	NA	NA	NA	ns
R254S	none	9.68	0.09	Increasing	0	NA	NA	NA	ns
M296T	none	3.22	-0.7	Decreasing	0	NA	NA	NA	ns
G297L	none	3.74	-0.9	Decreasing	0	NA	NA	NA	ns
N602D	none	6.84	0.12	Increasing	0	NA	NA	NA	ns
P605S	none	5.72	0.07	Increasing	0	NA	NA	NA	ns
Y608D	none	3.31	-1.9	Decreasing	0	NA	NA	NA	ns
L611R	none	3.52	-1.18	Decreasing	0	NA	NA	NA	ns
A614G	none	8.39	0.17	Increasing	0	NA	NA	NA	ns
S620T	none	3.43	0.64	Increasing	0	NA	NA	NA	ns
A621D	none	4.92	0.27	Increasing	0	NA	NA	NA	ns
M624V	none	8.44	-0.24	Decreasing	0	NA	NA	NA	ns
I666V	none	5.89	-0.28	Decreasing	0	NA	NA	NA	ns
W668L	none	3.26	-1.66	Decreasing	0	NA	NA	NA	ns
D675H	none	6.99	0.11	Increasing	0	NA	NA	NA	ns
K681T	none	9.36	-0.05	Decreasing	0	NA	NA	NA	ns
R693C	none	3.77	-0.95	Decreasing	0	NA	NA	NA	ns
F698V	none	6.39	-0.55	Decreasing	0	NA	NA	NA	ns
E703Q	none	3.05	1.27	Increasing	0	NA	NA	NA	ns

L704W	none	6.06	-0.4	Decreasing	0	NA	NA	NA	ns
R705L	none	3.04	-0.54	Decreasing	0	NA	NA	NA	ns
D714G	none	2.84	-1.14	Decreasing	0	NA	NA	NA	ns
F720S	none	8.53	-0.74	Decreasing	0	NA	NA	NA	ns
L43R	none	3.87	-1.29	Decreasing	0.01	NA	NA	NA	ns
W91G	none	7.83	-0.41	Decreasing	0.01	NA	NA	NA	ns
W91L	none	7.83	-0.45	Decreasing	0.01	NA	NA	NA	ns
H116A	none	6.49	0.02	Increasing	0.01	NA	NA	NA	ns
G124T	none	9.73	-0.35	Decreasing	0.01	NA	NA	NA	ns
G125S	none	8.84	-0.7	Decreasing	0.01	NA	NA	NA	ns
M126I	none	8.23	0.13	Increasing	0.01	NA	NA	NA	ns
P131A	none	3.03	-1.44	Decreasing	0.01	NA	NA	NA	ns
N133S	none	3.89	-0.89	Decreasing	0.01	NA	NA	NA	ns
K157Q	none	3.07	-0.11	Decreasing	0.01	NA	NA	NA	ns
K158S	none	4.96	-0.16	Decreasing	0.01	NA	NA	NA	ns
P286L	none	7.46	-0.11	Decreasing	0.01	NA	NA	NA	ns
P288H	none	5.64	-0.84	Decreasing	0.01	NA	NA	NA	ns
A290P	none	3.95	-0.23	Decreasing	0.01	NA	NA	NA	ns
W300G	none	8.4	-0.44	Decreasing	0.01	NA	NA	NA	ns
A614T	none	8.39	0.61	Increasing	0.01	NA	NA	NA	ns
F657S	none	9.29	-0.89	Decreasing	0.01	NA	NA	NA	ns
G680D	none	8.44	-0.39	Decreasing	0.01	NA	NA	NA	ns
K681Q	none	9.36	-0.07	Decreasing	0.01	NA	NA	NA	ns
G124S	none	9.73	-0.22	Decreasing	0.02	NA	NA	NA	ns
M624I	none	8.44	-0.22	Decreasing	0.02	NA	NA	NA	ns
H116Q	none	6.49	-0.3	Decreasing	0.03	NA	NA	NA	ns

P718S	none	7.41	-0.13	Decreasing	0.07	NA	NA	NA	ns
H116T	none	6.49	-0.08	Decreasing	0.08	NA	NA	NA	ns
K157R	none	3.07	-0.09	Decreasing	0.08	NA	NA	NA	ns
I165M	none	8.81	-0.37	Decreasing	0.09	NA	NA	NA	ns

Table 6.B.1: Mutations close to KatG PPI
Two-hundred and sixty single amino acid variation (SAV) mutations lying within 10Å of the Protein-Protein interface (PPI) and their corresponding PPI affinity changes ($\Delta\Delta$G) measured by mCSM-PPI2. The estimated effect are categorised as Destabilising ($\Delta\Delta$G<0) and Stabilising ($\Delta\Delta$G>0). The genomic measures of minor allele frequency (MAF), Odds Ratio, P-values, and FDR adjusted P-values are shown. Statistical significance indicated as: *P < 0.05, **P < 0.01, ***P < 0.001, ****P < 0.0001, ns: >0.05. The table is arranged by Odds Ratio to show mutation with the highest OR at the top for mutations at the PPI. Columns with NA indicate insufficient data to calculate Odds Ratio and P-values. Abbreviations used: $\Delta\Delta$G: change in Gibbs free energy in Kcal/mol, FDR: false discovery rate, ns: not significant, INH: isoniazid.

6.C Average stability comparisons for lineages

Lineage comparisons	Samples (n)	Adjusted P-values	*Adjusted P-values* Significance
L1 vs L2	L1 (3745), L2 (12809)	<0.0001	****
L1 vs L3	L1 (3745), L3 (4782)	<0.0001	
L1 vs L4	L1 (3745), L4 (5103)	<0.0001	
L2 vs L3	L2 (12809), L3 (4782)	<0.0001	
L2 vs L4	L2 (12809), L4 (5103)	<0.0001	
L3 vs L4	L3 (4782), L4 (5103)	<0.0001	
Within Lineage comparisons			
L1: R vs S	R (n=346), S (n=3399)	<0.0001	****
L2: R vs S	R (n=3705), S (n=9104)	<0.0001	
L3: R vs S	R (n=741), S (n=4041)	<0.0001	

Table 6.C.1: Lineage comparisons for KatG mutations
Kolmogorov-Smirnoff (KS) test reporting the statistical differences in distributions between *M. tuberculosis* lineages when assessed based on average stability changes ($\Delta\Delta$G) measured by mCSM-DUET, FoldX, DeepDDG, and Dynamut2. Lineage comparisons were performed for samples containing mutations associated with sensitivity (R: Resistant, S: Sensitive). These comparisons were performed for R and S samples between and within lineages. Statistical significance thresholds used are *P<0.05, **P<0.01, ***P<0.001, ****P<0.0001. Abbreviations used: $\Delta\Delta$G: change in Gibbs free energy in Kcal/mol, Adj. P-values: Bonferroni adjusted P-values, n=number of samples.

Chapter 7

Alr-cycloserine results

7.1 Background

7.1.1 Mechanism of action of cycloserine

Cycloserine, or more specifically D-cycloserine (DCS), is a bacteriostatic antibiotic used in the treatment of MDR TB. It is a cyclic analogue of D-alanine and is thought to competitively target at least two bacterial enzymes alanine racemase (*alr*) and D-alanine ligase (*Ddl*). DCS inhibits the action of these two crucial enzymes involved in the peptidoglycan synthesis an important component of any bacterial cell wall.[1] The *alr* gene codes for a pyridoxal 5'-phosphate-dependent enzyme involved in the conversion of L-alanine to D-alanine, which in turn serves as a substrate for *Ddl* which then joins two D-alanine together residues with a dipeptide bond.[2,3] DCS, by targeting these enzymes, prevents formation of the D-alanine residues, as well as the dipeptide derivative, thus removing components required for peptidoglycan biosynthesis.[2,3]

7.1.2 Cycloserine resistance in *M. tuberculosis*

The rate of emergence of DCS resistance has thus far remained low since its initial use more than half a century ago, due to a low mutation rate associated with *alr*.[4,5] DCS has been used in all regimens of MDR-TB treatment since 2018[6] where MDR-TB accounts for upwards of half a million cases per year. Overexpression and SAVs in *alr* have been reported to be associated with resistance.[5,7,8] Mutations in *ddl*,[3] as well as other targets are also involved in DCS resistance.[8,9] Despite the low rise of resistance, the burden of resistance-conferring mutations, however, remains high in *alr*, with potential compensatory mechanisms alleviating the associated high fitness costs.[5] SAVs in Alr: E373G, L113R,[10] as well as M343T, Y388D, R397L are implicated in DCS resistance, where SAVs M343T, and Y388D in particular have been associated with XDR *M. tuberculosis* strains.[11] It was noted that the equivalent positions in the Nakatani, *et. al.*, study are offset by -24.

More recently, it has been shown that DCS is a slow but reversible covalent inhibitor of Alr enzymes contrary to the prior understanding.[12] In slow growing bacteria like *M. tuberculosis*, DCS is unable to effectively inhibit Alr causing reactivation through a previously unrecognised pathway involving DCS-ring opening and subsequent reactivation of *M. tuberculosis* Alr.[12] An overview of the mechanism of action and resistance development in DCS is shown in **Figure 1**.

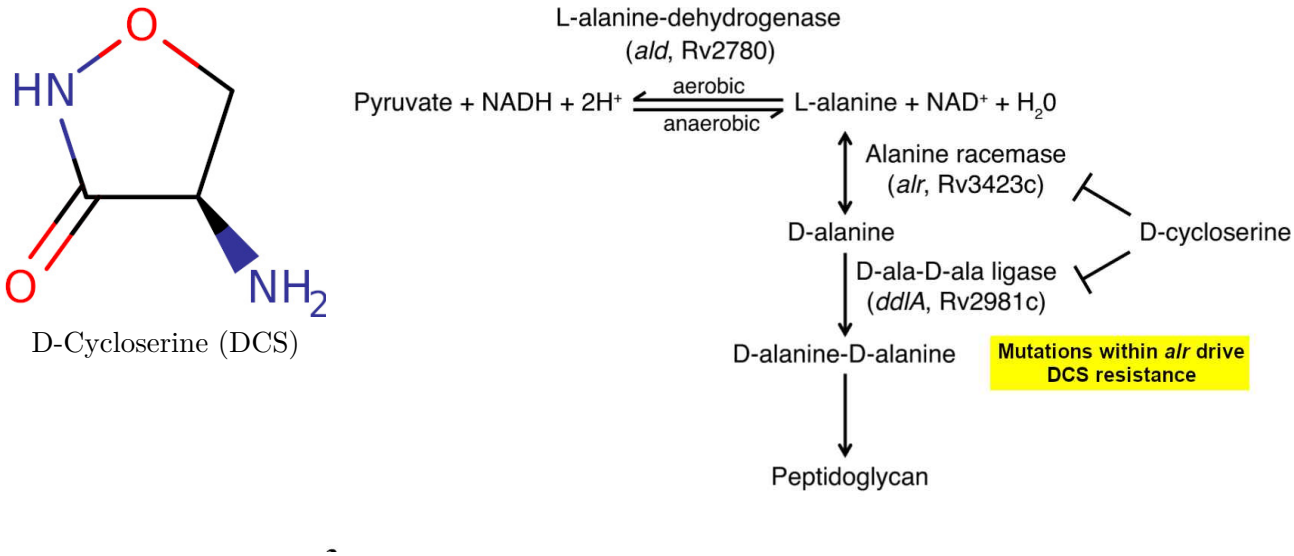


Figure 1: Chemical structure and mechanism of action and resistance for cycloserine
The chemical structure of cycloserine (DCS) appears at the top left and is sourced from DrugBank (ID:DB00260), along with its mechanism of action shown: L alanine is converted to D-alanine via alanine racemase *alr* where D-alanine molecules are joined by D-ala-D-ala ligase to produce the dipeptide D-alanine-D-alanine which is incorporated into peptidoglycan of the *M. tuberculosis* cell wall. It is also highlighted that mutations within *alr* are primarily responsible for driving DCS resistance. Figure adapted from Desjardins, *et. al.*.[13]

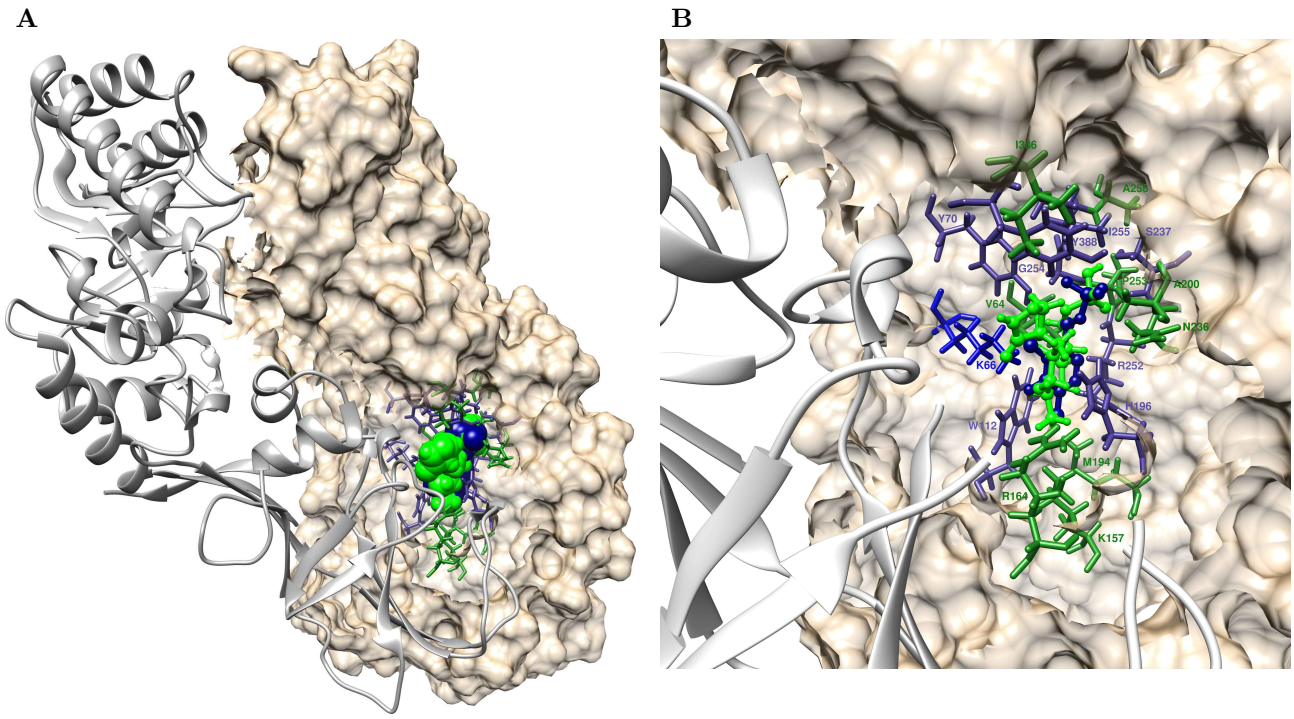


Figure 2: Active site description of *M. tuberculosis* Alr with DCS and PLP bound
Overall description of Alr-DCS complex. **A)** homo-dimer Alr in complex with DCS and co-factor PLP bound. Chain A of Alr is shown as surface representation in tan, while chain B appears as grey ribbons. DCS appears as green spheres, and co-factor PLP is shown as navy blue spheres, **B)** Close-up view of all interacting residues coloured green for DCS and blue for co-factor PLP, and labelled accordingly. The figure is generated using UCSF Chimera version 1.14. DCS: cycloserine, PLP: pyridoxal phosphate.

7.1.3 Description of the Alr-DCS complex

Alr is shown to form a native homo-dimer. DCS covalently binds to cofactor pyridoxal 5'-phosphate (PLP), disrupting the Alr-PLP covalent bond formation and thus inhibiting Alr activity.[14] Only recently has it been shown that this is not an entirely irreversible reaction in slow growing bacteria like *M. tuberculosis* where the slow growth rate is used as one of the key mechanisms to avoid antibiotic induced toxicity effects.[12]

Interactions of Alr

Molecular interactions with residues in *alr*, DCS, and co-factor PLP were identified using LigPlus, PLIP and Arpeggio, resulting in a total of twenty-seven interacting residues:

- Twenty three residues at sites 64, 66, 70, 112, 157, 164, 194, 196, 200, 236, 237, 252, 253, 254, 255, 256, 295, 314, 342, 343, 344, 386, 388 were identified to interact with DCS, with residues M343 and Y388 directly involved in the active site of Alr.
- Ten residues at sites 66, 70, 112, 196, 227, 237, 252, 254, 255, and 388 were identified to be interacting with co-factor PLP, with residue L66 forming an essential covalent bond with DCS.

An overview of the Alr homo-dimer structural complex with all interactions identified is described in **(Figure 2)**.

7.2 Structural and genomic insights into cycloserine resistance

7.2.1 Mutational landscape of Alr

Limited active site residues exhibited SAVs with mutations distributed across Alr

A total of 271 SAVs were found in the protein coding region of Alr (Genomic id: Rv3423c, coding region: 3840194-3841420), and appear to be widely distributed across the protein **(Figure 3)**, with mutations present in 186 unique positions for a maximum of 4 SAVs at any one site **(Figure 4)**. Not all active site residues were associated with SAVs. Of those that were, sites with single aa mutations were the most common.

Mapping mutations on Alr highlight **(Figure 4)** the following:

Sites with DCS interactions were associated with a maximum of 3 SAVs (sites marked in green)

- Single mutation: 157, 237, 253, 295, 314, 343, 344 with 237 sharing interaction with PLP
- Budding resistant hotspots: A256 and Y388 with Y388 sharing interaction with DCS

- Hotspots with three mutations: A200

Sites with PLP interactions associated with a maximum of 2 point mutations (sites marked in navy blue)

- Single mutation: 237 sharing interaction with DCS
- Budding resistant hotspots: Y388 sharing interaction with DCS

The majority (57%, n=155) of the mutational effects resulted in electrostatic changes.

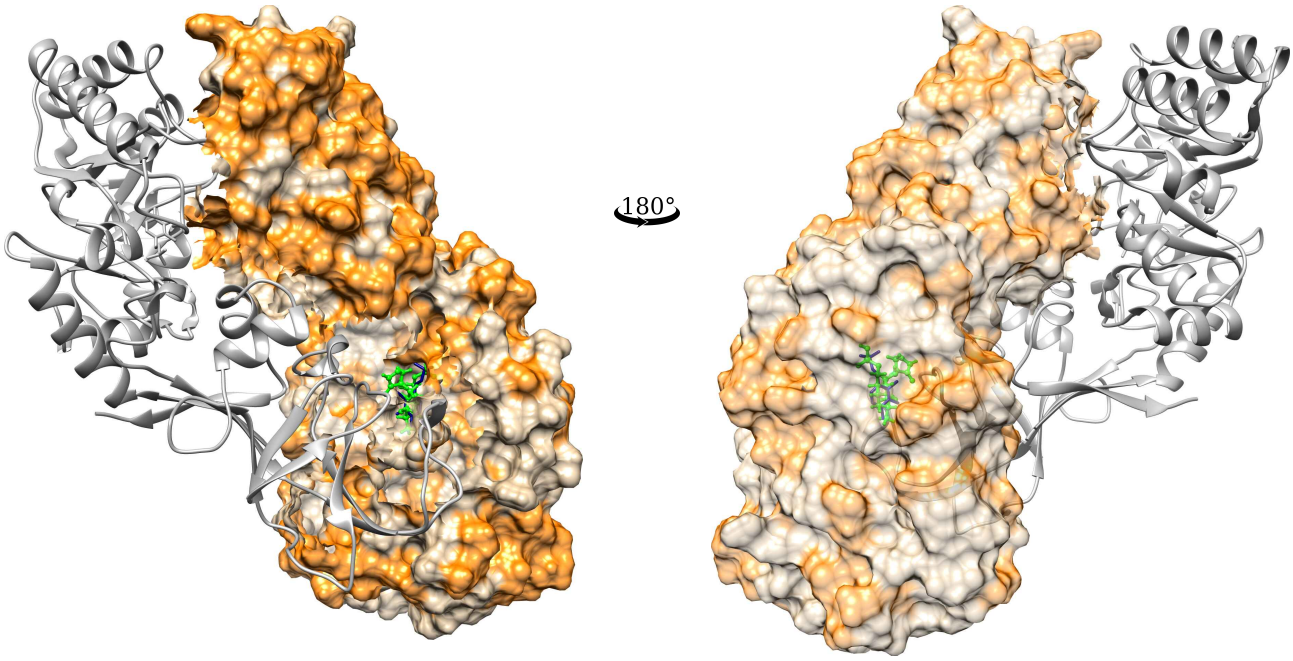


Figure 3: Mutational landscape of *M. tuberculosis* Alr
An overview of all mutational sites on *M. tuberculosis* Alr chain A appearing as surface representation in tan colour with chain B shown as grey ribbons. The left and right panels are opposing representations (rotated 180°) of Alr. The drug (DCS) is shown in green as ball-and-stick in the binding pocket and co-factor PLP is shown in navy blue sticks. The figure is generated using UCSF Chimera version 1.14. Abbreviations used: SAV: single amino acid variation, DCS: cycloserine, PLP: pyridoxal phosphate.

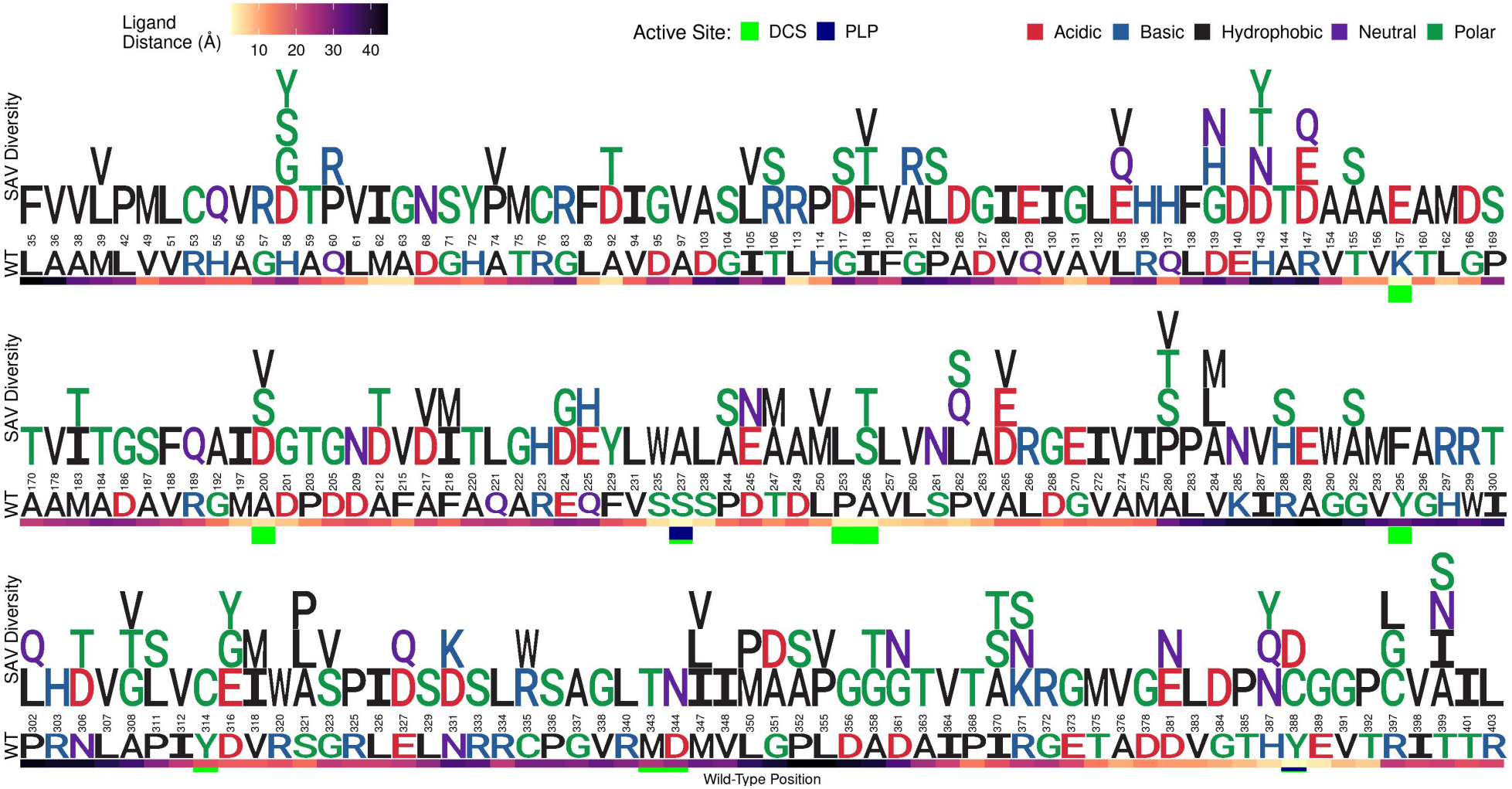


Figure 4: Sites associated with SAVs in *M. tuberculosis* Alr

Logo plot showing 188 unique sites/positions associated with 271 SAVs in the *M. tuberculosis* Alr. The horizontal axis shows the wild-type positions associated with SAVs in Alr and the vertical axis shows all the mutant residues observed in our data highlighting SAV diversity at a given site. Residues are coloured according to the amino acid (aa) property, where acidic aa appear in red, basic aa appear in blue, hydrophobic aa in black, neutral aa in purple, and polar aa in dark green. The structural positions associated with SAVs in Alr are indicated on the horizontal axis. The wild-type (WT) residues also coloured according to aa property appear under the respective position markings. The heat bar underneath the WT residues indicate the distance of that position from DCS according to the magma colour gradient, where light yellow indicates sites closer to DCS (ligand distance in Angstroms). The positions are further annotated to reflect active site residues involved in interactions with DCS (green), and co-factor PLP (navy blue). The figure is generated using R statistical software version 4.0.2, ggplot2 package. Abbreviations used: SAV: single amino acid variation, DCS: cycloserine, PLP: pyridoxal phosphate.

7.2.2 Mutational outcome from protomer stability changes and evolutionary conservation

Mutations were destabilising for overall protomer stability without affecting protein function

Over 75% of mutations had a destabilising effect on overall protomer stability when measured by the different computational tools **(Figure 5A-D)**, with DeepDDG estimating ~87% (n=235) mutations as destabilising, followed by Dynamut2 predicting ~82% (n=223) mutations as destabilising, and both mCSM-DUET and FoldX estimating 78% (n=213) mutations as destabilising. Based on evolutionary conservation, over 50% of mutations were predicted to have a non-deleterious impact (effect) on protein function indicated by PROVEAN and SNAP2 scores, where SNAP2 predicted a higher number (n=187) of SAVs compared with PROVEAN (n=143) with neutral effect **(Figure 5E** and **Figure 5F)**.

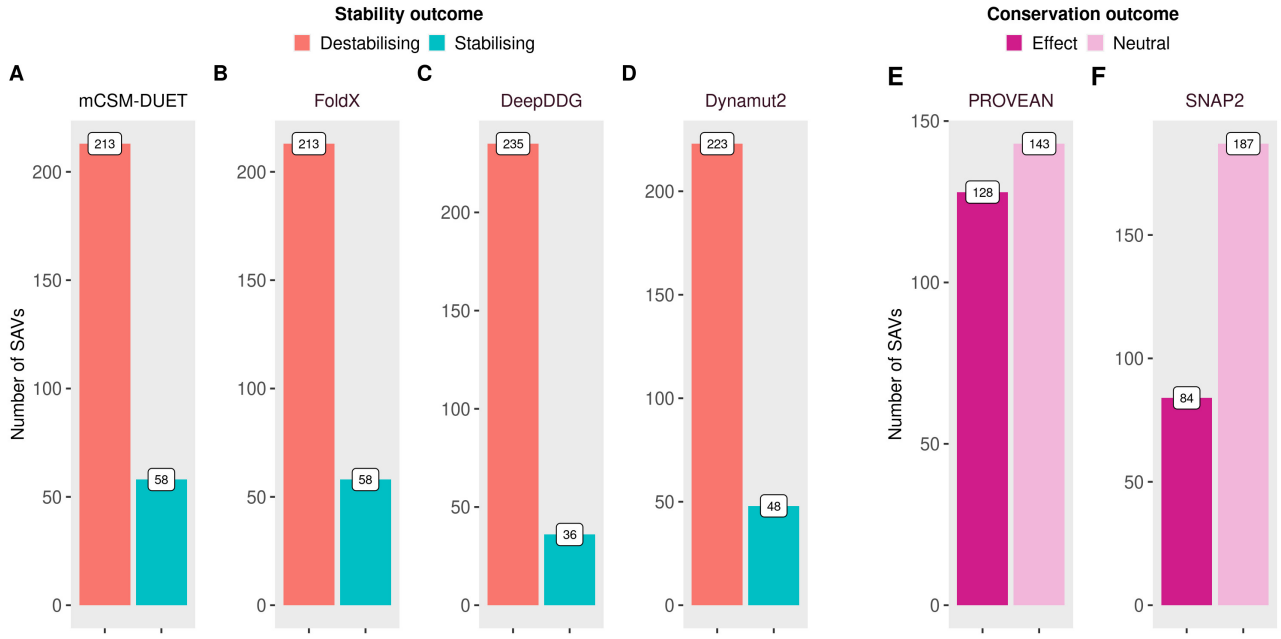


Figure 5: Protein stability outcome of SAVs in *M. tuberculosis* Alr
Mutational impact of SAVs on overall protein stability and evolutionary conservation changes for 271 SAVs, **A-D)** Barplots showing number of SAVs categorised as destabilising (red) or stabilising (blue) according to protein stability changes (ΔΔG Kcal/mol) as measured by four computational tools: mCSM-DUET, FoldX, DeepDDG, and Dynamut2, **E-F)** Number of SAVs categorised as Effect/Deleterious (magenta) or Neutral (pink) according to evolutionary conservation changes as estimated by computational tools: PROVEAN and SNAP2. The figures are generated using R statistical software version 4.0.2, ggplot2 package. Abbreviations used: ΔΔG: change in Gibbs free energy, SAV: single amino acid variation.

Evolutionary and structure-based predictors provide different insights into understanding mutational impact. Mutational impact in this context is considered to be its effect on protein stability, drug binding affinity, other binding affinities such as PPI or nucleic acid, and functional effects arising from protein sequence variations. The first three mutational consequences are assessed by structure based predictors relying on the 3D structure of a protein, while the last is assessed by sequence based

predictors relying mainly on evolutionary conservation trends across many proteins using multiple sequence alignments. The sequence based predictors are aimed at predicting pathogenicity or change of molecular function, structure based tools rely on estimating variant effects in relation to structure damage, corresponding to stability changes, as protein stability is considered the basic characteristic affecting function, activity, and regulation. Predictors such as ConSurf are able to use both structural and sequence information to identify important functional regions conserved in proteins. A variant classified as 'deleterious' to protein conservation may display gain-of-function in the presence of a drug through optimised protein stability. Thus, different methodological strategies benefit from complementary information when assessing specific proteins.

Active site residues were all destabilising for protomer stability except S237

When assessing the impact of mutations on protomer stability changes due to mutations, the estimates from all four tools: mCSM-DUET, FoldX, DeepDDG, and Dynamut2 were considered together and averaged to gain an understanding of the consensus mutational effect **(Figure 6)**. All active site residues associated with SAVs were destabilising except DCS (and PLP) interacting residue S237 with a single mutation (S237A) resulting in a stabilising effect **(Figure 7)**. Sites 58, 143, 321, 381 with multiple SAVs were stabilising for overall protomer stability.

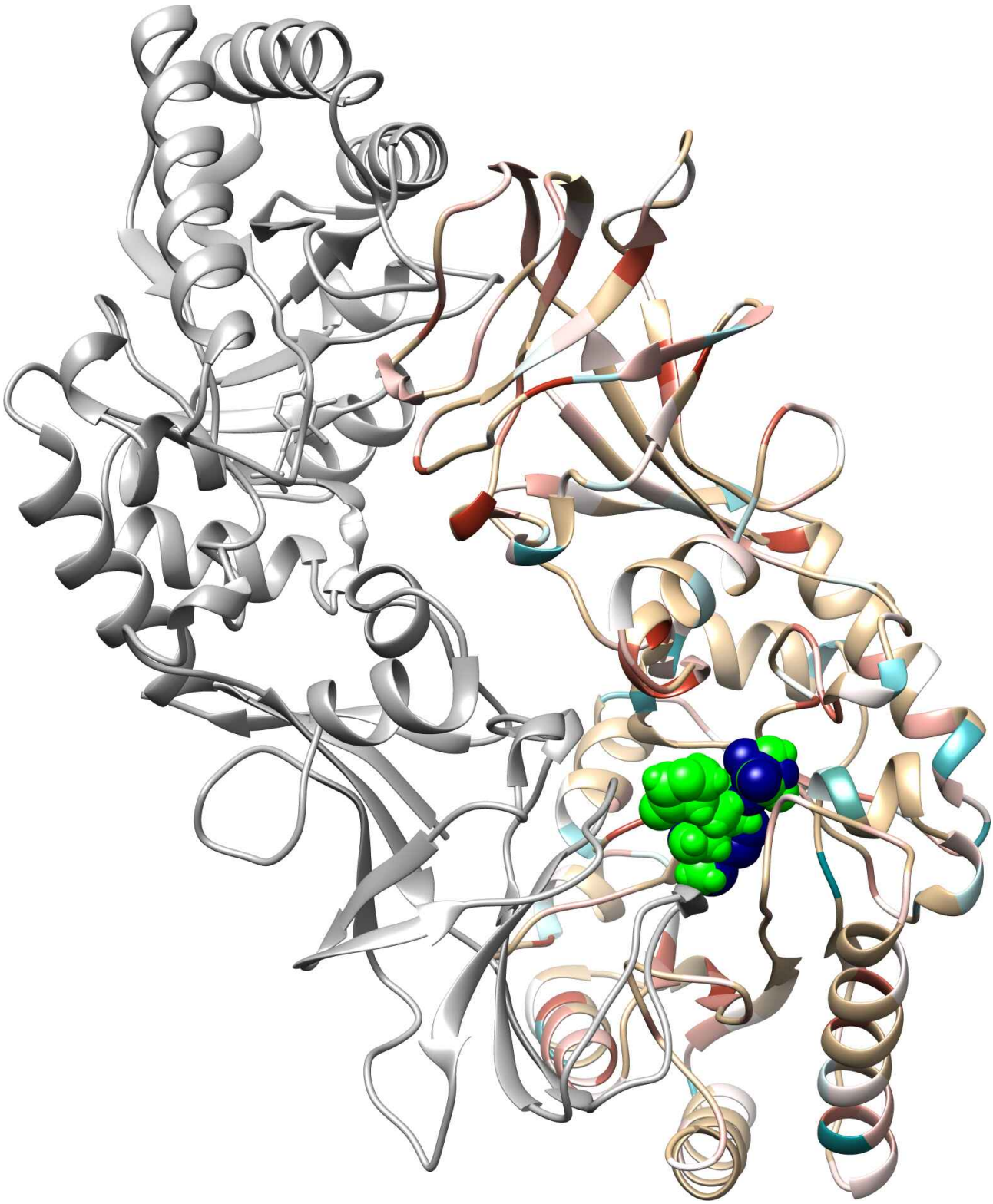


Figure 6: Average protein stability effects of SAVs mapped onto the *M. tuberculosis* Alr protein structure

The protein stability changes (ΔΔG Kcal/mol) of SAV mutations measured by mCSM-DUET, FoldX, DeepDDG, and Dynamut2 were averaged and mapped onto Alr positions with mutations. Destabilising mutational sites are depicted in red while stabilising mutational sites appear in blue where the colour intensity reflects the extent of the effect, ranging from -1 (most destabilising) to +1 (most stabilising). DCS is shown as green spheres in the binding pocket, with co-factor PLP shown in navy blue spheres. The figure is rendered using UCSF Chimera version 1.14. Abbreviations used: ΔΔG: change in Gibbs free energy in Kcal/mol, SAV: single amino acid variation, PLP: pyridoxal phosphate, DCS: cycloserine.

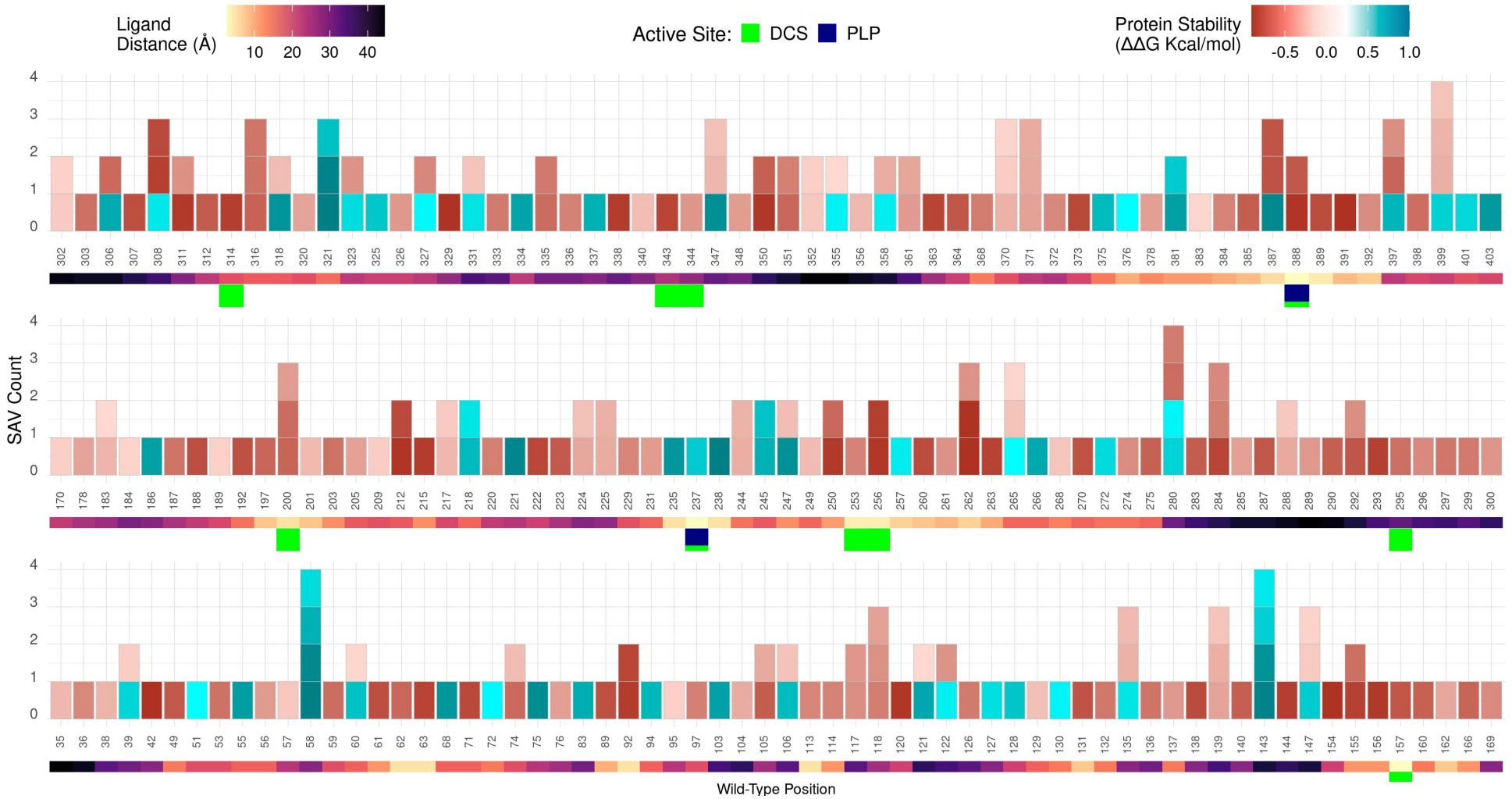


Figure 7: Average protein stability effect for individual SAVs occurring in *M. tuberculosis alr*

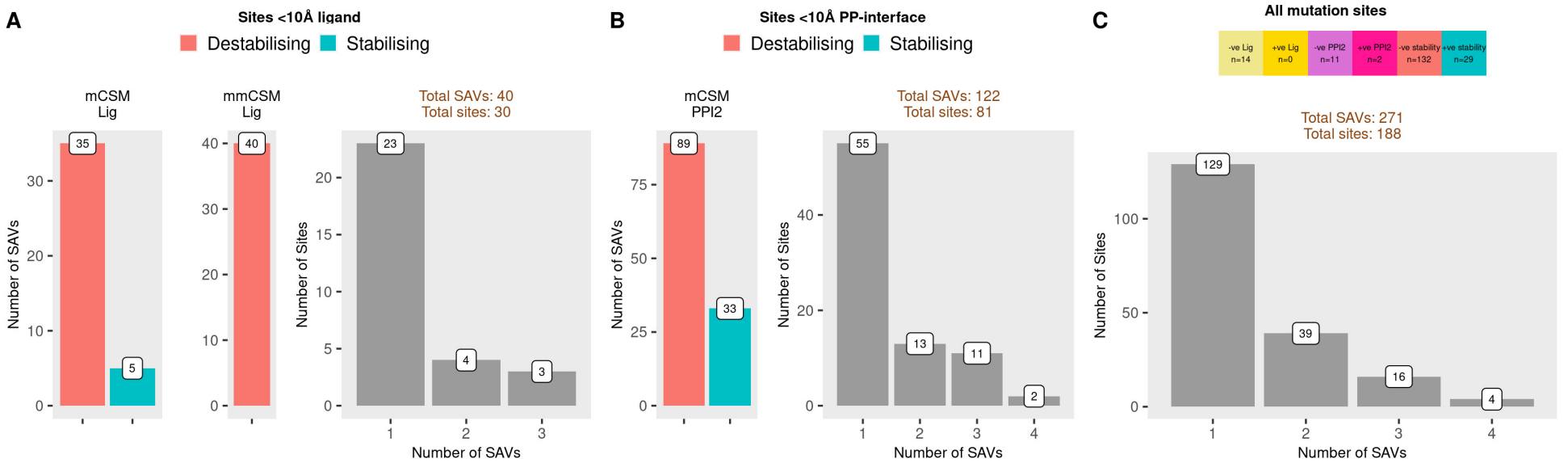
Barplot showing the number of single amino acid variation (SAV) mutation at each position in Alr coloured by the average protein stability effect, where the horizontal axis shows the wild-type positions associated with SAVs, and the vertical axis shows the number of SAVs at that position. For a given position, each corresponding SAV is coloured by the average protein stability effect calculated across estimates (ΔΔG Kcal/mol) from mCSM-DUET, FoldX, DeepDDG, and Dynamut2. The structural positions associated with SAVs in Alr are indicated on the horizontal axis. The heat bar underneath the positions indicates the distance of that position from DCS according to the magma colour gradient where light yellow indicates sites closer to DCS (ligand distance in Angstroms). The positions are further annotated to reflect active site residues involved in interactions with DCS in green, and co-factor PLP in navy blue. The barplot figures are generated using R statistical software version 4.0.2, ggplot2 package. Abbreviations used: ΔΔG: change in Gibbs free energy in Kcal/mol, SAV: single amino acid variation, PLP: pyridoxal phosphate, DCS: cycloserine.

7.2.3 Mutation consequences on affinity changes and prominent mutational effects

Mutations decrease binding affinity of DCS as well as the dimer interface

When considering SAV induced DCS binding affinity changes for sites within 10Å of DCS, 15% (n=40) of mutations were identified. These mutations occurred at 30 distinct sites, with most (n=23) sites showing single mutations. Of these, nearly 88% (n=35) had a destabilising effect on DCS binding affinity as measured by mCSM-lig and all 40 mutations were destabilising when measured by mmCSM-lig **(Figure 8A** top panel, Appendix Table 7.A.1**)**. The 30 mutational sites with their average affinity outcome impact were mapped onto the Alr (chain A) which showed mild-to-moderate destabilising mutational consequences **(Figure 8A** bottom panel**)**. Inspecting the dimer interface of Alr highlighted 45% (n=122) of mutations to be within 10Å of the PPI as estimated by mCSM-PPI2, with 73% (n=89) of mutational consequences being destabilising **(Figure 8B** top panel, Appendix Table 7.B.1**)**. The mutations at the dimer interface showed mild-to-moderate destabilising effects on visual inspection **(Figure 8B** bottom panel**)**.

Of the total 188 unique sites in Alr displaying SAVs, only 10% of sites (n=20) had multiple SAVs, 39 sites were considered budding resistant hotspots, with the majority (n=129) of sites exhibiting single mutations **(Figure 8C** top panel**)**. The most prominent effects on DCS binding were from reduced affinity (destabilising effect) to DCS from mutations at 14 surrounding sites **(Figure 8C**, yellow text boxes, and bottom panel**)**. Similarly, the dimer surface of Alr was chiefly affected by destabilising mutations, which reduced affinity for the second Alr protomer from mutations at 11 surrounding sites **(Figure 8C**, pink text boxes, and bottom panel**)**. All other sites were largely (n=132) affected by destabilising mutations **(Figure 8C**, blue and red text boxes, and bottom panel**)**.


A
Sites <10Å ligand
Destabilising
Stabilising
mCSM Lig
mmCSM Lig
Total SAVs: 40
Total sites: 30
Number of SAVs
Number of Sites
35
5
40
23
4
3
B
Sites <10Å PP-interface
Destabilising
Stabilising
mCSM PPI2
Total SAVs: 122
Total sites: 81
89
33
55
13
11
2
C
All mutation sites
-ve Lig n=14
+ve Lig n=0
-ve PPI2 n=11
+ve PPI2 n=2
-ve stability n=132
+ve stability n=29
Total SAVs: 271
Total sites: 188
129
39
16
4

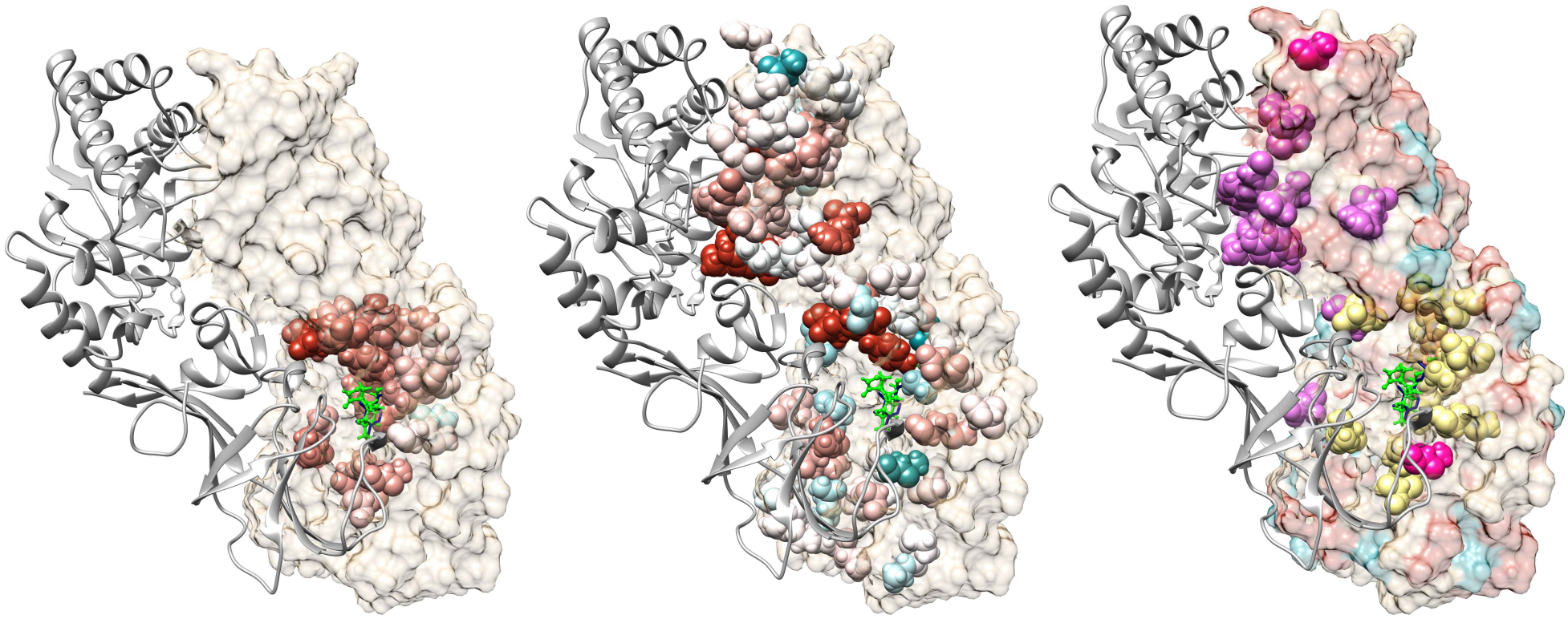


Figure 8: Mutational impact on DCS binding affinity, protein-protein interaction on Alr and sites with the most prominent mutational effects within *M. tuberculosis* Alr
The top panel displays barplots showing the mutational outcome of affinity changes and their corresponding site frequency, while the bottom panel shows the corresponding mutational impact mapped onto the Alr (chain A appearing in tan colour, while chain B is shown as grey ribbons. DCS is shown in green as ball-and-stick in the binding site, while co-factor PLP appears in navy blue. **A)** Mutational impact on DCS binding affinity (log fold change) from mCSM-lig and mmCSM-lig where 40 mutations, corresponding to 30 sites within 10Å of DCS, **B)** Mutational impact on protein-protein (PP) binding affinity ($\Delta\Delta$G) for 122 mutations, corresponding to 81 sites within 10Å of the PPI. For both parts A) and B), red denotes destabilising mutational sites while blue denotes stabilising mutational sites, and the colour intensity reflects the extent of the effect ranging from -1 (most destabilising) to +1 (most stabilising), **C)** Most prominent mutational effect for all 271 SAVs at 188 sites, prioritised in order of increasing effect size: mCSM/mmCSM-lig, mCSM-NA, protomer stability changes. Mutational effects are coloured according to the effect type with brighter colours indicate stabilising effects. Sites marked in yellow indicate changes due to ligand (DCS) binding affinity with light yellow indicating destabilising effect, pink areas indicate changes due to PPI affinity with bright pink highlighting stabilising and light pink areas indicating destabilising mutational effects. Protomer stability changes are coloured with blue indicating stabilising and red indicating destabilising mutational consequences. The corresponding number of mutation sites contributing to the different effect types are indicated in the text box at the top, and coloured accordingly. The barplot figures are generated using R statistical software version 4.0.4, ggplot2 package. The structure figures are generated using Chimera version 1.14. Abbreviations used: Å: Angstroms, $\Delta\Delta$G: Change in Gibbs free energy in Kcal/mol, SAV: single amino acid variation, DCS: cycloserine, PLP: pyridoxal phosphate.

7.2.4 Mutational association with DCS resistance and flexibility

Most mutations occur in highly variable regions of Alr, and are not associated with high flexibility

Mutational association with resistance according to aggregate DST data showed only 2 mutations as Resistant. These were L113R and M343T (**Figure 9A)**, where L113R occurred at 7.35Å of DCS and M343T occurred >10Å of DCS, but close to the dimer interface at 3.73Å (Appendix Tables 7.A.1 and 7.B.1).

DCS and PLP are located in the conserved regions of Alr **(Figure 9B1)**. Sites close to DCS as well as the dimer surface are moderate-to-highly conserved **(Figure 9B1)**. ConSurf scores are calculated for each site on the protein, and range from 1 (rapidly evolving, variable sites) to 9 (slowly evolving, conserved sites). Most mutations (n=54) occurred in the most variable region of Alr (ConSurf score 1) (**Figure 9B2)**, but the two resistant mutations were in moderate-to-highly conserved regions of Alr **(Figure 9B1)**.

The local flexibility in Alr in relation to DCS resistance was also analysed with thickness of the ribbon/tube (thinthick) corresponding to the extent of flexibility. Normal mode analysis of the Alr protein highlighted that residues interacting with DCS and PLP were not associated with high flexibility **(Figure 10A1)**, where regions associated with high flexibility were not associated with any SAVs **(Figure 10A2)**.

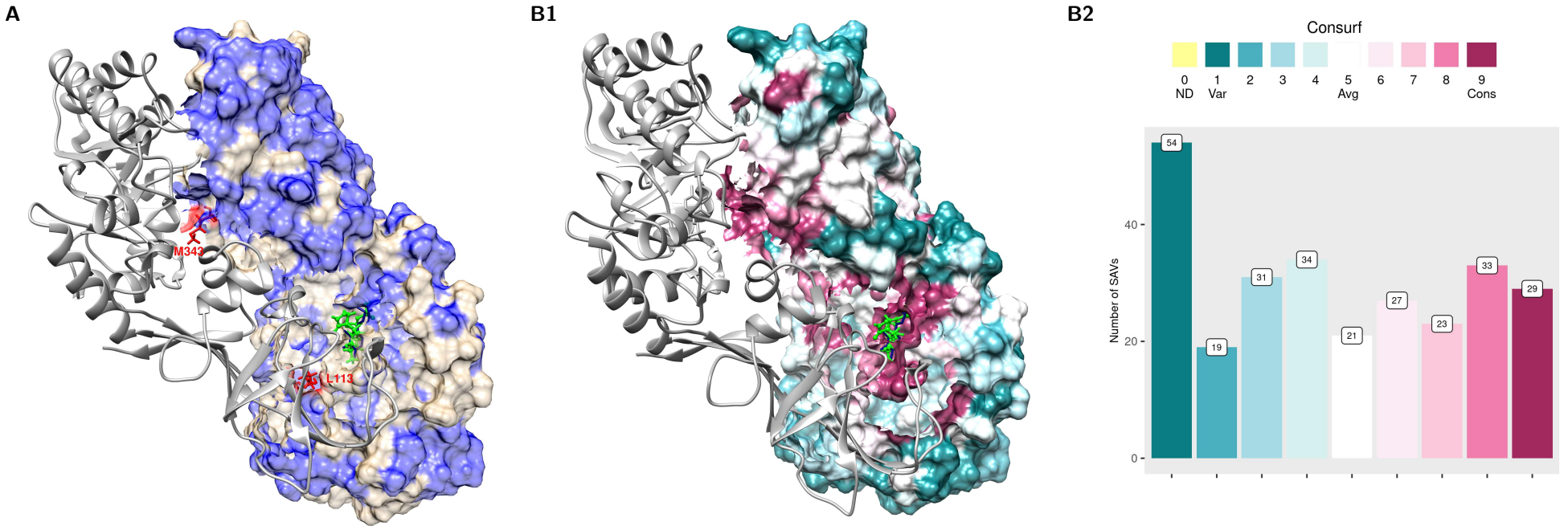


Figure 9: Mutational association with cycloserine resistance and evolutionary conservation in *M. tuberculosis* Alr

Mutational landscape of *M. tuberculosis* Alr according to different measures where **A)** All sites associated with SAVs in Alr chain A (surface representation in tan colour), along with chain B appearing as grey ribbons. DCS is represented as green ball-and-stick in the binding pocket and co-factor PLP is shown in navy blue sticks. Sites are coloured according to association with resistance for one or more SAVs, where red denotes sites with exclusively resistant mutations (n=2), and blue indicates sites with exclusively sensitive mutations (n=186). There were no sites with both sensitive and resistant mutations, **B1)** Alr chain A coloured according to ConSurf Scores where maroon indicates conserved sites and teal indicates variable sites. DCS appears as green ball-and-stick in the binding pocket and co-factor PLP is shown in navy blue sticks, B2) shows the number of mutations associated with ConSurf values that range from 1 (variable) in teal to 9 (conserved) in maroon, where 0 denotes insufficient data/not defined (ND). The barplot figure is generated using R statistical software version 4.0.4, ggplot2 package. The structure figures were generated using UCSF Chimera version 1.14. Abbreviations used: SAV: single amino acid variation, DCS: cycloserine, PLP: pyridoxal phosphate, DCS: cycloserine.

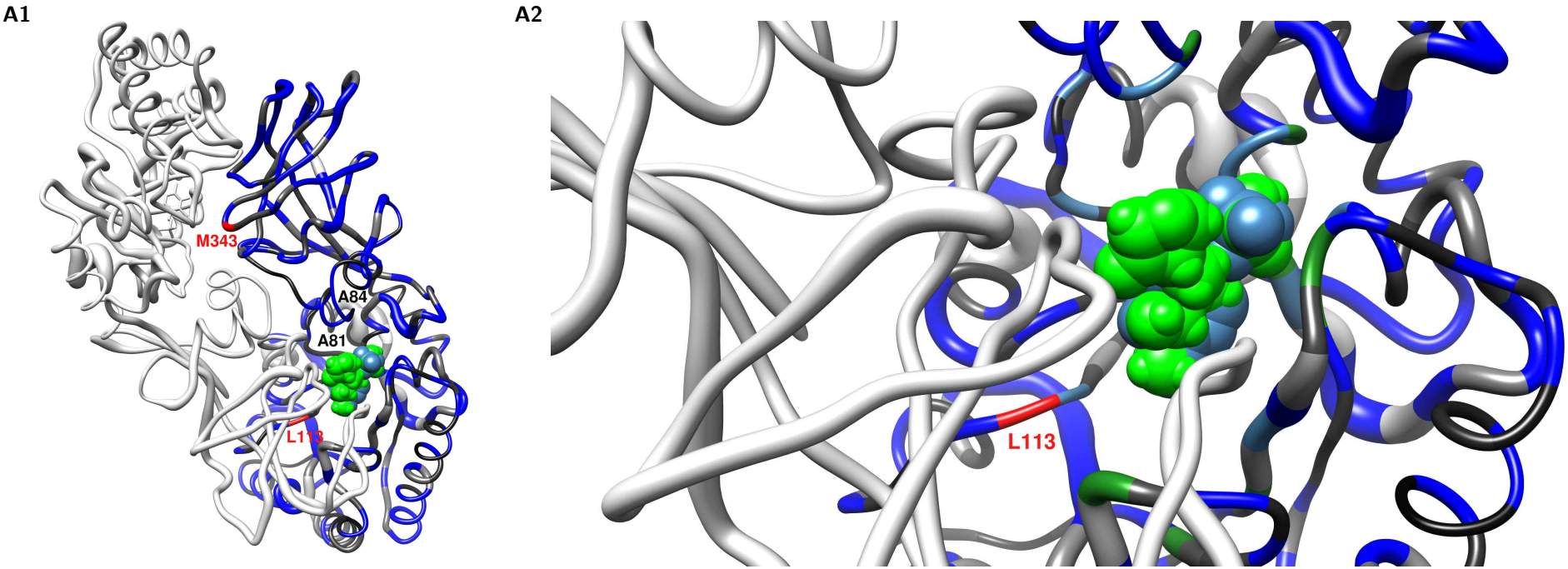


Figure 10: Mutational association with cycloserine resistance and local protein flexibility of *M. tuberculosis* Alr
Mutational landscape of *M. tuberculosis* Alr according to flexibility in Alr according to normal mode analysis (NMA), measuring atomic deformation according to protein dynamics to denote flexibility associated at sites in Alr. The magnitude of the flexibility is represented from thin (low flexibility) to thick (high flexibility) tubes coloured to show mutational association with resistance, red: resistant sites, blue: sensitive sites, black: sites with no SAVs. Resistant residues (using standard one-letter amino acid code) are indicated in red, and two others which are associated with moderate to high flexibility appear in black to denote these sites were not associated with SAVs in our data. **A1)** Overview of the Alr homo-dimer protein according to NMA flexibility, with chain A shown as tubes coloured as described above, and chain B appearing as grey coloured tubes, **A2)** Close-up view to show that highly resistant mutations L113R is a region of low flexibility, while areas coloured green (interacting with DCS) and steel blue (interacting with PLP) are also regions of low flexibility. DCS is denoted as green spheres. The structure figures were generated using UCSF Chimera version 1.14. Abbreviations used: SAV: single amino acid variation, DCS: cycloserine, PLP: pyridoxal phosphate, DCS: cycloserine.

7.2.5 Relating mutational frequency and biophysical and evolutionary conservational changes

Correlation analysis was performed to understand the relationship between frequently occurring mutations as assessed by MAF and their association stability (mCSM-DUET, FoldX, DeepDDG, Dynamut2), conservation (ConSurf, SNAP2, PROVEAN), affinity changes (mCSM-lig/mmCSM-lig, and mCSM-PPI2), distance to ligand (Lig-Dist), and protein-protein interface (PPI-Dist). Mutations were not separated into resistant and sensitive mutations due to there being only two resistant mutations **(Figures 11** and **12)**. Analyses focused on determining the strength of association without regard for the direction of the association due to dissimilarity of threshold criteria used by the various estimators.

Frequently occurring mutations were weakly related to stability and evolutionary conservation and not related to affinity changes

Mutational frequency was overall weakly associated with only DeepDDG stability changes (ρ=0.15, P<0.05). All other measures of stability changes (mCSM-DUET, FoldX and Dynamut2), as well as distance to ligand and dimer interface, were not associated (ρ=0.0, P>0.05) **(Figure 11)**.

Similarly, mutational frequency was very weakly associated with PROVEAN estimates (ρ=0.12, P<0.05) **(Figure 12** left panel**)**, while affinity changes (mCSM-/mmCSM-lig and mCSM-PPI2) were not found to be associated with mutational frequency **(Figure 12** right panel**)**.

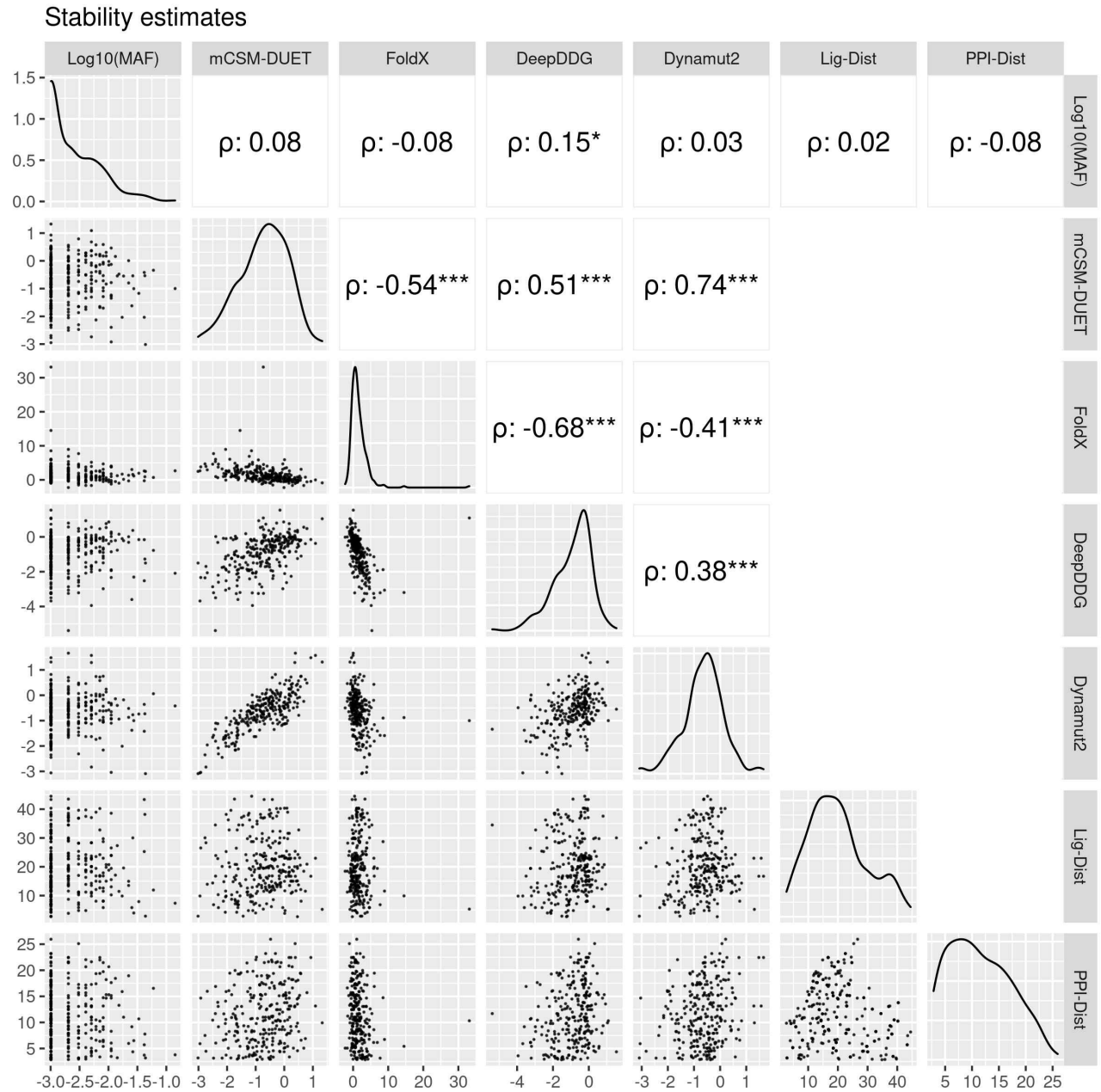


Figure 11: Correlation of protein stability changes and genomics measures
Pairwise correlation between minor allele frequency (MAF), protein stability changes ($\Delta\Delta$G) estimated using DUET, FoldX, DeepDDG, and Dynamut2, and distance to DCS, and the dimer interface for 271 SAVs. The upper panel in both plots includes the pairwise Spearman (ρ) correlation values along with their statistical significance (.P<0.10, *P<0.05, **P<0.01, ***P<0.001). The diagonal in each plot displays the density distribution of the corresponding parameter. The figure is generated using R statistical software version 4.0.4, ggplot2 package. Abbreviations used: Å: Angstroms, $\Delta\Delta$G: change in Gibbs free energy in Kcal/mol, SAV: single amino acid variation, Lig-Dist: distance to ligand in Å, PPI-Dist: distance to protein-protein interface in Å, DCS: cycloserine.

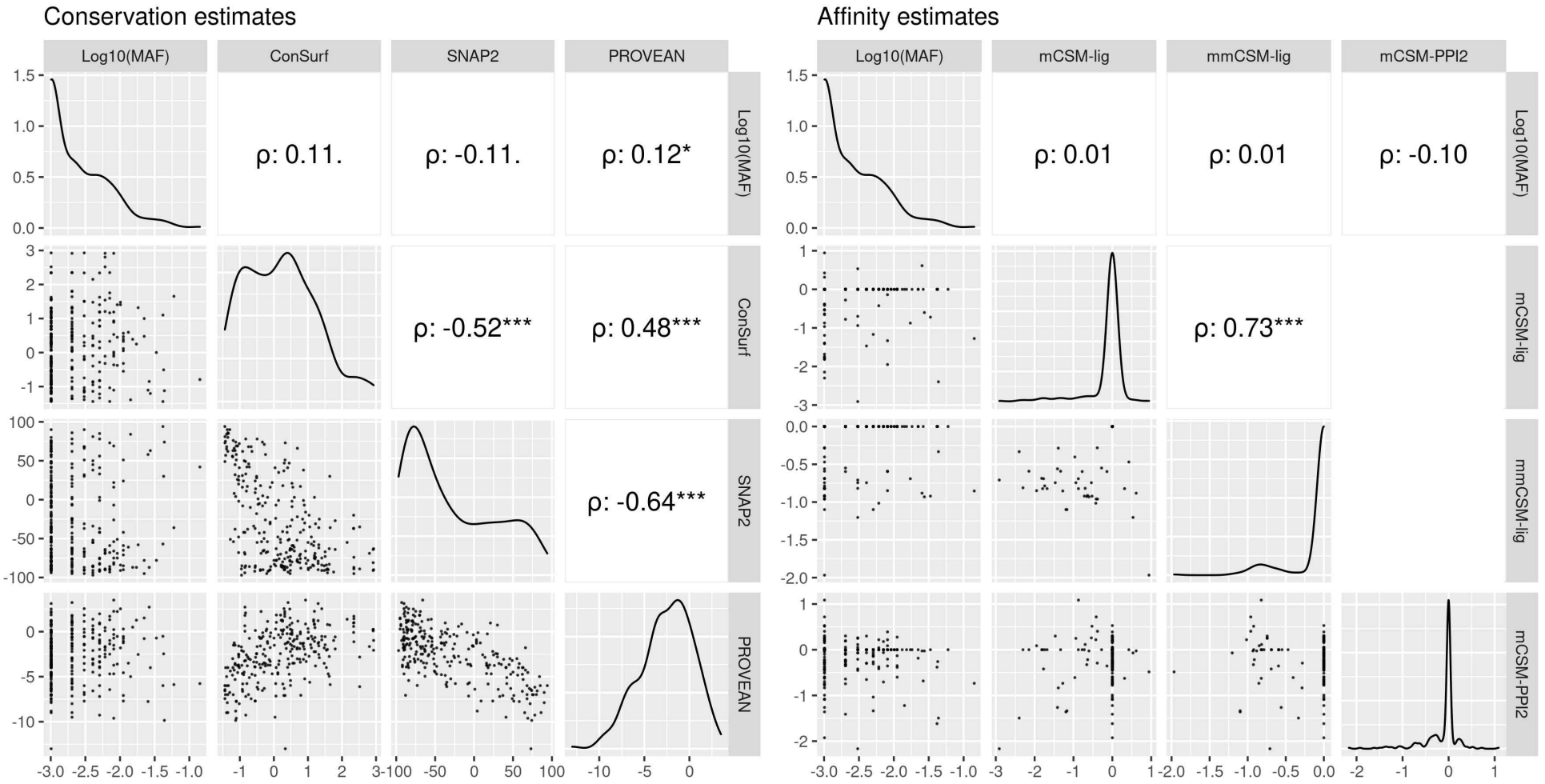


Figure 12: Correlation of evolutionary conservation, affinity changes, and genomics measures
Pairwise correlations of evolutionary conservation, affinity changes, and genomic measure of minor allele frequency (MAF) for 271 SAVs. **Left panel:** Evolutionary conservation predictors: ConSurf, SNAP2, and PROVEAN, **Right panel:** DCS binding affinity changes, estimated as log fold change (mCSM-lig) of 40 SAVs lying within 10Å of DCS, and protein-protein (PP) affinity changes ($\Delta\Delta$G) measured using mCSM-PPI2 of 122 SAVs lying within 10Å of the dimer interface. All corresponding affinity measures for mutations located more than 10Å of DCS, and the dimer interface were given a value of 0 to allow all identified SAVs to be used for analysis, while respecting the distance threshold for the respective tools. The upper panel in both plots includes the pairwise Spearman (ρ) correlation values along with their statistical significance (.P<0.10, *P<0.05, **P<0.01, ***P<0.001). The diagonal in each plot displays the density distribution of the corresponding parameter. The figure is generated using R statistical software version 4.0.4, ggplot2 package. Abbreviations used: Å: Angstroms, $\Delta\Delta$G: change in Gibbs free energy in Kcal/mol, SAV: single amino acid variation, Lig-Dist: distance to ligand in Å, PPI-Dist: distance to protein-protein interface in Å, DCS: cycloserine.

7.2.6 Associating mutations with Odds Ratio and extreme effects

L113R, not an active site residue was highly associated with DCS resistance

Based on DST data available for only 15 (out of 271) SAVs, mutational association with resistance was further estimated using Odds Ratio (OR), with values above 1 suggesting association with DCS resistance. The higher the OR, the greater the likelihood of a given mutation being resistant. All 15 mutations were predicted to be associated with DCS resistance **(Figure 13)**, compared with the 2 mutations according to the available DST data. Mutation L113R, one of the two resistant mutations according to DST, was the most frequently occurring (MAF=14%) **(Table 1)**, and highly associated with DCS resistance (OR=51.87) **(Figure 13)**. Further, it was also the single mutation at the site suggesting selective pressure from the drug. This was followed by residue Y388D involved with DCS and PLP binding showing the next highest association with DCS resistance (OR=16.50), followed thereon by mutations: A358T, L283P, H297R, N331D, T399I, S321L, E140D, L138F, A200V (OR=4.08), M343T (OR=1.91), S261N (OR=1.64), D139H (OR=1.36). Of these A200V and M343T are active site residues **(Figure 13)**.

Mutations with extreme effects did not include active site residues

As mentioned above, L113R was the most frequently occurring and highly associated with DCS resistance (OR=51.87) **(Figure 13, Table 1)**. Mutation E389G was the most destabilising for DCS, as well as for dimer interface binding affinity. The most stabilising mutation for the dimer interface was mutation L162M. Mutations V391G and S238L were the most destabilising and stabilising mutations respectively for overall protomer stability. None of these residues was involved in DCS and/or PLP interactions **(Table 1)**.

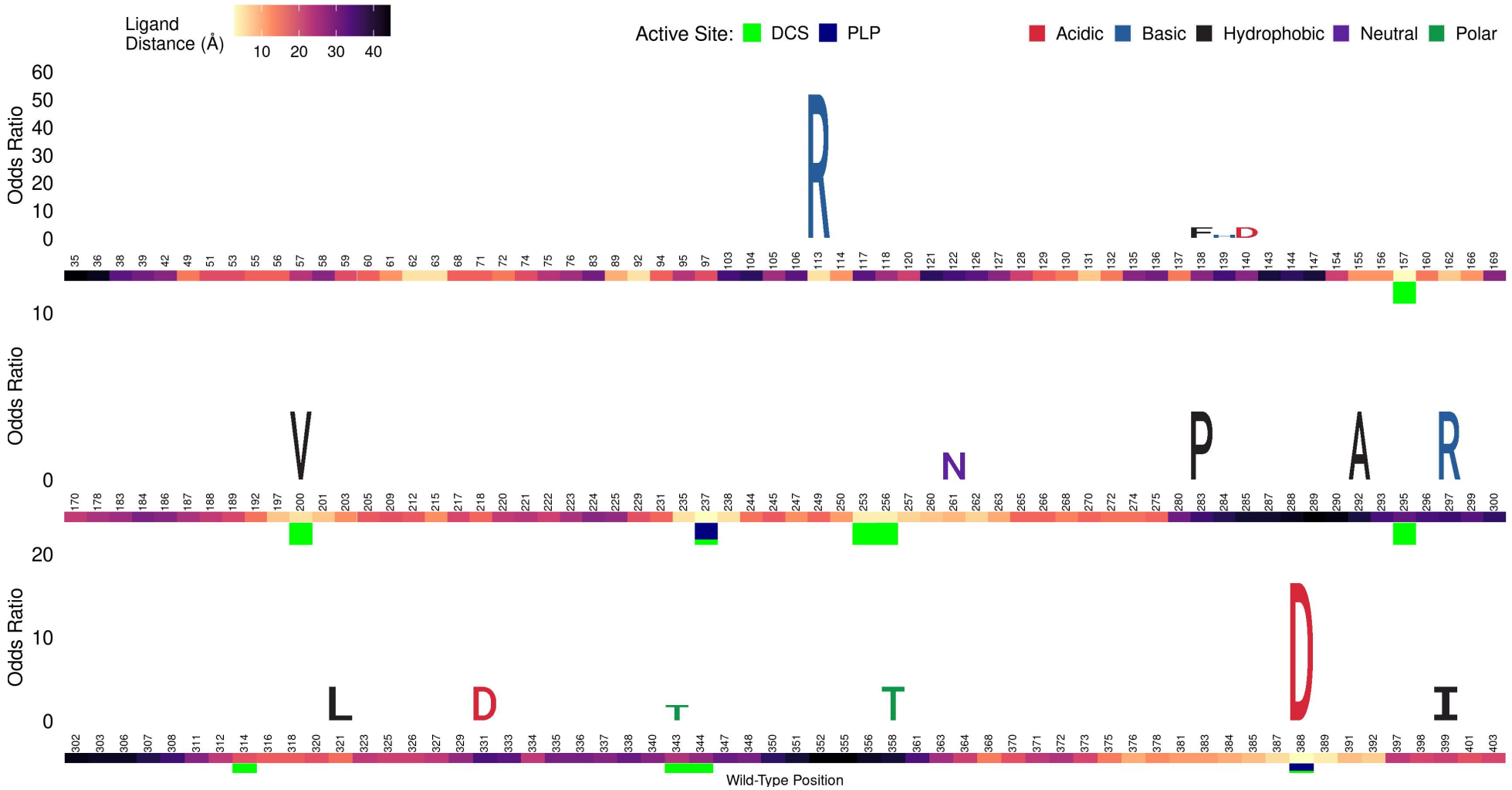


Figure 13: Logo plot showing mutational sites and their association with resistance according to Odds Ratio

Logo plot showing 15 SAVs by mutational site according to their association with DCS resistance calculated using Odds Ratio (OR). The vertical axis represents the OR, where letters denote mutant residues which are proportional to their corresponding OR, highlighting the most resistant mutation at each site and overall. The mutant residues are coloured according to the amino acid (aa) properties as denoted where red denotes acidic aa, basic aa appear in blue, hydrophobic aa in black, neutral aa in purple, and polar aa in dark green. The structural positions associated with SAVs with OR are indicated on the horizontal axis. The heat bar underneath the positions indicate the distance of that position from DCS according to the magma colour gradient, where light yellow indicates sites closer to DCS (ligand distance in Angstroms). The positions are further annotated to reflect residues involved in interactions with DCS (green), and co-factor PLP (navy blue). Empty positions imply missing DST data preventing OR calculations. The figure is generated using R statistical software version 4.0.2, ggplot2 package. Abbreviations used: SAV: single amino acid variation, DCS: cycloserine, PLP: pyridoxal phosphate.

Mutation	Mutational effect	Mutational effect value	Lig-Dist (Å)	PPI2-Dist (Å)	Interacting partner
L113R	Mutation with highest OR	OR = 51.87	7.35	3.79	none
L113R	Most frequent mutation	MAF (%) = 14.20	7.35	3.79	none
V391G	Most Destabilising for protomer	ΔΔG = -0.64	9.85	5.92	none
S238L	Most Stabilising for protomer	ΔΔG = 0.71	5.21	9.92	none
E389G	Most Destabilising for DCS binding affinity	Log fold change = -0.68	7.50	2.95	none
	Most Destabilising for PPI affinity	ΔΔG = -2.17	7.50	2.95	none
L162M	Most Stabilising for PPI affinity	ΔΔG = 1.09	7.81	3.72	none

Table 1: Mutations with extreme effects
Mutations (SAVs) with extreme effects related to Odds Ratio (OR), mutational frequency (MAF), stability and affinity changes. For affinity changes only mutations within 10Å of DCS for DCS binding affinity, and Protein-Protein (PP) interface for PPI affinity were considered. The protomer stability changes are the average effect of all four estimates (mCSM-DUET, FoldX, DeepDDG and Dynamut2) combined, and the DCS binding affinity changes are the average effect of the two mCSM based tools (mCSM-lig and mmCSM-lig) combined. Changes in PP affinity correspond to estimates from mCSM-PPI. The estimated effects were categorised as Destabilising (log fold affinity change/ΔΔG<0) and Stabilising (log fold affinity change/ΔΔG>0). Abbreviations used: Å: Angstroms, ΔΔG: change in Gibbs free energy in Kcal/mol, MAF: minor allele frequency, SAV: single amino acid variation, Lig-Dist: distance to ligand, PPI-Dist: distance to protein-protein interface, DCS: cycloserine, PLP: pyridoxal phosphate.

7.2.7 Alr mutations in lineages

A small minority of samples contained mutations in Alr with lineage 1 showing the highest SAV diversity despite the lowest number of samples

Only 4% of samples (n=1,310) contained mutations in the protein coding region of *alr*, with 1025 samples contributing to the four main *M. tuberculosis* lineages (Lineages 1-4). Most samples belonged to lineage 4 (n=535), followed by lineage 2 (n=296), lineage 3 (n=106) and finally by lineage 1 (n=88) with the least number of samples **(Figure 14A)**. However, lineage 1 was high in its SAV diversity (55%, n=48), followed by lineage 3 (45% (n=48), and thereafter by lineages 4 and 2 with approximately similar SAV diversity (25%, n=135 vs 23%, n=68) SAVs respectively **(Figure 14B)**.

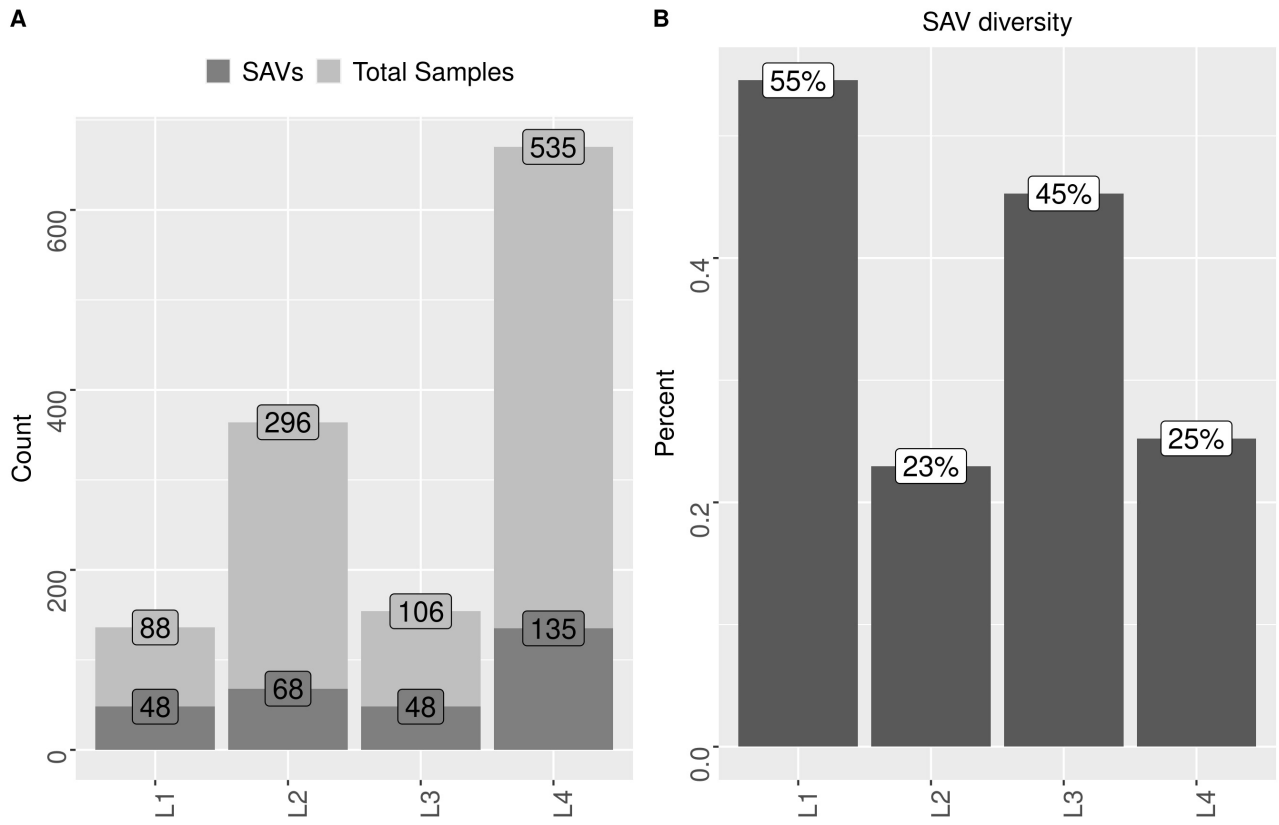


Figure 14: Lineage samples with *alr* mutations
Total number of samples (n=1,025) along with the number of mutations associated with DCS resistance in the four *M. tuberculosis* lineages (L1-L4). **A)** The dark grey bars show the number of mutations (SAVs), while the light grey bar show the total number of samples in each lineage, **B)** Mutational diversity in each lineage. Abbreviations used: SAV: single amino acid variation, DCS: cycloserine.

7.3 Chapter summary

With only a small minority of samples exhibiting SAVs, the contribution of SAVs in *alr* driven DCS resistance appears low. SAV mutations in Alr are spread across the protein with limited involvement of active site residues and, as such, mutations extend beyond the active site. Though most mutational effects destabilise the protomer overall, they occur in the highly variable regions of Alr that are not associated with high flexibility and, as such, are unlikely to affect protein function. This reinforces Alr's essential role as the target for DCS binding. The prominent mutational effects are related to a decrease in DCS binding affinity with nearly 15% mutations occurring close to DCS. Similarly, 45% mutations occurring at the dimer interface reducing the PP affinity highlights the importance of molecular interactions at the interface in maintaining the functions of the complex. The single mutation L113R at site L113, occurred most frequently and is strongly associated with DCS resistance. This suggests that site L113 is under selective pressure exerted by DCS, despite not being an active site residue. Mutations at active site residues M343T and Y388D observed in XDR strains were associated with DCS resistance, though association of DCS resistance for Y388D was much stronger (OR=16.50)

than observed for M343T (OR=1.91), suggesting that these mutations are acquired independently. Together, these findings suggest the delayed onset of *alr*-driven DCS resistance is mediated by factors involving interactions with other genes, and the possible role of evolutionary convergence for certain mutations, rather than being driven by a lack of SAVs within *alr*.

References

[1] Francis C. Neuhaus and Judy L. Lynch. *The Enzymatic Synthesis of D-Alanyl-D-alanine. III. On the Inhibition of D-Alanyl-D-alanine Synthetase by the Antibiotic D-Cycloserine**. ACS Publications. May 1, 2002. DOI: 10.1021/bi00892a001. URL: https://pubs.acs.org/doi/pdf/10.1021/bi00892a001 (visited on 08/19/2022).

[2] U. Strych et al. "Characterization of the Alanine Racemases from Two Mycobacteria". In: *FEMS microbiology letters* 196.2 (Mar. 15, 2001), pp. 93–98. ISSN: 0378-1097. DOI: 10.1111/j.1574-6968.2001.tb10547.x.

[3] Gareth A. Prosser and Luiz Pedro S. de Carvalho. "Kinetic Mechanism and Inhibition of Mycobacterium Tuberculosis D-Alanine:D-Alanine Ligase by the Antibiotic d-Cycloserine". In: *The FEBS Journal* 280.4 (2013), pp. 1150–1166. ISSN: 1742-4658. DOI: 10.1111/febs.12108.

[4] Israel G. Epstein, K.G.S. Nair, and Linn J. Boyd. "The Treatment of Human Pulmonary Tuberculosis With Cycloserine:* Progress Report**". In: *Diseases of the Chest* 29.3 (Mar. 1956), pp. 241–257. ISSN: 00960217. DOI: 10.1378/chest.29.3.241.

[5] Dimitrios Evangelopoulos et al. "Comparative Fitness Analysis of D-cycloserine Resistant Mutants Reveals Both Fitness-Neutral and High-Fitness Cost Genotypes". In: *Nature Communications* 10.1 (1 Sept. 13, 2019), p. 4177. ISSN: 2041-1723. DOI: 10.1038/s41467-019-12074-z.

[6] WHO consolidated guidelines DR-TB. *WHO Consolidated Guidelines on Drug-Resistant Tuberculosis Treatment*. Geneva: World Health Organization, 2019.

[7] Giovanna Riccardi, Maria Rosalia Pasca, and Silvia Buroni. "Mycobacterium Tuberculosis: Drug Resistance and Future Perspectives". In: *Future Microbiology* 4.5 (June 2009), pp. 597–614. ISSN: 1746-0913. DOI: 10.2217/fmb.09.20.

[8] Jiazhen Chen et al. "Identification of Novel Mutations Associated with Cycloserine Resistance in Mycobacterium Tuberculosis". In: *Journal of Antimicrobial Chemotherapy* 72.12 (Dec. 1, 2017), pp. 3272–3276. ISSN: 0305-7453, 1460-2091. DOI: 10.1093/jac/dkx316.

[9] Jeffrey M. Chen et al. "A Point Mutation in cycA Partially Contributes to the D-cycloserine Resistance Trait of Mycobacterium Bovis BCG Vaccine Strains". In: *PLoS ONE* 7.8 (Aug. 17, 2012), e43467. ISSN: 1932-6203. DOI: 10.1371/journal.pone.0043467.

[10] Stephanie Portelli et al. "Understanding Molecular Consequences of Putative Drug Resistant Mutations in Mycobacterium Tuberculosis". In: *Scientific Reports* (2018). ISSN: 20452322. DOI: 10.1038/s41598-018-33370-6.

[11] Yoshio Nakatani et al. "Role of Alanine Racemase Mutations in Mycobacterium Tuberculosis D -Cycloserine Resistance". In: *Antimicrobial Agents and Chemotherapy* 61.12 (Dec. 2017), e01575–17. ISSN: 0066-4804, 1098-6596. DOI: 10.1128/AAC.01575-17.

[12] Cesira de Chiara et al. "D-Cycloserine Destruction by Alanine Racemase and the Limit of Irreversible Inhibition". In: *Nature Chemical Biology* 16.6 (6 June 2020), pp. 686–694. ISSN: 1552-4469. DOI: 10.1038/s41589-020-0498-9.

[13] Christopher A. Desjardins et al. "Genomic and Functional Analyses of Mycobacterium Tuberculosis Strains Implicate Ald in D-cycloserine Resistance". In: *Nature Genetics* 48.5 (May 2016), pp. 544–551. ISSN: 1546-1718. DOI: 10.1038/ng.3548.

[14] Timothy D Fenn et al. "A Side Reaction of Alanine Racemase: Transamination of Cycloserine". In: *Biochemistry* 42.19 (May 1, 2003), pp. 5775–5783. ISSN: 1520-4995. DOI: 10.1021/bi027022d.

Appendix for Chapter 7

7.A Mutations close to cycloserine

Mutation	Interacting partner	Lig-Dist (Å)	mCSM-lig affinity	mCSM-lig outcome	mmCSM-lig affinity	mmCSM-lig outcome	MAF (%)	Odds Ratio	P-value	Adjusted P-value	Adjusted P-value significance
L113R	none	7.35	-1.28	Destabilising	-0.85	Destabilising	14.2	51.87	<0.0001	<0.001	****
Y388D	drug and plp	2.86	-2.4	Destabilising	-0.33	Destabilising	4.36	16.5	0.11	0.82	ns
A200V	drug	4.8	-0.7	Destabilising	-0.82	Destabilising	0.3	4.09	>1	>1	ns
S261N	none	7.91	-0.72	Destabilising	-0.92	Destabilising	3.35	1.64	0.5	1	ns
M62I	none	6.56	-1.68	Destabilising	-0.82	Destabilising	0.1	NA	NA	NA	ns
A63G	none	7.14	-2.15	Destabilising	-0.59	Destabilising	0.1	NA	NA	NA	ns
L89F	none	8.63	-1.82	Destabilising	-0.69	Destabilising	0.1	NA	NA	NA	ns
A92T	none	6.82	-0.54	Destabilising	-0.74	Destabilising	0.1	NA	NA	NA	ns
T155A	none	9.55	-2.3	Destabilising	-0.81	Destabilising	0.1	NA	NA	NA	ns
T155S	none	9.55	-1.74	Destabilising	-0.79	Destabilising	0.1	NA	NA	NA	ns
L162M	none	7.81	-0.87	Destabilising	-0.82	Destabilising	0.1	NA	NA	NA	ns
G166D	none	9.9	0.94	Stabilising	-1.97	Destabilising	0.1	NA	NA	NA	ns
M197I	none	6.94	0.31	Stabilising	-0.82	Destabilising	0.1	NA	NA	NA	ns
A200S	drug	4.8	-0.38	Destabilising	-0.96	Destabilising	0.1	NA	NA	NA	ns
D201G	none	7.16	-0.97	Destabilising	-0.4	Destabilising	0.1	NA	NA	NA	ns
S235W	none	5.3	0.42	Stabilising	-0.47	Destabilising	0.1	NA	NA	NA	ns
S237A	drug and plp	2.68	-0.64	Destabilising	-0.91	Destabilising	0.1	NA	NA	NA	ns
S238L	none	5.21	-0.39	Destabilising	-0.28	Destabilising	0.1	NA	NA	NA	ns
A256S	drug	4.22	-0.42	Destabilising	-0.96	Destabilising	0.1	NA	NA	NA	ns
A256T	drug	4.22	-0.52	Destabilising	-0.92	Destabilising	0.1	NA	NA	NA	ns
V257L	none	6.15	-1.79	Destabilising	-0.82	Destabilising	0.1	NA	NA	NA	ns
L260V	none	6.86	-1.77	Destabilising	-0.69	Destabilising	0.1	NA	NA	NA	ns

H387Q	none	8.05	-0.66	Destabilising	-0.93	Destabilising	0.1	NA	NA	NA	ns
Y388C	drug	2.86	-1.41	Destabilising	-0.56	Destabilising	0.1	NA	NA	NA	ns
V391G	none	9.85	-1.39	Destabilising	-0.29	Destabilising	0.1	NA	NA	NA	ns
T392P	none	8.58	-1.19	Destabilising	-1.1	Destabilising	0.1	NA	NA	NA	ns
P253L	drug	3.84	-0.78	Destabilising	-0.6	Destabilising	0.2	NA	NA	NA	ns
V263A	none	8.71	-0.28	Destabilising	-0.55	Destabilising	0.2	NA	NA	NA	ns
A92D	none	6.82	-0.94	Destabilising	-0.74	Destabilising	0.3	NA	NA	NA	ns
P262S	none	5.83	0.53	Stabilising	-1.2	Destabilising	0.3	NA	NA	NA	ns
E389G	none	7.5	-2.91	Destabilising	-0.71	Destabilising	0.3	NA	NA	NA	ns
A200D	drug	4.8	-1.47	Destabilising	-0.74	Destabilising	0.41	NA	NA	NA	ns
T385P	none	9.82	-1.17	Destabilising	-1.1	Destabilising	0.51	NA	NA	NA	ns
H387N	none	8.05	-0.43	Destabilising	-1.02	Destabilising	0.61	NA	NA	NA	ns
A131G	none	8.11	-1.33	Destabilising	-0.59	Destabilising	0.81	NA	NA	NA	ns
P262L	none	5.83	-0.14	Destabilising	-0.6	Destabilising	0.81	NA	NA	NA	ns
H387Y	none	8.05	-1.95	Destabilising	-0.85	Destabilising	0.81	NA	NA	NA	ns
L61V	none	9.84	-0.88	Destabilising	-0.69	Destabilising	1.72	NA	NA	NA	ns
P262Q	none	5.83	0.61	Stabilising	-0.88	Destabilising	2.54	NA	NA	NA	ns
K157E	drug	3.96	-0.6	Destabilising	-0.93	Destabilising	2.74	NA	NA	NA	ns

Table 7.A.1: Mutations close to DCS
Forty mutations single amino acid variation (SAV) mutations lying within 10Å of DCS and their corresponding ligand affinity changes (log fold change) measured by mCSM-Lig and mmCSM-lig. The estimated effect are categorised as Destabilising (log fold affinity change<0) and Stabilising ($\Delta\Delta G>0$). The genomic measures of minor allele frequency (MAF), Odds Ratio (OR), OR related P-values, and FDR adjusted P-values are shown. Statistical significance is indicated as: *P < 0.05, **P < 0.01, ***P < 0.001, ****P < 0.0001, ns:>0.05. The table is arranged by OR to show mutation with the highest OR at the top for mutations close to DCS. Columns with NA indicate insufficient data to calculate OR. Abbreviations used: FDR: false discovery rate, ns: not significant, DCS: cycloserine.

7.B Mutations close to the protein-protein interface

Mutation	Interacting partner	PPI2-Dist (Å)	mCSM-PPI2 (ΔΔG)	mCSM-PPI2 outcome	MAF (%)	Odds Ratio	P-value	Adjusted P-value	Adjusted P-value significance
L113R	none	3.79	-0.74	Decreasing	14.2	51.87	<0.0001	<0.0001	****
Y388D	drug and plp	3.18	-1.5	Decreasing	4.36	16.5	0.11	0.82	ns
L283P	none	3.8	-1.08	Decreasing	0.1	4.09	>1	>1	ns
G292A	none	3.48	-0.75	Decreasing	0.2	4.09	>1	>1	ns
H297R	none	2.89	-0.25	Decreasing	0.2	4.09	>1	>1	ns
A200V	drug	5.16	0.12	Increasing	0.3	4.09	>1	>1	ns
A358T	none	7	0.26	Increasing	0.3	4.09	>1	>1	ns
T399I	none	7.9	-0.07	Decreasing	0.41	4.09	>1	>1	ns
S321L	none	5.06	-0.06	Decreasing	0.51	4.09	>1	>1	ns
E140D	none	6.38	-0.06	Decreasing	0.71	4.09	>1	>1	ns
M343T	drug	3.73	-1.62	Decreasing	4.16	1.91	0.4	>1	ns
L35F	none	4.49	0.01	Increasing	0.1	NA	NA	NA	ns
A36V	none	7.63	0.15	Increasing	0.1	NA	NA	NA	ns
D68N	none	6.89	0.31	Increasing	0.1	NA	NA	NA	ns
A92T	none	3.14	0.26	Increasing	0.1	NA	NA	NA	ns
V94I	none	7.95	-0.3	Decreasing	0.1	NA	NA	NA	ns
D95G	none	6.05	-0.51	Decreasing	0.1	NA	NA	NA	ns
I118T	none	8.3	-0.1	Decreasing	0.1	NA	NA	NA	ns
I118V	none	8.3	-0.23	Decreasing	0.1	NA	NA	NA	ns
L162M	none	3.72	1.09	Increasing	0.1	NA	NA	NA	ns
G166D	none	5.41	-0.48	Decreasing	0.1	NA	NA	NA	ns
A170T	none	6.75	0.18	Increasing	0.1	NA	NA	NA	ns

M197I	none	3.7	-0.63	Decreasing	0.1	NA	NA	NA	ns
A200S	drug	5.16	0.4	Increasing	0.1	NA	NA	NA	ns
D201G	none	5.21	-0.61	Decreasing	0.1	NA	NA	NA	ns
S237A	drug and plp	9.85	-0.22	Decreasing	0.1	NA	NA	NA	ns
S238L	none	9.92	-0.29	Decreasing	0.1	NA	NA	NA	ns
A256S	drug	9.79	0.72	Increasing	0.1	NA	NA	NA	ns
A256T	drug	9.79	0.51	Increasing	0.1	NA	NA	NA	ns
L260V	none	8.03	-0.42	Decreasing	0.1	NA	NA	NA	ns
A280S	none	5.4	0.28	Increasing	0.1	NA	NA	NA	ns
A280V	none	5.4	0.03	Increasing	0.1	NA	NA	NA	ns
V284A	none	7.76	-0.26	Decreasing	0.1	NA	NA	NA	ns
V284M	none	7.76	-0.38	Decreasing	0.1	NA	NA	NA	ns
R288H	none	5.7	-0.05	Decreasing	0.1	NA	NA	NA	ns
A289E	none	8.16	0.53	Increasing	0.1	NA	NA	NA	ns
Y295F	drug	2.95	-0.59	Decreasing	0.1	NA	NA	NA	ns
G296A	none	2.84	-0.78	Decreasing	0.1	NA	NA	NA	ns
W299R	none	6.77	-0.7	Decreasing	0.1	NA	NA	NA	ns
N306D	none	9.35	-0.27	Decreasing	0.1	NA	NA	NA	ns
A308T	none	8.25	0.24	Increasing	0.1	NA	NA	NA	ns
A308V	none	8.25	0.2	Increasing	0.1	NA	NA	NA	ns
Y314C	drug	3.15	-1.93	Decreasing	0.1	NA	NA	NA	ns
D316G	none	6.15	-0.99	Decreasing	0.1	NA	NA	NA	ns
R320W	none	2.94	0.02	Increasing	0.1	NA	NA	NA	ns
S321A	none	5.06	-0.16	Decreasing	0.1	NA	NA	NA	ns
G323S	none	8.63	-0.02	Decreasing	0.1	NA	NA	NA	ns
G323V	none	8.63	0.03	Increasing	0.1	NA	NA	NA	ns

R325P	none	8.6	-1.09	Decreasing	0.1	NA	NA	NA	ns
V338G	none	7.49	-0.37	Decreasing	0.1	NA	NA	NA	ns
M347L	none	4.36	-0.75	Decreasing	0.1	NA	NA	NA	ns
D361G	none	6.41	-0.43	Decreasing	0.1	NA	NA	NA	ns
D361N	none	6.41	-0.43	Decreasing	0.1	NA	NA	NA	ns
D381N	none	9.21	-0.44	Decreasing	0.1	NA	NA	NA	ns
G384D	none	4.42	0.22	Increasing	0.1	NA	NA	NA	ns
H387Q	none	6.59	-0.33	Decreasing	0.1	NA	NA	NA	ns
Y388C	drug and plp	3.18	-1.26	Decreasing	0.1	NA	NA	NA	ns
V391G	none	5.92	-0.84	Decreasing	0.1	NA	NA	NA	ns
T392P	none	3.34	-1.36	Decreasing	0.1	NA	NA	NA	ns
R397C	none	3.37	-0.75	Decreasing	0.1	NA	NA	NA	ns
G71S	none	6.73	-0.32	Decreasing	0.2	NA	NA	NA	ns
A97V	none	9.94	0.3	Increasing	0.2	NA	NA	NA	ns
G117D	none	6.63	0.19	Increasing	0.2	NA	NA	NA	ns
G117S	none	6.63	-0.01	Decreasing	0.2	NA	NA	NA	ns
L135V	none	8.63	-0.35	Decreasing	0.2	NA	NA	NA	ns
D205G	none	7.74	-0.25	Decreasing	0.2	NA	NA	NA	ns
A280P	none	5.4	0.06	Increasing	0.2	NA	NA	NA	ns
G290W	none	7.45	-0.1	Decreasing	0.2	NA	NA	NA	ns
V293M	none	3.6	-1.41	Decreasing	0.2	NA	NA	NA	ns
R303H	none	9.22	-0.01	Decreasing	0.2	NA	NA	NA	ns
N306T	none	9.35	-0.27	Decreasing	0.2	NA	NA	NA	ns
L307V	none	5.05	0.19	Increasing	0.2	NA	NA	NA	ns
D316E	none	6.15	-0.64	Decreasing	0.2	NA	NA	NA	ns
D316Y	none	6.15	-0.35	Decreasing	0.2	NA	NA	NA	ns

S321P	none	5.06	-0.07	Decreasing	0.2	NA	NA	NA	ns
M347I	none	4.36	-0.23	Decreasing	0.2	NA	NA	NA	ns
V348I	none	8.78	-0.32	Decreasing	0.2	NA	NA	NA	ns
T399A	none	7.9	-0.1	Decreasing	0.2	NA	NA	NA	ns
T399N	none	7.9	-0.23	Decreasing	0.2	NA	NA	NA	ns
H72Y	none	8.39	0.05	Increasing	0.3	NA	NA	NA	ns
A92D	none	3.14	0.25	Increasing	0.3	NA	NA	NA	ns
P169S	none	8.42	-0.07	Decreasing	0.3	NA	NA	NA	ns
P262S	none	9.16	-0.07	Decreasing	0.3	NA	NA	NA	ns
I287V	none	3.19	0	Decreasing	0.3	NA	NA	NA	ns
R340L	none	5.89	-0.45	Decreasing	0.3	NA	NA	NA	ns
A358G	none	7	-0.42	Decreasing	0.3	NA	NA	NA	ns
E389G	none	2.95	-2.17	Decreasing	0.3	NA	NA	NA	ns
T399S	none	7.9	-0.21	Decreasing	0.3	NA	NA	NA	ns
A200D	drug	5.16	0.3	Increasing	0.41	NA	NA	NA	ns
K285N	none	2.97	-0.3	Decreasing	0.41	NA	NA	NA	ns
R397G	none	3.37	-0.65	Decreasing	0.41	NA	NA	NA	ns
I398V	none	9.57	-0.3	Decreasing	0.41	NA	NA	NA	ns
V132L	none	9.78	-0.27	Decreasing	0.51	NA	NA	NA	ns
R136H	none	4.13	-0.14	Decreasing	0.51	NA	NA	NA	ns
T160A	none	6.94	-0.31	Decreasing	0.51	NA	NA	NA	ns
I300T	none	8.54	-0.1	Decreasing	0.51	NA	NA	NA	ns
P311L	none	4.47	-1.04	Decreasing	0.51	NA	NA	NA	ns
P311S	none	4.47	-0.14	Decreasing	0.51	NA	NA	NA	ns
T385P	none	3.59	-1.34	Decreasing	0.51	NA	NA	NA	ns
L135E	none	8.63	-0.38	Decreasing	0.61	NA	NA	NA	ns

V318M	none	4.56	-0.42	Decreasing	0.61	NA	NA	NA	ns
D381E	none	9.21	0.04	Increasing	0.61	NA	NA	NA	ns
H387N	none	6.59	0.22	Increasing	0.61	NA	NA	NA	ns
A308G	none	8.25	0.18	Increasing	0.71	NA	NA	NA	ns
R397L	none	3.37	-0.71	Decreasing	0.71	NA	NA	NA	ns
L135Q	none	8.63	-0.21	Decreasing	0.81	NA	NA	NA	ns
Q137H	none	6.31	0.22	Increasing	0.81	NA	NA	NA	ns
P262L	none	9.16	-0.36	Decreasing	0.81	NA	NA	NA	ns
A280T	none	5.4	0.39	Increasing	0.81	NA	NA	NA	ns
H387Y	none	6.59	0.09	Increasing	0.81	NA	NA	NA	ns
H114P	none	4.93	-0.6	Decreasing	1.12	NA	NA	NA	ns
I118F	none	8.3	0.28	Increasing	1.12	NA	NA	NA	ns
G292S	none	3.48	-0.13	Decreasing	1.12	NA	NA	NA	ns
I312V	none	6.03	-0.35	Decreasing	1.12	NA	NA	NA	ns
V318I	none	4.56	-0.39	Decreasing	1.12	NA	NA	NA	ns
D344N	drug	3.21	-1.22	Decreasing	1.42	NA	NA	NA	ns
V383L	none	5.96	-0.19	Decreasing	2.23	NA	NA	NA	ns
P262Q	none	9.16	-0.03	Decreasing	2.54	NA	NA	NA	ns
M347V	none	4.36	-0.68	Decreasing	2.64	NA	NA	NA	ns
K157E	drug	6.5	-0.03	Decreasing	2.74	NA	NA	NA	ns
R288S	none	5.7	-0.38	Decreasing	4.16	NA	NA	NA	ns
V284L	none	7.76	-0.26	Decreasing	4.26	NA	NA	NA	ns

Table 7.B.1: Mutations close to Alr PPI
One hundred and twenty two single amino acid variation (SAV) mutations lying within 10Å of the protein-protein interface (PPI) and their corresponding PPI affinity changes ($\Delta\Delta$G) measured by mCSM-PPI2. The estimated effect are categorised as Destabilising ($\Delta\Delta$G<0) and Stabilising ($\Delta\Delta$G>0). The genomic measures of minor allele frequency (MAF), Odds Ratio (OR), OR related P-values, and FDR adjusted P-values are shown. Statistical significance is indicated as: *P < 0.05, **P < 0.01, ***P < 0.001, ****P < 0.0001, ns: >0.05. The table is arranged by OR to show mutation with the highest OR at the top for mutations at the PPI. Columns with NA indicate insufficient data to calculate OR. Abbreviations used: $\Delta\Delta$G: change in Gibbs free energy in Kcal/mol, FDR: false discovery rate, ns: not significant, DCS: cycloserine.

Chapter 8

RpoB-rifampicin results

8.1 Background

8.1.1 Mechanism of action of rifampicin

Rifampicin (RFP) is an antibiotic that is used in the treatment of both active and latent TB.[1] It is used with other antibiotics such as isoniazid, pyrazinamide and ethambutol. RFP in combination with isoniazid forms the basis of treatment for MDR-TB.[2] RFP inhibits the elongation of messenger RNA by binding to the β-subunit of the bacterial RNA polymerase (RpoB RNAP).[3] The target for RFP is the *rpoB* gene encoding a DNA-dependent RNA polymerase enzyme.[4] It has been shown that RFP binding to *rpoB* induces hydroxyl radical formation in susceptible but not resistant bacilli, contributing to the bactericidal activity of RFP.[5]

8.1.2 Rifampicin resistance in *M. tuberculosis*

SAVs in *rpoB* are the major contributing factor to RFP resistance development.[6–8] About 96% of *rpoB* SAVs occur frequently within a 81 base pair (bp) region, spanning codons 507-533 (encoding 27 amino acids). This region is commonly known as the Rifampicin Resistance Determining Region (RRDR).[7] RFP resistant mutations are particularly found in codons 518, 523-529, and 531[9,10] **(Figure 1)**. Despite this, mutations beyond the RRDR have been investigated, and reported to being linked with RFP resistance.[11,12] Further, the role of compensatory mutations in *rpoA* and *rpoC* in restoring any fitness penalty from mutations in *rpoB* have been shown to result in strains with a high degree of transmissibility.[13–15]

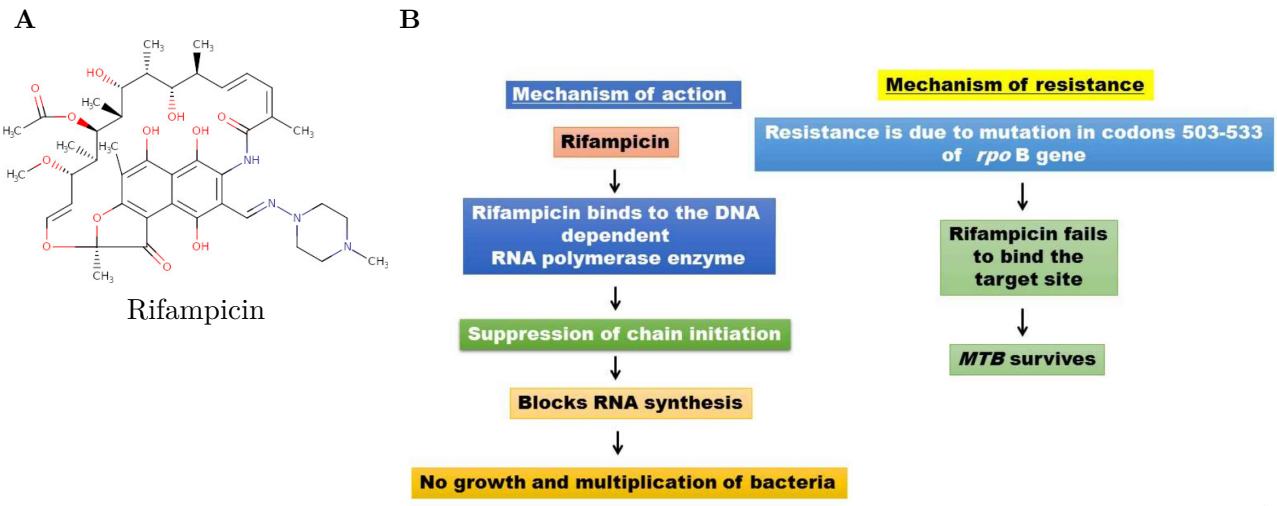


Figure 1: Chemical structure and mechanism of action and resistance for rifampicin
A) The chemical structure of rifampicin (RFP) is shown at the top left and is sourced from DrugBank (ID: DB01045), and shows an aromatic core linked by aliphatic chains, **B)** An overview of the mechanism of action and resistance for Rifampicin. Figure adapted from Sheikh *et. al.*.[16]

8.1.2.1 Active site description of the RpoB RNAP-RFP complex

The genomic location and coding region of *rpoB* was retrieved from the Mycobrowser database (Rv0667c: 759807-763325). An experimentally determined 3D atomic structure of the RpoB RNAP complex from *M. tuberculosis* containing a 3 nucleotide RNA with RFP bound is available as the PDB entry 5UHC as a native hetero-hexamer.[17] RFP is bound at the β subunit of *rpoB* (chain C), while the transcribed DNA i.e. nucleic acid (NA) is present at the cleft between *rpoB*, and *rpoC* (chain D, subunit β'). The other chains in the complex comprise of *rpoA* (chain A and B, subunits α1 and α2), *rpoZ* (chain E, subunit ω), and *rpoD* (chain F, subunit SigA).[17]

Interactions with RFP

Molecular interactions between RFP and RpoB RNAP were identified using LigPlus, PLIP and Arpeggio resulting in a total of twenty-one interaction sites: 70, 428, 429, 430, 431, 432, 433, 435, 445, 448, 450, 452, 453, 458, 459, 483, 487, 491, 604, 607, 674. An overview of the RpoB RNAP structural complex with all interactions identified is shown in **Figure 2**.

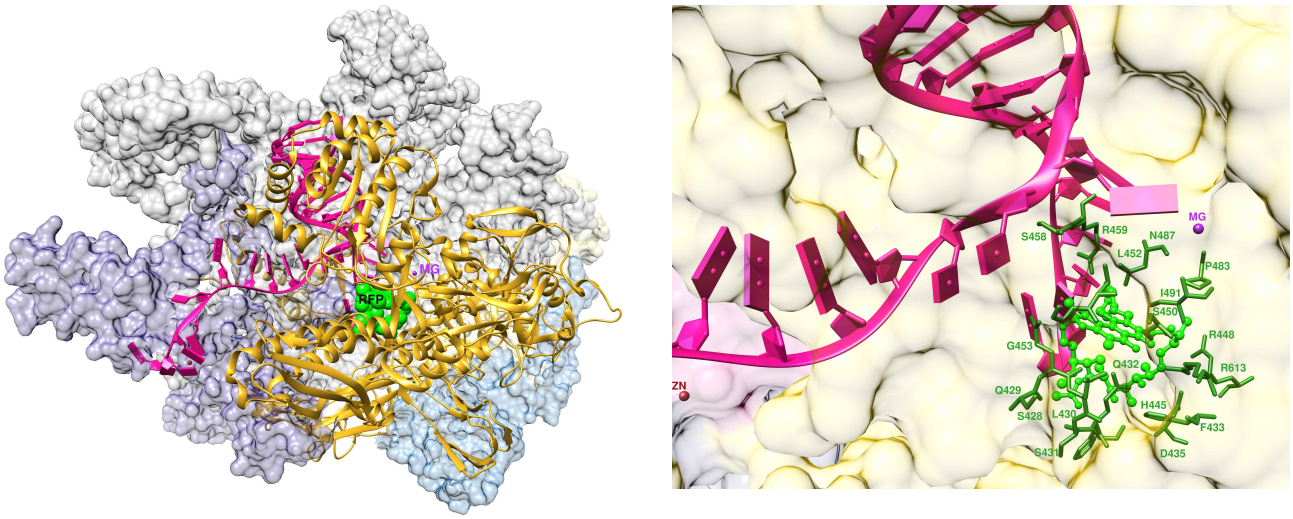


Figure 2: Description of *M. tuberculosis* RpoB RNA polymerase β subunit complex with rifampicin bound
Left panel shows the overall description of hetero-hexamer RpoB RNA polymerase β subunit in complex with RFP (PDB-ID: 5UHC) shown as surface representation, with chain C (*rpoB*) appearing as gold ribbon. The transcribed DNA i.e. nucleic acid (NA) is present at the cleft between subunit complex. Chains A, B and D appear in steel blue, light yellow, and dark grey surfaces. RFP is shown in green as spheres, with nucleic acid appearing in bright pink. Right panel shows a close-up view of RpoB RNAP interactions with RFP shown as green ball-and-stick. RFP interacting residues are indicated in green, with nucleic acid in bright pink, along with Mg^{2+} and Zn^{2+} ions which formed part of the complex. Abbreviations used: RFP: rifampicin.

8.2 Structural and genomic insights into rifampicin resistance

8.2.1 Mutational landscape of RpoB RNAP

Mutations peak around the active site with residue H445 showing a maximum of 13 SAVs

A total of 1132 SAVs were located in the protein coding region of *rpoB* (Rv0667c: 759807-763325). The mutational landscape is distributed across the protein **(Figure 3)** with mutations present in 631 unique positions, with a maximum of 13 SAVs at a single site **(Figure 4)**.

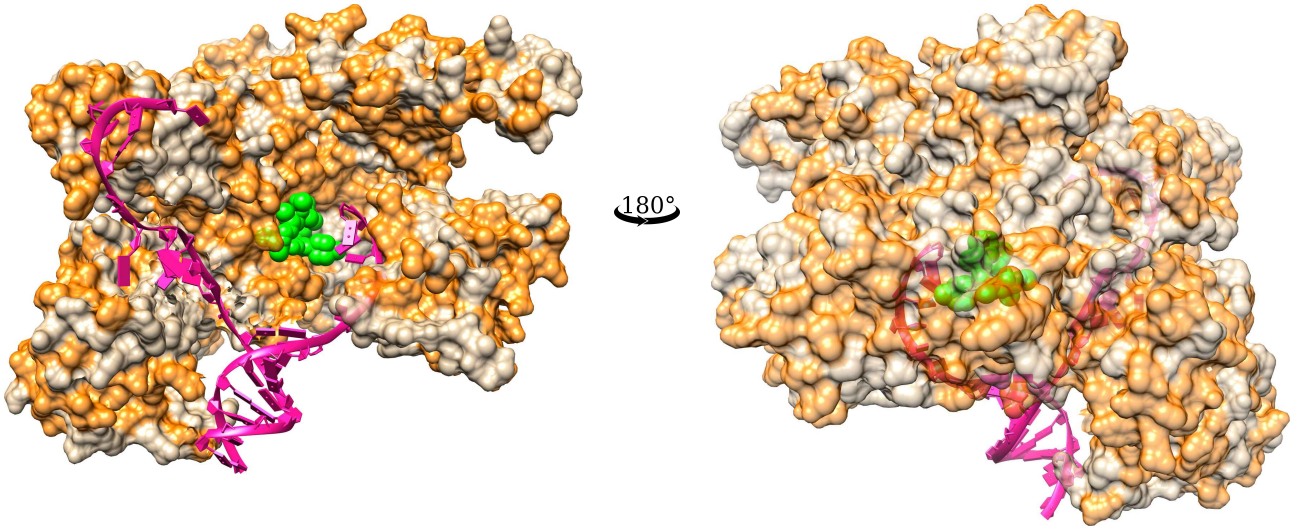


Figure 3: Mutational landscape of *M. tuberculosis* RpoB RNA polymerase β subunit
An overview of all mutational sites on *M. tuberculosis* RpoB RNA polymerase β subunit (PDB-ID: 5UHC, chain C) appearing as surface representation in tan colour, nucleic acid (NA) bound in complex is shown in bright pink, while the RFP appears as green spheres. The left and right panels are opposing representations (rotated 180°) of RFP, with RFP shown as green spheres. The figure is generated using UCSF Chimera version 1.14. Abbreviations used: SAV: single amino acid variation, RFP: rifampicin.

Mapping mutations by position on the RpoB RNAP **(Figure 4)** highlight the following:

Sites interacting with RFP were associated with a maximum of 13 SAVs at a single site (sites marked in green)

- Single mutations: F433, S458, R459, and N487
- Budding resistant hotspots: R448, G453, P483
- Hotspots with three mutations: S428, L430
- Hotspots with four mutations: Q429, S431, L452
- Hotspots with five mutations: Q432
- Hotspots with seven mutations: I491
- Hotspots with ten mutations: D435

- Hotspots with eleven mutations: S450
- Hotspots with thirteen: H445

Sites away from RFP also showed a maximum of 5 SAVs at a single site. A majority (56%, n=639) of the mutational effects resulted in electrostatic changes **(Figure 4)**.

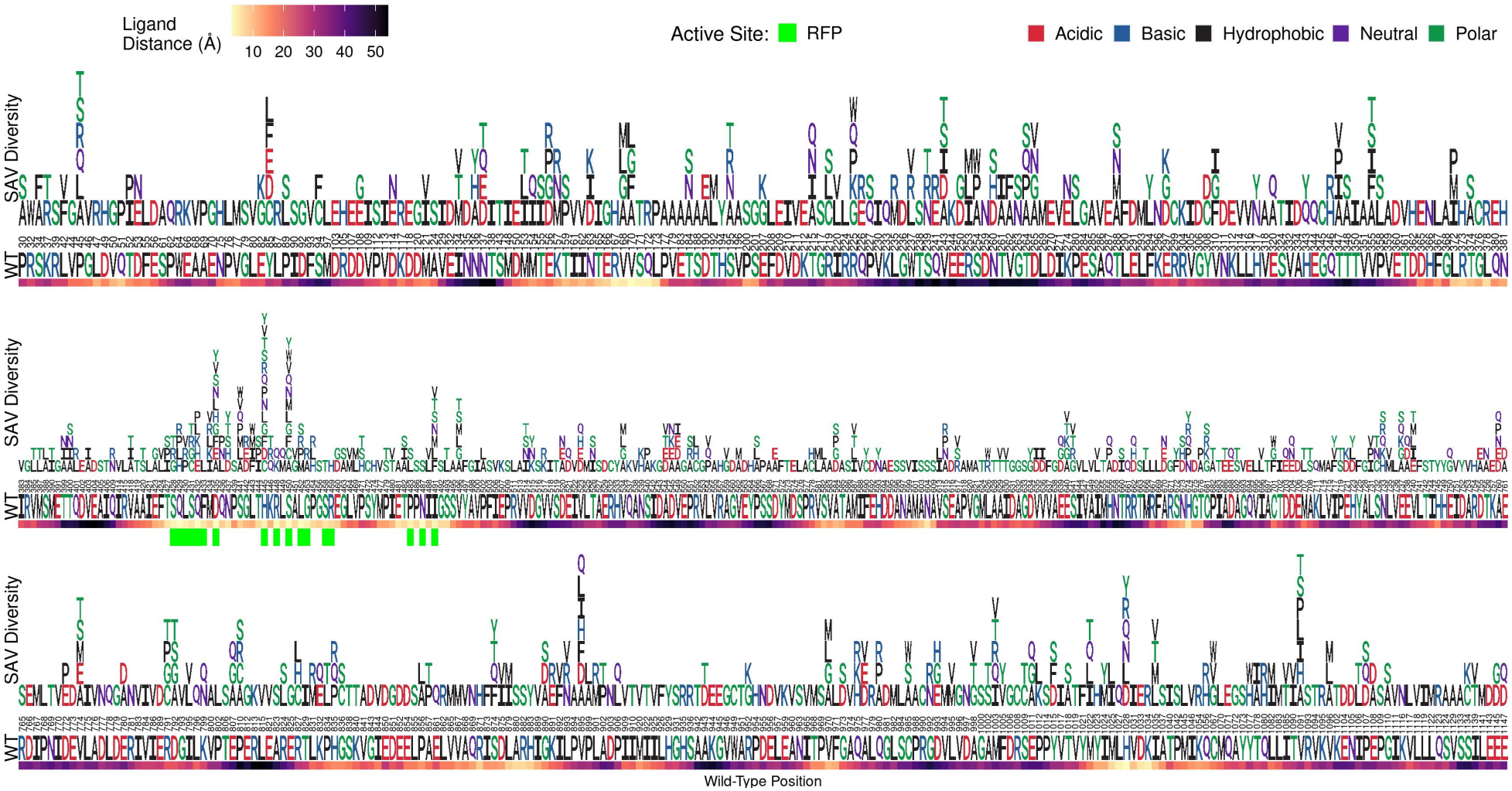


Figure 4: Sites associated with SAVs in *M. tuberculosis* RpoB RNA polymerase β subunit complex

Logo plot showing 631 unique sites/positions associated with 1132 SAVs in *M. tuberculosis* RpoB RNA polymerase β subunit. The horizontal axis shows the wild-type positions associated with SAVs in RpoB RNA polymerase β subunit and the vertical axis shows all the mutant residues observed in our data highlighting SAV diversity at a given site. Residues are coloured according to the amino acid (aa) property where acidic aa appear in red, basic aa appear in blue, hydrophobic aa in black, neutral aa in purple, and polar aa in darkgreen. The structural positions associated with SAVs in RpoB RNA polymerase β subunit are indicated on the horizontal axis. The wild-type (WT) residues also coloured according to aa property appear under the respective position markings. The heat bar underneath the WT residues indicate the distance of that position from RFP according to the magma colour gradient where light yellow indicates sites closer to RFP (ligand distance in Angstroms). The positions are further annotated to reflect active site residues involved in interactions with RFP (green). The figure is generated using R statistical software version 4.0.2, ggplot2 package. Abbreviations used: SAV: single amino acid variation, RFP: rifampicin.

8.2.2 Mutational outcome from protomer stability changes and evolutionary conservation

Most mutational consequences are destabilising for protomer stability with possible consequences on protein function

Most mutations had a destabilising effect on the overall protomer stability when measured by the different computational tools **(Figure 5A-4D)**, with DeepDDG estimating 88% (n=997) as destabilising, followed by ∼81% of mutations estimated by Dynamut2 (n=687) as destabilising, about 78% (n=774) estimated as destabilising by mCSM-DUET, followed by FoldX predicting ∼69% (n=778) mutations as destabilising. Based on an analysis of evolutionary conservation, PROVEAN estimated nearly equal numbers of mutations as deleterious (n=564) and neutral (n=568) **(Figure 5E)** while SNAP2 estimated 61% (n=691) with non-deleterious impact **(Figure 5F)**.

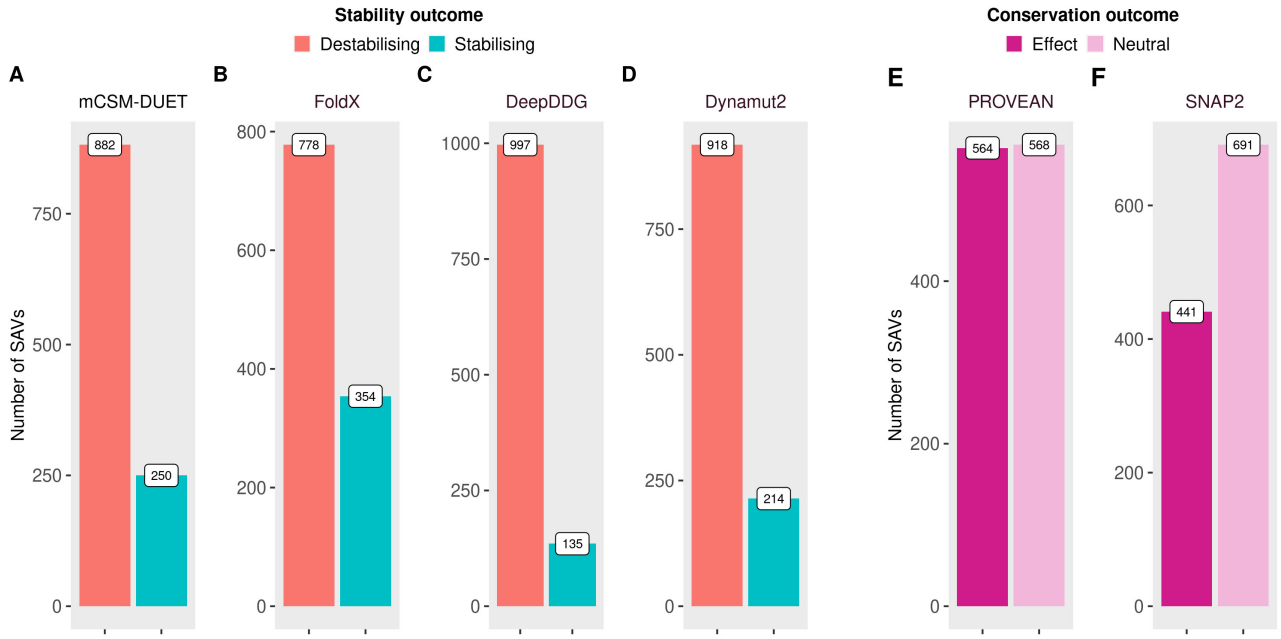


Figure 5: Protein stability outcome of SAVs in *M. tuberculosis* RpoB RNA polymerase β subunit Mutational impact on overall protein stability and evolutionary conservation changes for 1132 SAVs, **A-D)** Barplots showing number of SAVs categorised as destabilising (red) or stabilising (blue) according to protein stability changes (ΔΔG Kcal/mol) measured by four computational tools: mCSM-DUET, FoldX, DeepDDG, and Dynamut2, **E-F)** Number of SAVs categorised as Effect/Deleterious (magenta) or Neutral (pink) according to evolutionary conservation changes estimated by computational tools: PROVEAN, and SNAP2. The figure is generated using R statistical software version 4.0.2, ggplot2 package. Abbreviations used: ΔΔG: change in Gibbs free energy, SAV: single amino acid variation.

Evolutionary and structure-based predictors provide different insights into understanding mutational impact. Mutational impact in this context is considered to be its effect on protein stability, drug binding affinity, other binding affinities such as PPI or nucleic acid, and functional effects arising from protein sequence variations. The first three mutational consequences are assessed by structure based predictors relying on the 3D structure of a protein, while the last is assessed by sequence based

predictors relying mainly on evolutionary conservation trends across many proteins using multiple sequence alignments. The sequence based predictors are aimed at predicting pathogenicity or change of molecular function, structure based tools rely on estimating variant effects in relation to structure damage, corresponding to stability changes, as protein stability is considered the basic characteristic affecting function, activity, and regulation. Predictors such as ConSurf are able to use both structural and sequence information to identify important functional regions conserved in proteins. A variant classified as 'deleterious' to protein conservation may display gain-of-function in the presence of a drug through optimised protein stability. Thus, different methodological strategies benefit from complementary information when assessing specific proteins.

8.2.3 Mutational consequences on affinity changes and prominent mutational effects

Mutations reduce binding affinity for RFP, and PPI without largely affecting affinity for NA

About 15% (n=168) of SAVs inducing changes in RFP binding affinity were within 10Å of RFP. These mutations occurred at 54 distinct sites, with most sites (n=16) showing single mutations. Nearly 74% of mutations (n=168) were predicted to result in a destabilising effect on RFP binding affinity measured by mCSM-lig, and all (n=168) mutations were destabilising according to mmCSM-lig **(Figure 6A** top panel, Appendix Table 8.A.1**)**. When the 36 mutational sites with their average effect on RFP binding affinity were mapped onto the RpoB RNAP, these were shown to result in mild-moderate destabilising mutational consequences **(Figure 6A** bottom panel**)**.

Analysing sites close to the nucleic acid highlighted around 17% (n=195) mutations, corresponding to 86 distinct sites within 10Å of the NA measured by mCSM-NA, with 58% (n=114) of mutations classed as destabilising (**Figure 6B** top panel**)**. While sites around the NA showed a mixture of stability effects on visual inspection, sites with stabilising mutations appear in the immediate surrounding areas to the NA, followed by mild-moderate destabilising mutations, with strongly destabilising mutations located further away **(Figure 6B** bottom panel, Appendix Table 8.B.1**)**.

Considering the hetero-hexamer RpoB RNAP, the PPI surfaces covered nearly 60% (n=674) mutations at 367 distinct sites located within 10Å of the PPI measured by mCSM-PPI2, with 70% (n=467) of mutations resulting in destabilising effects **(Figure 6C** top panel**)**. Sites around the PPI showed a mixture of effects but predominately with mild stability impact **(Figure 6C** bottom panel, Appendix Table 8.C.1**)**. Of the total 631 unique sites in RpoB RNAP displaying SAVs, about 58% of sites (n=364) showed single mutations, with 23% (n=146) of sites as budding resistant hotspots **(Figure 7**

top panel).

The most prominent effects on RFP binding were from reduced (destabilising) affinity to RFP contributed by mutations from 31 surrounding sites. Four sites contributing mutations that increased the binding affinity to RFP (**Figure 7**, yellow text boxes, and bottom panel). Sites around the NA had mutations that contributed nearly equally towards decreasing (n=11) and increasing NA affinity (n=10), with stabilising mutations appearing to be located farther away from NA on visual inspection, with destabilising mutational sites appearing to be comparatively closer to the NA (**Figure 7**, yellow text boxes, and bottom panel). Sites close to the PPI had mostly destabilising mutations from 54 surrounding sites, while 27 had a stabilising mutational impact (**Figure 7**, pink text boxes, and bottom panel). All other sites were largely (n=418) affected by destabilising mutations (**Figure 7**, blue and red text boxes, and bottom panel) impacting protomer stability.

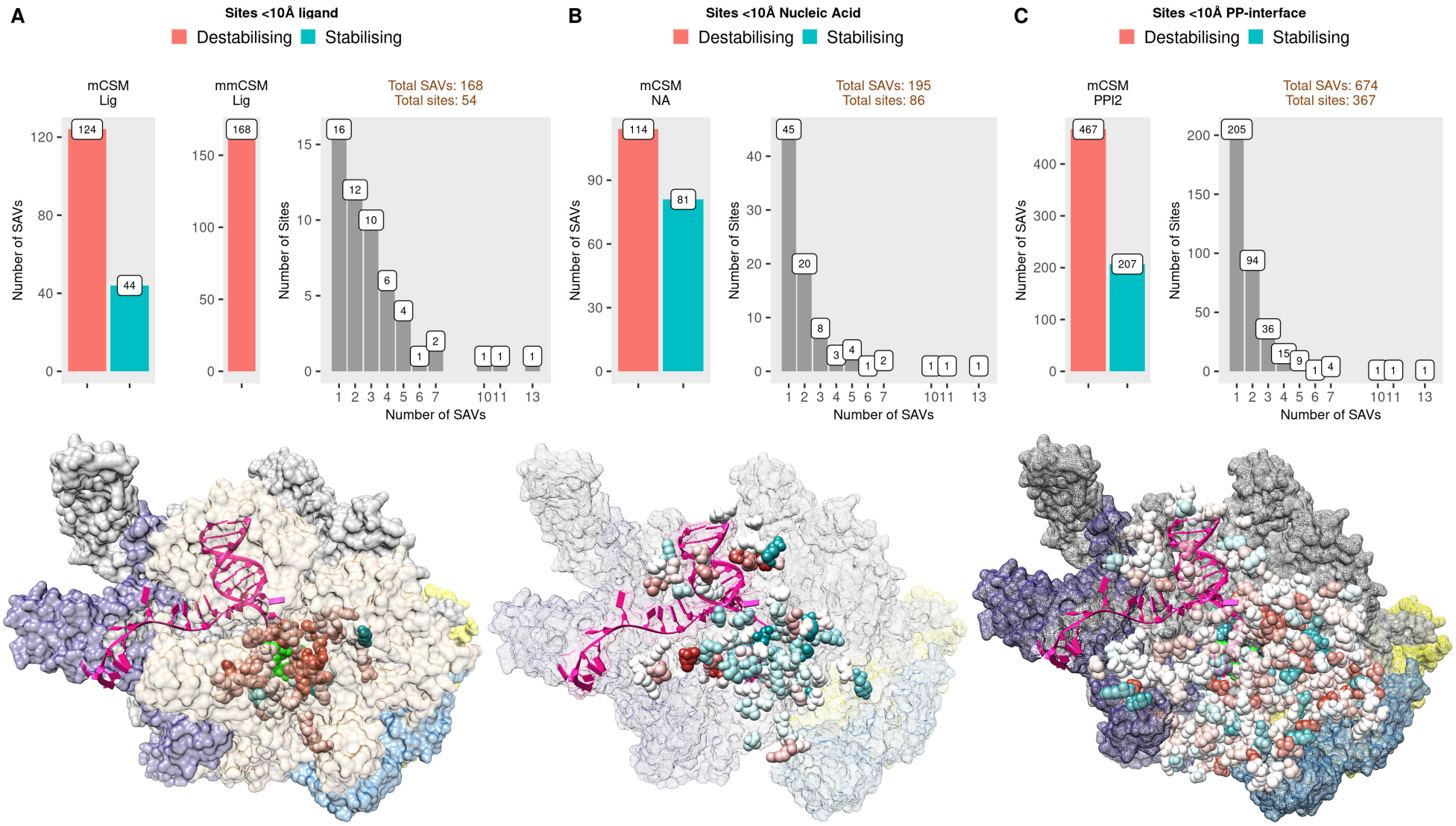


Figure 6: Mutational impact on binding affinities for RFP, nucleic acid, and protein-protein interface in *M. tuberculosis* RpoB RNA polymerase β subunit

The top panel displays barplots showing the mutational outcome of affinity changes and their corresponding site frequency, while the bottom panel shows the corresponding mutational impact mapped onto the RpoB RNA polymerase β subunit (PDB-ID: 5UHC, chain C) shown as surface representation. Chains A, B and D appear in steel blue, light yellow, and dark grey. RFP is shown as green spheres while the nucleic acid (NA) fragment is indicated in pink. **A)** Mutational impact on RFP binding (log fold change) from mCSM-lig and mmCSM-lig with 168 mutations at 54 sites within 10Å of RFP, **B)** Mutational impact on NA binding affinity ($\Delta\Delta$G) for 195 mutations at 86 sites within 10Å of the NA, **C)** Mutational impact on protein-protein (PP) affinity ($\Delta\Delta$G) for 674 mutations at 367 sites within 10Å of the PPI. For all parts, red denotes destabilising and blue denotes stabilising mutational sites, and the colour intensity reflects the extent of the effect: -1 (most destabilising) to +1 (most stabilising). Barplots are generated using R statistical software version 4.0.4, ggplot2 package. The structure figures are generated using Chimera version 1.14. Abbreviations used: Å: angstroms, $\Delta\Delta$G: change in Gibbs free energy in kcal/mol, SAV: single amino acid variation, RFP: rifampicin.

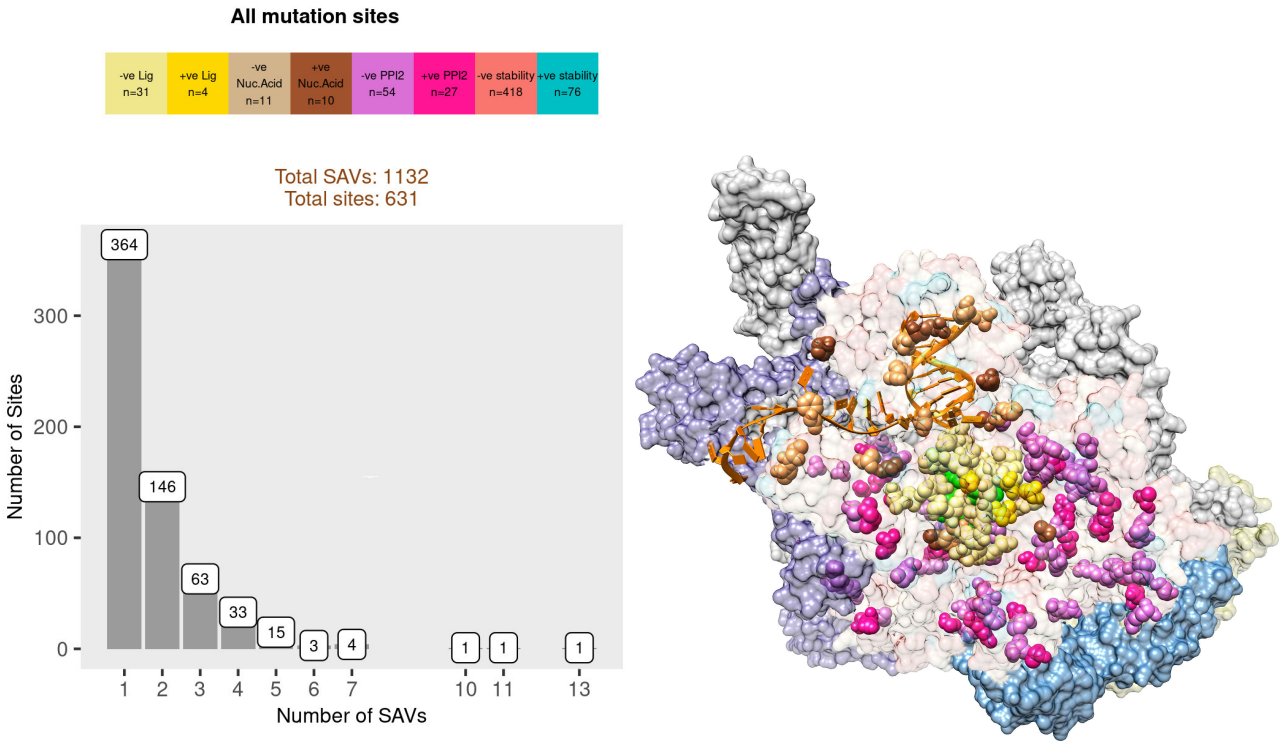


Figure 7: Prominent mutational effects in *M. tuberculosis* RpoB RNA polymerase β subunit Most prominent mutational effect for all 1132 SAVs located in 631 sites prioritised in order of increasing effect size: mCSM/mmCSM-lig, mCSM-NA, mCSM-PPI2, followed by overall stability changes. The left panel shows a barplot displaying the overall frequency of 1132 SAVs with respect to 631 sites, with the coloured bars denoting the site frequency with respect to the most prominent effects. Mutational effects are coloured according to the effect type with brighter colours representing stabilising mutational effects. Sites marked in yellow indicate changes due to ligand (RFP) affinity with light yellow showing destabilising and bright yellow indicating stabilising effect, brown areas indicate changes in nucleic acid (NA) affinity with light brown indicating destabilising and dark brown denoting stabilising effects, pink areas indicate changes due to PPI binding affinity with bright pink highlighting stabilising and light pink areas indicating destabilising mutational effects. Protomer stability changes are coloured with blue indicating stabilising and red indicating destabilising mutational consequences. The corresponding number of mutation sites contributing to these changes are indicated in the text box at the top, and coloured accordingly. The barplot figures are generated using R statistical software version 4.0.4, ggplot2 package. The structure figures are generated using Chimera version 1.14. Abbreviations used: Å: angstroms, $\Delta\Delta$G: change in Gibbs free energy in kcal/mol, SAV: single amino acid variation, RFP: rifampicin.

8.2.4 Mutational association with RFP resistance and flexibility

Resistant mutations appear to be concentrated around RFP and the PPI areas as well as in low-to-mild flexibility regions

Mutational association with resistance according to aggregate DST data showed a minority (30%, n=329) of mutations as resistant. Mutational sites were mapped onto the RpoB RNAP to highlight the location of sites with exclusively resistant (red) and sensitive (blue) mutations while sites displaying both resistant and sensitive mutations were coloured purple. For RpoB RNAP, there were 120 sites with resistant mutations, 201 sites with sensitive mutations, and 86 sites with both resistant and sensitive mutations **(Figure 8A)**.

Resistant mutations appear to cluster around sites close to RFP, NA and the PPI, while sensitive

mutations were spread across the structure **(Figure 8A** and **8B)**. ConSurf scores are calculated for each site on the protein, and range from 1 (rapidly evolving, variable sites) to 9 (slowly evolving, conserved sites). Most mutations (n=270) occurred in the highly variable regions of RpoB RNAP (ConSurf score 1) **(Figure 8B** right panel**)**, in line with observation that sensitive mutations were distributed across the RpoB RNAP **(Figure 8A)**. As such resistant mutations were located in the conserved regions surrounding RFP **(Figure 8B** left panel**)**.

Further, the local flexibility in RpoB RNAP in relation to RFP resistance was also analysed with thickness of the ribbon/tube (thinthick) corresponding to the extent of flexibility. Normal mode analysis of RpoB RNAP highlighted that regions associated with SAVs were in low-mild flexibility **(Figure 9** left panel**)** and the key active site residue H445 (site with the highest SAV diversity: 13 SAVs) and S450 (most frequently occurring mutation) were regions of low flexibility **(Figure 9** right panel**)**. All mutations at H445 were resistant, while S450 had both sensitive and resistant mutations. Other active site residues S428, Q429, S431, D435 and R459 were also regions of mild flexibility **(Figure 9** right panel**)**. Non-active site residues S458 and N437 showing mild flexibility contained a single sensitive mutation (S458T) and five resistant mutations (N437D, N437H, N437S, N437T, N437Y) respectively **(Figure 9** left panel**)**.

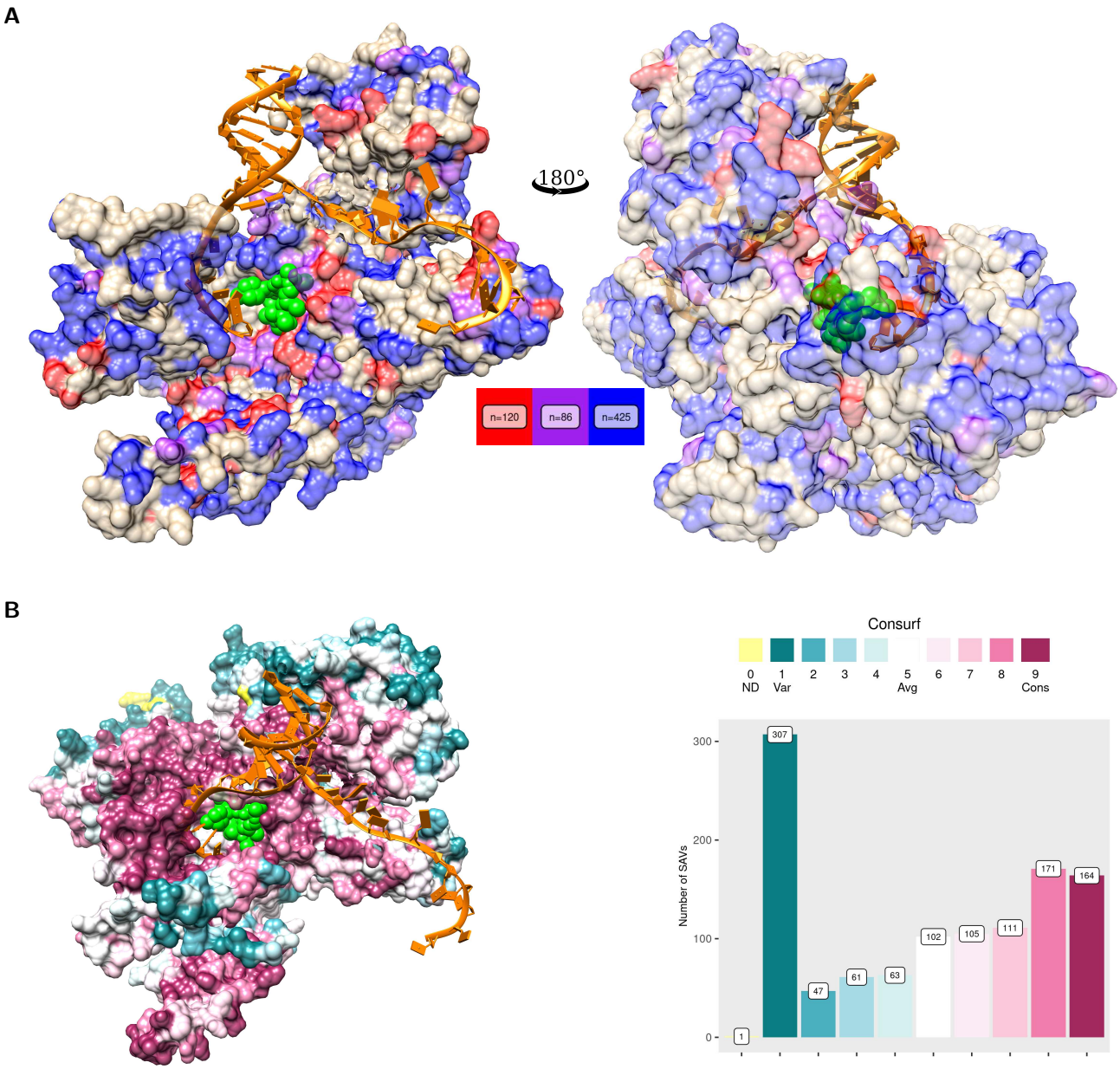


Figure 8: Mutational association with rifampicin resistance and evolutionary conservation in *M. tuberculosis* RpoB RNA polymerase β subunit

Mutational landscape of *M. tuberculosis* RpoB RNA polymerase β subunit according to different measures where **A)** All sites associated with SAVs on *M. tuberculosis* RpoB RNA polymerase β subunit (chain C, PDB-ID: 5UHC) appearing as surface representation in tan colour). RFP appears in green either as spheres or ball-and-stick representation, while nucleic acid (NA) is shown in orange to aid visibility, **A)** The left panel shows all mutational sites associated with resistant (red, n=120 sites), sensitive (blue, n=425 sites), while common sites with both resistant and sensitive mutations appear in purple (n=86). The corresponding right panel depicts the structure rotated by 180°, **B)** Left panel shows RpoB RNA polymerase β subunit, chain C coloured according to ConSurf scores where maroon indicates conserved sites and teal indicates variable sites with RFP (green spheres) located in the conserved binding pocket. Yellow areas reflect sites with uncertainty due to insufficient data for ConSurf score calculation. The barplot on the right panel shows the the number of mutations associated with ConSurf values that range from 1 (variable) in teal to 9 (conserved) in maroon, where 0 denotes insufficient data/not defined (ND). The barplot figures are generated using R statistical software version 4.0.4, ggplot2 package. All structure figures were generated using UCSF Chimera version 1.14. Abbreviations used: SAV: single amino acid variation, RFP: rifampicin.

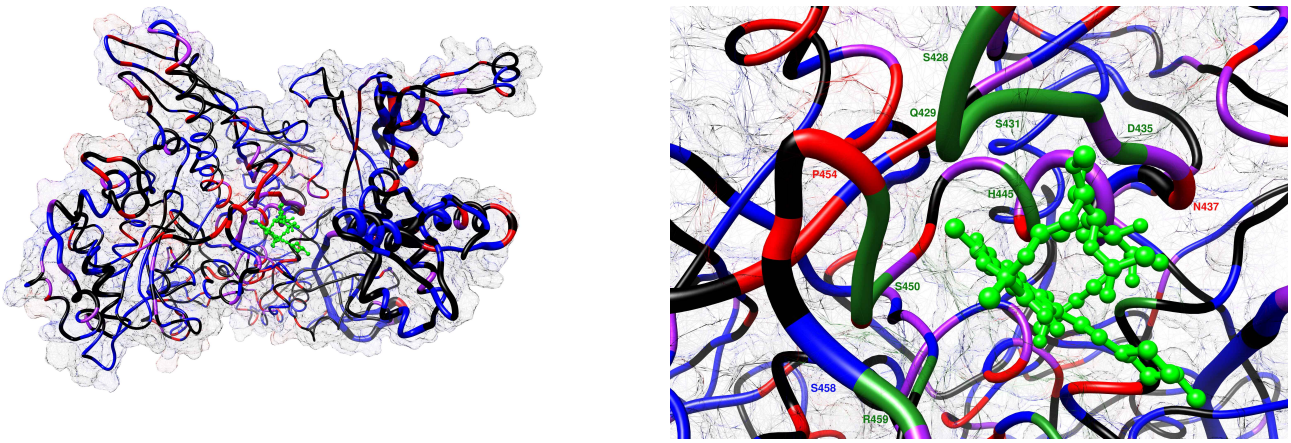


Figure 9: Mutational association with rifampicin resistance and local protein flexibility of *M. tuberculosis* RpoB RNA polymerase β subunit
Mutational landscape of *M. tuberculosis* RpoB RNA polymerase β subunit according to flexibility in RpoB RNA polymerase β subunit, chain C according to normal mode analysis (NMA), measuring atomic deformation according to protein dynamics to denote flexibility associated at sites in RpoB RNA polymerase β subunit. The magnitude of flexibility is represented from thin (low flexibility) to thick (high flexibility) tubes. Left panel: The tubes are further coloured to show mutational association with RFP resistance, red: resistant sites, blue: sensitive sites, purple: shared sites, black: sites with no SAVs, Right panel: Close up view of the RFP binding pocket to denote RFP interacting residues associated with mild-to-moderate flexibility labelled using the standard one-letter amino acid code for the wild-type residue. All structure figures were generated using UCSF Chimera version 1.14. Abbreviations used: SAV: single amino acid variation, RFP: rifampicin.

8.2.5 Relating mutational frequency and biophysical and evolutionary conservational changes

Correlation analysis was performed to understand the relationship between frequently occurring mutations as assessed by MAF and their association with stability (mCSM-DUET, FoldX, DeepDDG, Dynamut2), conservation (ConSurf, SNAP2, PROVEAN), affinity changes (mCSM-lig/mmCSM-lig, mCSM-NA, and mCSM-PPI2), and distances to ligand (Lig-Dist), nucleic acid (NA-Dist) and protein-protein interface (PPI-Dist). A combined analysis with all mutations, as well as separately for resistant (R) and sensitive (S) mutations was performed **(Figures 10 and 11)**. Analyses focused on determining the strength of association without regard for the direction of the association due to dissimilarity of threshold criteria used by the various estimators.

Frequently occurring sensitive mutations were weakly related to protomer stability changes, while frequently occurring resistant mutations were weakly related to distance from nucleic acid, and moderately related to distance from RFP

Frequently occurring mutations were not related to protomer stability changes (P>0.05) according to mCSM-DUET and Dynamut2, though weak association were noted when assessed by FoldX (ρ_{R+S}<0.1, P<0.01) and DeepDDG (ρ_{R+S}=0.1, P<0.001). FoldX showed a weak link with sensitive mutations (ρ_{S}=-0.14, P<0.001), similarly to DeepDDG which also showed weak associations for sensitive mutations (ρ_{S}=0.27, P<0.001) **(Figure 10)**. Frequently occurring mutations were overall weakly

associated with distance from the drug (ρ_{R+S}=-0.11, P<0.001), though resistant mutations driving this association were moderately associated (ρ_{R}=-0.34, P<0.001). Mutational frequency was not associated with distance to nucleic-acid (ρ_{R+S}<0.1, P<0.05) with only weak association for resistant mutations noted ($\rho_{R<}$=-0.15, P<0.01). Mutational frequency was not associated with distance to the dimer interface (ρ_{R+S}<0.1 and $\rho_{R/S}$<0.1, P>0.05).

The different computational tools showed good consensus (moderate to strong associations) amongst their predicted estimates, both overall (0.3≤ρ_{R+S}<0.8, P<0.001), as well as individually for resistant and sensitive mutation groups (0.3≤$\rho_{R/S}$<0.8, P<0.001). As expected, mCSM-DUET and Dynamut2 were strongly correlated as these tools share common methodology (ρ=0.74, P<0.001) **(Figure 10)**. Of note, the negative sign associated with FoldX correlations with other estimators is due to the inverse classification criteria used by these tools (See Chapter 2: Methods for details).

Frequently occurring resistant mutations were weakly associated with evolutionary conservation, while frequently occurring sensitive mutations were moderately associated with protein functional effects

Overall, there was no association with mutational frequency and rate of evolution estimated from ConSurf (ρ_{R+S}<0.1, P>0.05), though resistant mutations showed a weak association (ρ_{R}=-0.17, P<0.01). Frequently occurring mutations were weakly associated with functional changes in protein as estimated by SNAP2 and PROVEAN (ρ_{R+S}<0.3, P<0.001) with both sensitive and resistant mutations contributing to this association (0.1 <$\rho_{R/S}$<0.3, P<0.001). There was good agreement (moderate to strong association) between estimates across the three conservation predictors both overall (ρ_{R+S}>0.6, P<0.001) and individually in sensitive and resistant mutation groups ($\rho_{R/S}$>0.5, P<0.001) **(Figure 11** left panel**)**.

Frequently occurring sensitive mutations were weakly related to PPI affinity changes, while frequently occurring resistant mutations were weakly related to RFP affinity changes

Frequently occurring mutations were weakly related to RFP affinity changes according to mmCSM-lig (ρ_{R+S}=-0.15, P<0.001), but not according to mCSM-lig (ρ_{R+S}<0.1, P<0.01). Further, only resistant mutations appeared to drive this weak association for mCSM (ρ_{R}=-0.15, P<0.01) and moderate association for mmCSM-lig (ρ_{R}:-0.31, P<0.001) **(Figure 11** right panel**)**. No association between mutational frequency and NA affinity was noted, either overall or individually in sensitive and resistant mutation groups (ρ_{R+S} and $\rho_{R/S}$<0.1, P>0.05). However, weak associations were noted with PP affinity changes for sensitive mutations (ρ_{s}=0.13, P<0.001) **(Figure 11** right panel**)**. Estimates from mCSM- and mmCSM-lig were not as strongly correlated (ρ_{R+S}<0.6, $\rho_{R/S}$<0.5, P<0.001) despite the underlying shared methodology **(Figure 11** right panel**)**.

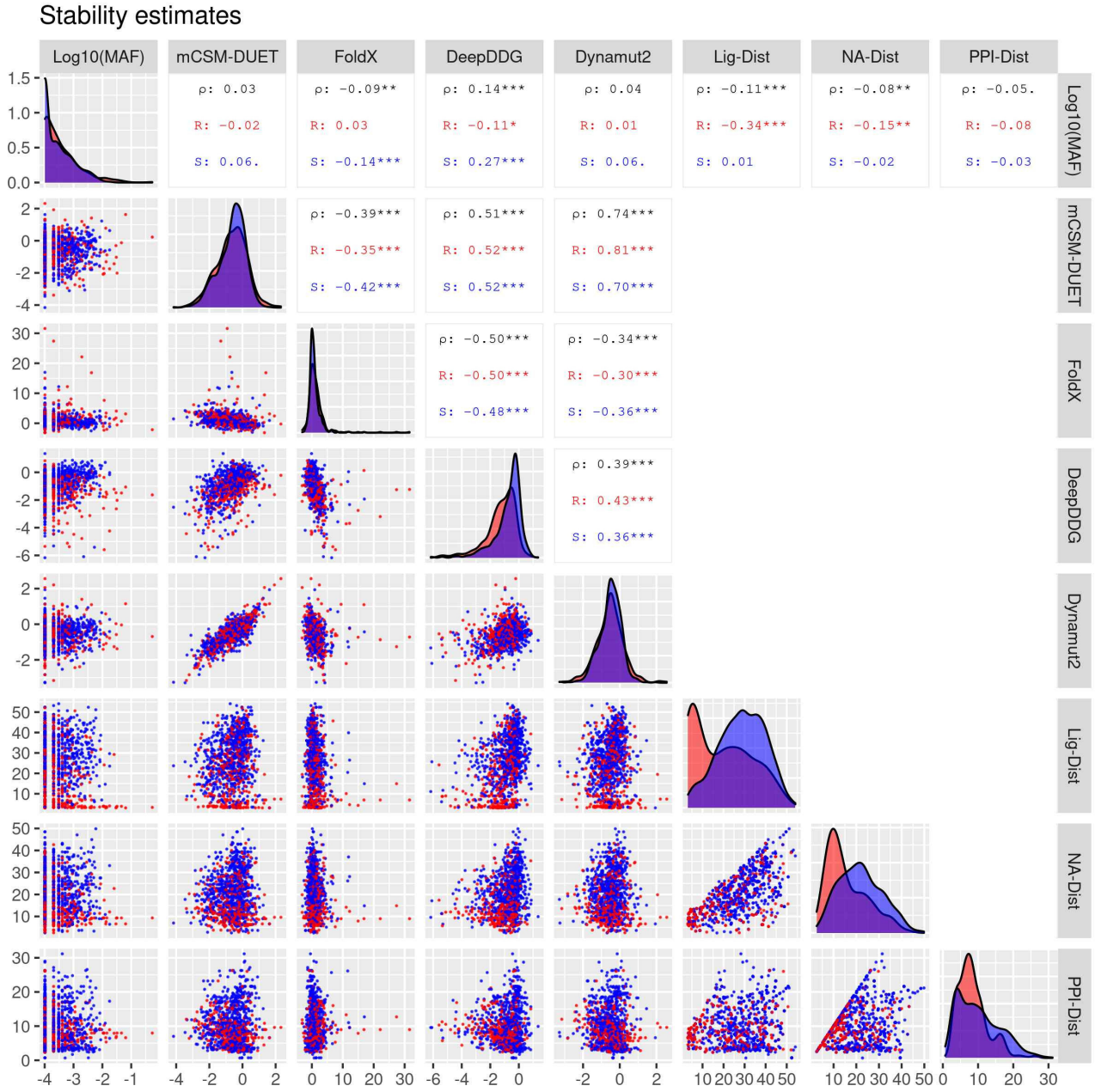


Figure 10: Correlation of protein stability changes and genomics measures
Pairwise correlations between minor allele frequency (MAF), protein stability changes (ΔΔG) estimated using DUET, FoldX, DeepDDG, and Dynamut2, and distance to RFP, and the protein-protein interface for 1132 SAVs. The upper panel in both plots include the pairwise Spearman (ρ) correlation values along with their statistical significance (.P<0.10, *P<0.05, **P<0.01, ***P<0.001). Three correlation values appear in each plot where black denotes the overall correlation with both resistant (R) and sensitive (S) mutations, while red denotes correlation estimates for resistant mutations, and blue denotes correlation estimates for sensitive mutations. The points in the lower panel represent SAVs, where red dots denote resistant mutations and blue represent sensitive mutations individually. The diagonal in each plot displays the density distribution of the corresponding parameter split by the two mutation groups. The figure is generated using R statistical software version 4.0.4, ggplot2 package. Abbreviations used: Å: Angstroms, ΔΔG: change in Gibbs free energy in Kcal/mol, SAV: single amino acid variation, Lig-Dist: distance to ligand in Å, NA-Dist: distance to nucleic-acid in Å, PPI-Dist: distance to protein-protein interface in Å, RFP: rifampicin.

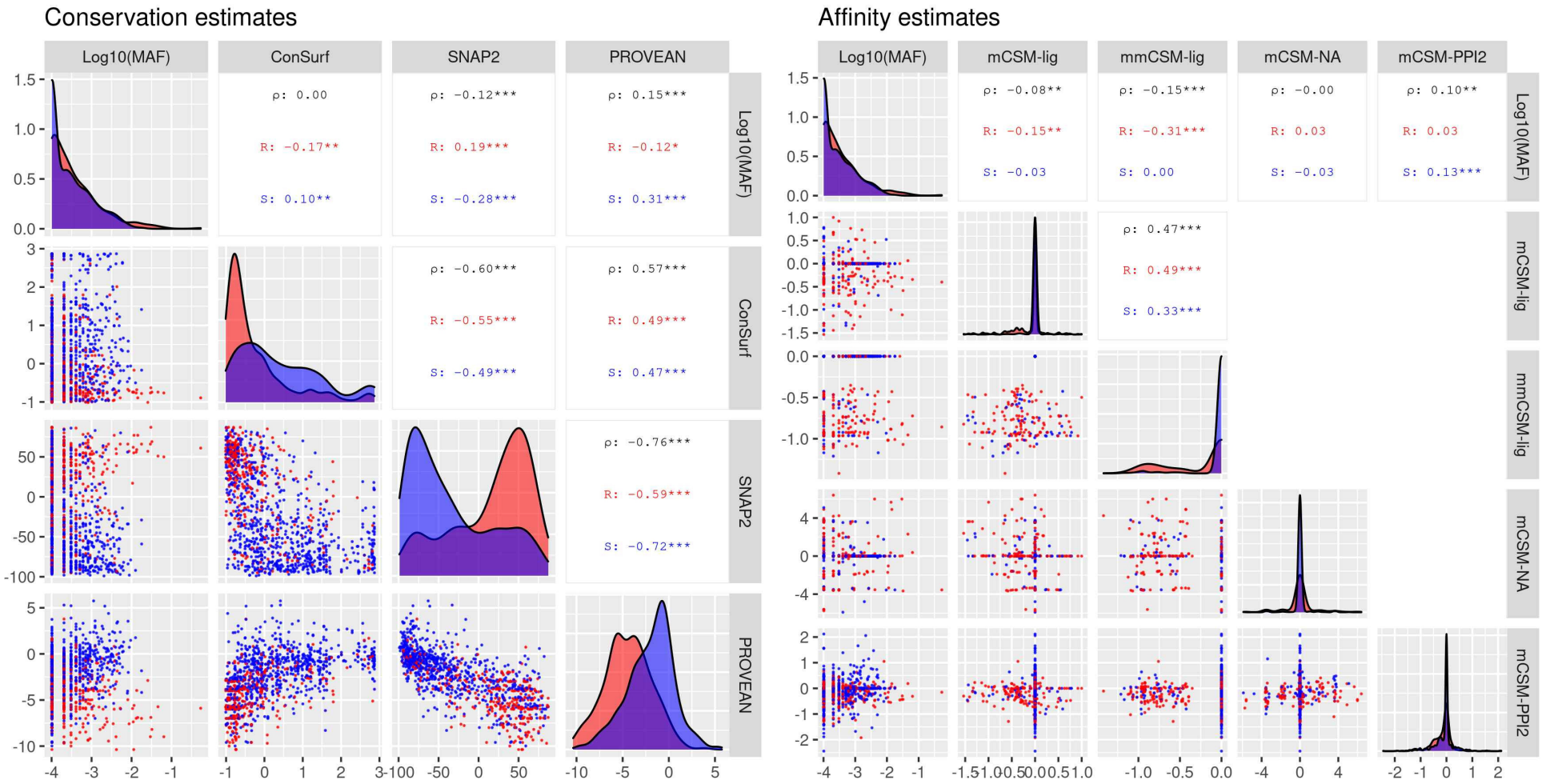


Figure 11: Correlation of evolutionary conservation, affinity changes, and genomics measures

Pairwise correlations of evolutionary conservation, affinity changes, and genomic measure of minor allele frequency (MAF) for 1132 SAVs. **Left panel:** Evolutionary conservation predictors: ConSurf, SNAP2, and PROVEAN, **Right panel:** RFP binding affinity changes estimated as log fold change (mCSM-lig and mmCSM-lig) of 54 SAVs lying within 10Å of RFP, nucleic acid (NA) affinity changes (mCSM-NA) estimated as ΔΔG 195 mutations within 10Å of NA, and protein-protein (PP) binding affinity changes (mCSM-PPI2) estimated as ΔΔG for 674 SAVs lying within 10Å of PPI. All corresponding affinity measures for mutations located more than 10Å of RFP, and the PPI were given a value of 0 to allow complete SAVs to be used for analysis, while respecting the distance threshold for the respective tools. The upper panel in both plots include the pairwise Spearman (ρ) correlation values along with their statistical significance (.P<0.10, *P<0.05, **P<0.01, ***P<0.001). Three correlation values appear in each plot where black denotes the overall correlation with both resistant (R) and sensitive (S) mutations, while red denotes correlation estimates for resistant mutations, and blue denotes correlation estimates for sensitive mutations. The points in the lower panel represent SAVs, where red dots denote resistant mutations and blue represent sensitive mutations individually. The diagonal in each plot displays the density distribution of the corresponding parameter split by the two mutation groups. The figure is generated using R statistical software version 4.0.4, ggplot2 package. Abbreviations used: Å: Angstroms, ΔΔG: change in Gibbs free energy in Kcal/mol, SAV: single amino acid variation, Lig-Dist: distance to ligand in Å, NA-Dist: distance to nucleic-acid in Å, PPI-Dist: distance to protein-protein interface in Å, RFP: rifampicin.

8.2.6 Comparing resistant and sensitive mutations

Resistant mutations occur marginally less frequently, are closer to RFP, NA and PPI, are destabilising for protomer stability, and also likely to affect protein function

Resistant mutations were destabilising compared with sensitive mutations for changes in protomer stability as measured by FoldX (P<0.001) and DeepDDG (P<0.0001), but not by mCSM-DUET and Dynamut2 **(Figures 12A-12D)**. Resistant mutations were only slightly less frequent compared with sensitive mutations (P<0.05, **Figure 12E)**, were located significantly closer to RFP (P<0.0001, **Figure 12F)** without affecting drug binding affinity (P<0.05, **Figures 12L** and **12M)**. Further, resistant mutations were located closer to NA (P<0.0001, **Figure 12G)** and PPI (P<0.001, **Figure 12H)** resulting in marginal reduction in affinity of interaction at the PPI (P<0.05, **Figure 12O)**, but not to NA binding affinity (P>0.05, **Figure 12N)**. Resistant mutations were conserved (slower rate of evolution according to ConSurf) (P<0.0001, **Figure 12I)**, and were more likely to result in deleterious impact towards protein function when assessed by both PROVEAN and SNAP2 (P<0.0001, **Figures 12J** and **12K)**.

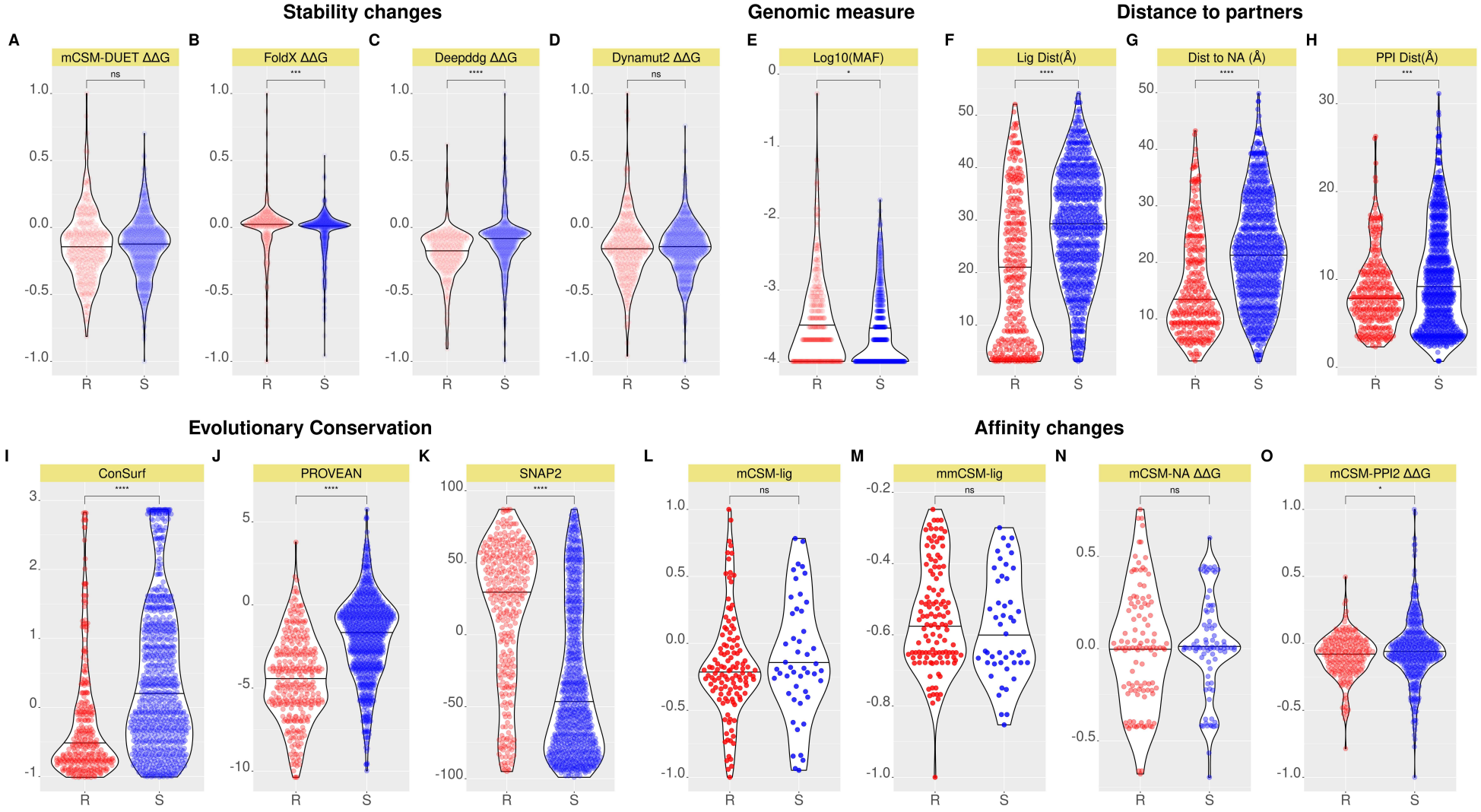

Stability changes
Genomic measure
Distance to partners
A
mCSM-DUET ΔΔG
ns
B
FoldX ΔΔG
***
C
Deepddg ΔΔG
****
D
Dynamut2 ΔΔG
ns
E
Log10(MAF)
*
F
Lig Dist(Å)
****
G
Dist to NA (Å)
****
H
PPI Dist(Å)
***
R
S
Evolutionary Conservation
Affinity changes
I
ConSurf
****
J
PROVEAN
****
K
SNAP2
****
L
mCSM-lig
ns
M
mmCSM-lig
ns
N
mCSM-NA ΔΔG
ns
O
mCSM-PPI2 ΔΔG
*

Figure 12: Comparison of resistant (R) and sensitive (S) mutations
Violin plots showing the distribution of features related to structural properties, genomic measure, evolutionary conservation for 1132 RpoB RNA polymerase β subunit SAVs. For affinity changes related to the ligand (RFP) binding affinity measured by mCSM- and mmCSM-lig, only mutations within 10Å of RFP (n=168) were considered. Similarly, for nucleic acid (NA) binding affinity, mutations within 10Å of nucleic acid (NA) estimated by mCSM-NA (n=195), and for protein-protein (PP) affinity changes estimated by mCSM-PPI2, mutations within 10Å of the PPI (n=674) were considered. Mutations were grouped as either resistant (R, n=127) or sensitive (S, n=731) and were compared using the Wilcoxon rank-sum (unpaired) test, with statistical significance indicated as: *P < 0.05, **P < 0.01, ***P < 0.001, ****P < 0.0001). Mutations in the resistant group appear as red dots, while those in the sensitive group appear as blue dots, and the horizontal line in the violin plots display the median value. The two mutations groups were compared based on **A-D)** Stability changes ($\Delta\Delta$G) estimated from four computational tools: mCSM-DUET, FoldX, DeepDDG and Dynamut2, **E)** genomic measure of average mutational occurrence (Log10MAF), **F-H)** Distance to ligand (Lig-Dist), Distance to Nucleic acid (Distance to NA), and Distance to the PPI (PPI-Dist), **I-K)** Evolutionary conservation measured by ConSurf (<0: Conserved, >0: Variable), PROVEAN >-2.5: Neutral, < -2.5: Deleterious) and SNAP2 (<=0: Neutral, >0: Effect) computational tools, **L-M)** Comparison of STR binding affinity changes from mCSM-lig and mmCSM-lig measured as log fold change for R (n=120) and S (n=48) mutations, **N)**, those for NA binding affinity changes (mCSM-NA) measured as $\Delta\Delta$G for R (n=110) and S (n=85) mutations, and for **O)** PP binding affinity changes (mCSM-PPI2) measured as $\Delta\Delta$G for R (n=17) and S (n=104) mutations. The figure is generated using R statistical software version 4.0.4. Abbreviations used: Å: Angstroms, $\Delta\Delta$G: change in Gibbs free energy in Kcal/mol, SAV: single amino acid variation, ns: not-significant, Lig-Dist: distance to ligand in Å, NA-Dist: distance to nucleic-acid in Å, PPI-Dist: distance to protein-protein interface/s in Å, RFP: rifampicin, MAF: minor allele frequency, R: resistant mutations, S: sensitive mutations.

8.2.7 Associating mutations with Odds Ratio and extreme effects

Mutations involving and surrounding the active site are associated with high OR and RFP resistance

Based on DST data available for 793 (out of 1132) SAVs, mutational association with resistance was further estimated using Odds Ratio (OR), with values above 1 suggesting association with RFP resistance. The higher the OR, the greater the likelihood of a given mutation being resistant. This resulted in a majority (77%, n=611/793) of mutations predicted to be associated with RFP resistance, much higher than observed in our data (30%, n=329/1132).

An overview of mutations in RpoB RNAP show that mutations involving active site residues (H445, D435, S450) and those within 10Å of RFP were the ones with the strongest association with RFP resistance: H445D (OR=1038.14), D435V (OR=863.35), and S450L (OR=573.58) **(Figure 13)**. These were followed by mutations L731P (OR=372.02), V170F (OR=297.78), R827C (OR=292.85), V534M (OR=199.42), I480V (OR=150.51) though not directly involved with the active site were located within 10Å of RFP **(Figure 13)**.

Mutations at active site residues H445D, S450L, and I491T occurred most frequently, showed the strongest link to RFP resistance and reduction in binding affinity to RFP

The most frequently occurring mutation S450L (MAF ∼54%) was located close to RFP (4Å), NA (9Å), and PPI (9Å). The most destabilising mutation for RFP binding affinity was active site residue I491T, while T444P was the most stabilising mutation for RFP binding affinity. Mutations with

other extreme effects like those affecting protomer stability (V168G: destabilising, T702I: stabilising), NA affinity (F367L: destabilising, R225W: stabilising), PPI (Y1073S: destabilising, I717Y: stabilising) were not involved with the active site **(Table 1)**.

Mutation	Mutational effect	Mutational effect value	Lig-Dist (Å)	NA-Dist (Å)	PPI-Dist (Å)	Interacting partner
H445D	Mutation with highest OR	OR = 1038.14	3.85	9.32	7.97	drug
S450L	Most frequent mutation	MAF (%) = 53.66	3.32	9.34	7.97	drug
V168G	Most Destabilising for protomer	ΔΔG = -0.59	7.19	13.59	12.01	no
T702I	Most Stabilising for protomer	ΔΔG = 0.55	19.48	19.87	10.82	no
I491T	Most Destabilising for RFP binding affinity	Log fold change = -0.83	3.62	7.59	5.35	drug
T444P	Most Stabilising for RFP binding affinity	Log fold change = 0.32	5.96	11.06	10.15	no
F367L	Most Destabilising for NA binding affinity	ΔΔG = -5.90	14.56	7.82	7.82	no
R225W	Most Stabilising for NA binding affinity	ΔΔG = 6.41	29.40	7.4	7.40	no
Y1073S	Most Destabilising for PPI affinity	ΔΔG = -2.45	27.35	16.38	2.98	no
I717Y	Most Stabilising for PPI affinity	ΔΔG = 2.12	28.14	20.25	3.15	no

Table 1: Mutations with extreme effects
Mutations (SAVs) with extreme effects related to Odds Ratio (OR), mutational frequency (MAF), stability and affinity changes. For affinity changes only mutations within 10Å of RFP for RFP binding affinity, NA for NA binding affinity, and Protein-Protein Interface (PPI) for PPI affinity were considered. The protomer stability changes are the average effect of all four estimates (mCSM-DUET, FoldX, DeepDDG and Dynamut2) combined, and the RFP binding affinity changes are the average effect of the two mCSM based tools (mCSM-lig and mmCSM-lig) combined. Changes in NA binding affinity and PP affinity correspond to estimates from mCSM-NA and mCSM-PPI respectively. The estimated effects were categorised as Destabilising (log fold affinity change/ΔΔG<0) and Stabilising (log fold affinity change/ΔΔG>0). Abbreviations used: Å: Angstroms, ΔΔG: change in Gibbs free energy in Kcal/mol, MAF: minor allele frequency, SAV: single amino acid variation, Lig-Dist: distance to ligand, PPI-Dist: distance to protein-protein interface, RFP: rifampicin
.

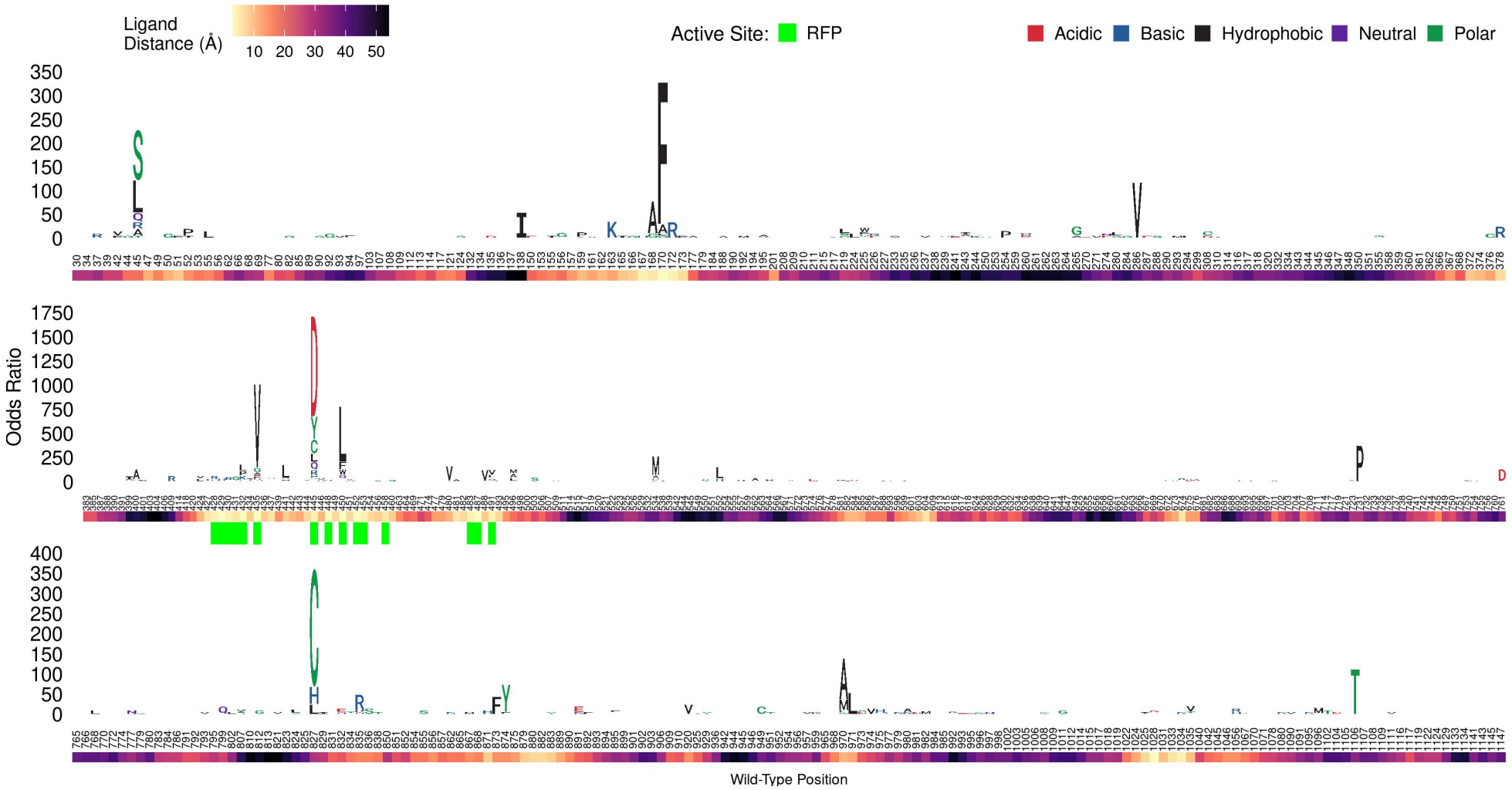


Figure 13: Logo plot showing mutational sites and their association with resistance according to Odds Ratio

Logo plot showing 793 SAVs by mutational site according to their association with RFP resistance calculated using Odds Ratio (OR). The vertical axis represents the OR where letters denote mutant residues which are proportional to their corresponding OR highlighting the most resistant mutation at each site and overall. The mutant residues are coloured according to the amino acid (aa) properties as denoted where red denotes acidic aa, basic aa appear in blue, hydrophobic aa in black, neutral aa in purple, and polar aa in darkgreen. The structural positions associated with SAVs with OR are indicated on the horizontal axis. The heat bar underneath the positions indicate the distance of that position from RFP according to the magma colour gradient where light yellow indicates sites closer to RFP (ligand distance in Angstroms). The positions are further annotated to reflect residues involved in interacting with RFP (green). The figure is generated using R statistical software version 4.0.2, ggplot2 package. Abbreviations used: SAV: single amino acid variation, RFP: rifampicin.

8.2.8 Relating lineage and protomer stability

Lineages 1 and 3 have high SAV diversity, with resistant mutations marginally stabilising for protomer stability

Nearly 45% of clinical isolates (n=15,898) consisted of SAVs in the protein coding region of *rpoB*, with 14,061 samples contributing to the four main *M. tuberculosis* lineages (Lineages 1-4). Most samples with RpoB RNAP mutations belonged to lineage 4 (n=6,861), followed by lineage 2 (n=5,121), lineage 3 (n=1,341) and finally by lineage 1 with the least number of samples (n=737) **(Figure 14A)**. However, lineages 1 and 3 displayed higher SAV diversity (Lineage 1: 26%, n=188; Lineage 3: 21%, n=486), compared with lineages 4 and 2 (Lineage 4: 11%, n=735; Lineage 2:25%, n=486) **(Figure 14B)**.

Resistant mutations for all lineages showed prominent peaks centred around ΔΔG 0.15 Kcal/mol corresponding to mildly stabilising protomer effects, with a smaller peak around ΔΔG 0.5 Kcal/mol corresponding to moderately stabilising protomer effects. Lineage 3 showed an additional peak with a ΔΔG -0.15 Kcal/mol corresponding to mildly destabilising protomer effects **(Figure 14C)**.

Sensitive mutations across all lineages were widely distributed around the moderately destabilising-to-mildly stabilising protomer effects (-0.3 Kcal/mol <ΔΔG < 0.5 Kcal/mol). Lineages 2 and 3 peaked with a ΔΔG of -0.12 Kcal/mol corresponding to marginally destabilising the protomer, with lineage 1 displaying a smaller peak with a ΔΔG of 0.35 Kcal/mol corresponding to moderately stabilising the protomer **(Figure 14C)**. Overall the distributions were significantly different between all lineages (adjusted P<0.0001), as well as in lineages between resistant and sensitive mutations (adjusted P<0.0001) except for lineage 1 (adjusted P>0.05) (Appendix Table 8.D.1).

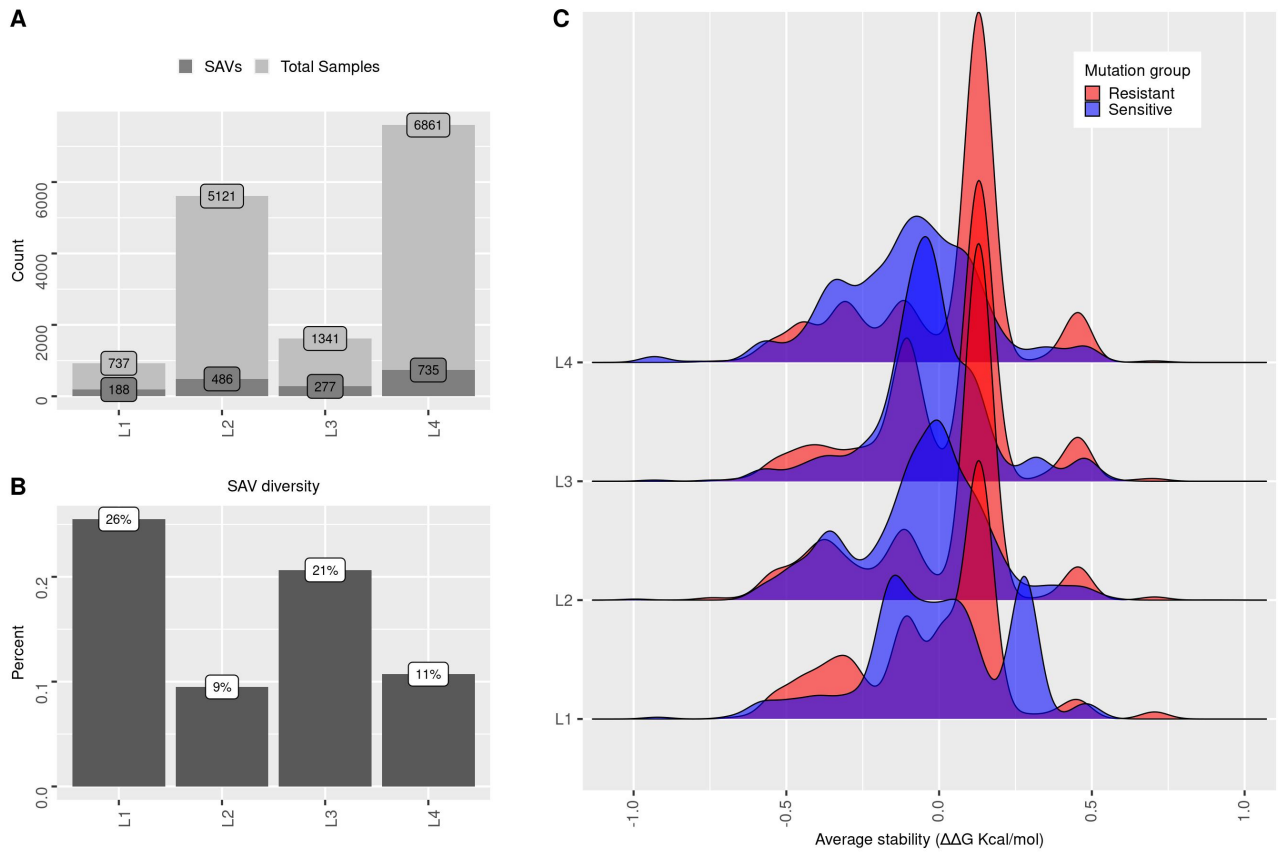


Figure 14: Lineage and Protomer stability distribution
Total number of samples (n=14,061) along with the number of mutations associated with RFP resistance in the four *M. tuberculosis* lineages (L1-L4). **A)** The dark grey bars show the number of mutations (SAVs), while the light grey bar show the total number of samples in each lineage, **B)** Mutational diversity in each lineage, **C)** Density distribution of lineages according to protein stability changes (ΔΔG). Estimates from four different computational tools: mCSM-DUET, FoldX, DeepDDG,and Dynamut2 were combined to calculate the average mutational stability impact for each SAV. The horizontal axis shows the average stability values (-1: highly destabilising and +1:highly stabilising) further coloured by mutational association with resistance: Red denotes resistant mutations (n=127, from 6,878 samples) and blue indicates sensitive mutations (n=731, from 6,657 samples) where the same sample may have mutations with differing sensitivities. The figure is generated using R statistical software version 4.0.4. Abbreviations used:ΔΔG: change in Gibbs free energy, SAV: single amino acid variation, RFP: rifampicin.

8.3 Chapter summary

Mutations in RpoB RNAP are prevalent in *M. tuberculosis*, with mutations at active site residues all contributing to RFP resistance. Most mutational consequences destabilise protomer stability with possible consequences on protein function, with frequently occurring mutations being weakly associated with evolutionary conservation. Resistant mutations appear to be concentrated around RFP, NA, and the PPI, as well as being located in regions of low-to-mild flexibility. Consequently, mutations reduce RFP binding affinity, as well as PP affinity, but do not affect NA binding affinity, indicating that mutations at the NA binding sites result in deleterious consequences with fitness costs being ameliorated by other factors. This suggests that mutations in the active site show a gain-of-function survival mechanism, able to tolerate extreme mutational consequences around RFP, NA, and the PPI

without a severe fitness penalty. As part of the larger hetero-hexamer complex, this large mutational tolerance, without compromising function, is likely to come from compensatory effects from *rpoC* and *rpoA* genes.

References

[1] Timothy R. Sterling. "Guidelines for the Treatment of Latent Tuberculosis Infection: Recommendations from the National Tuberculosis Controllers Association and CDC, 2020". In: *MMWR. Recommendations and Reports* 69 (2020). ISSN: 1057-59871545-8601. DOI: 10.15585/mmwr.rr6901a1.

[2] WHO consolidated guidelines DR-TB. *WHO Consolidated Guidelines on Drug-Resistant Tuberculosis Treatment*. Geneva: World Health Organization, 2019.

[3] M. E. Levin and G. F. Hatfull. "Mycobacterium Smegmatis RNA Polymerase: DNA Supercoiling, Action of Rifampicin and Mechanism of Rifampicin Resistance". In: *Molecular Microbiology* 8.2 (Apr. 1993), pp. 277–285. ISSN: 0950-382X. DOI: 10.1111/j.1365-2958.1993.tb01572.x.

[4] V. Donnabella et al. "Isolation of the Gene for the Beta Subunit of RNA Polymerase from Rifampicin-Resistant Mycobacterium Tuberculosis and Identification of New Mutations". In: *American Journal of Respiratory Cell and Molecular Biology* 11.6 (Dec. 1994), pp. 639–643. ISSN: 1044-1549. DOI: 10.1165/ajrcmb.11.6.7946393.

[5] Giovanni Piccaro et al. "Rifampin Induces Hydroxyl Radical Formation in Mycobacterium Tuberculosis". In: *Antimicrobial Agents and Chemotherapy* 58.12 (Dec. 2014), pp. 7527–7533. ISSN: 1098-6596. DOI: 10.1128/AAC.03169-14.

[6] N. Honore and S. T. Cole. "Molecular Basis of Rifampin Resistance in Mycobacterium Leprae". In: *Antimicrobial Agents and Chemotherapy* 37.3 (Mar. 1993), pp. 414–418. ISSN: 0066-4804. DOI: 10.1128/AAC.37.3.414.

[7] A Telenti. "Genetics and Pulmonary Medicine Bullet 5: Genetics of Drug Resistant Tuberculosis". In: *Thorax* 53.9 (Sept. 1, 1998), pp. 793–797. ISSN: 0040-6376. DOI: 10.1136/thx.53.9.793.

[8] J M Musser. "Antimicrobial Agent Resistance in Mycobacteria: Molecular Genetic Insights". In: *Clinical Microbiology Reviews* 8.4 (Oct. 1995), pp. 496–514. DOI: 10.1128/CMR.8.4.496.

[9] C. Cavusoglu, Y. Karaca-Derici, and A. Bilgic. "In-Vitro Activity of Rifabutin against Rifampicin-Resistant Mycobacterium Tuberculosis Isolates with Known rpoB Mutations". In: *Clinical Microbiology and Infection* 10.7 (July 1, 2004), pp. 662–665. ISSN: 1198-743X. DOI: 10.1111/j.1469-0691.2004.00917.x.

[10] Charambira Kelvin. "Increased Rifampicin Mono-Resistance Prevalence in Zimbabwe. Is the Higher Prevalence of Codon 523 to 529 Mutation in the rpoB Gene an Attributable Factor?" In: S.1 (2018), p. 1.

[11] Stephanie Portelli et al. "Prediction of Rifampicin Resistance beyond the RRDR Using Structure-Based Machine Learning Approaches". In: *Scientific Reports* 10.1 (1 Oct. 22, 2020), p. 18120. ISSN: 2045-2322. DOI: 10.1038/s41598-020-74648-y.

[12] Gilman Kit Hang Siu et al. "Mutations Outside the Rifampicin Resistance-Determining Region Associated with Rifampicin Resistance in Mycobacterium Tuberculosis". In: *The Journal of Antimicrobial Chemotherapy* 66.4 (Apr. 2011), pp. 730–733. ISSN: 1460-2091. DOI: 10.1093/jac/dkq519.

[13] Iñaki Comas et al. "Whole-Genome Sequencing of Rifampicin-Resistant Mycobacterium Tuberculosis Strains Identifies Compensatory Mutations in RNA Polymerase Genes". In: *Nature Genetics* 44.1 (2012), pp. 106–110. ISSN: 10614036. DOI: 10.1038/ng.1038.

[14] M. de Vos et al. "Putative Compensatory Mutations in the rpoC Gene of Rifampin-Resistant Mycobacterium Tuberculosis Are Associated with Ongoing Transmission". In: *Antimicrobial Agents and Chemotherapy* 57.2 (Feb. 2013), pp. 827–832. ISSN: 0066-4804. DOI: 10.1128/AAC.01541-12.

[15] Gerrit Brandis and Diarmaid Hughes. "Genetic Characterization of Compensatory Evolution in Strains Carrying rpoB Ser531Leu, the Rifampicin Resistance Mutation Most Frequently Found in Clinical Isolates". In: *Journal of Antimicrobial Chemotherapy* 68.11 (Nov. 1, 2013), pp. 2493–2497. ISSN: 0305-7453. DOI: 10.1093/jac/dkt224.

[16] Bashir A. Sheikh et al. "Development of New Therapeutics to Meet the Current Challenge of Drug Resistant Tuberculosis". In: *Current Pharmaceutical Biotechnology* 22.4 (Mar. 2021), pp. 480–500. ISSN: 13892010. DOI: 10.2174/1389201021666200628021702.

[17] Wei Lin et al. "Structural Basis of Mycobacterium Tuberculosis Transcription and Transcription Inhibition". In: *Molecular Cell* 66.2 (Apr. 20, 2017), 169–179.e8. ISSN: 1097-2765. DOI: 10.1016/j.molcel.2017.03.001.

Appendix for Chapter 8

8.A Mutations close to rifampicin

Mutation	Interacting partner	Lig-Dist (Å)	mCSM-lig	mCSM-lig outcome	mmCSM-lig	mmCSM-lig outcome	MAF (%)	Odds Ratio	P-value	Adjusted P-value	Adjusted P-value significance
H445D	drug	3.85	-0.58	Destabilising	-1.07	Destabilising	3.23	1,038.14	<0.0001	<0.0001	****
D435V	drug	3.10	-0.34	Destabilising	-0.72	Destabilising	6.40	863.35	<0.0001	<0.0001	****
S450L	drug	3.32	-0.40	Destabilising	-0.86	Destabilising	53.66	573.58	<0.0001	<0.0001	****
V170F	no	3.52	-0.37	Destabilising	-0.76	Destabilising	0.81	297.78	<0.0001	<0.0001	****
H445Y	drug	3.85	0.06	Stabilising	-0.95	Destabilising	4.58	237.23	<0.0001	<0.0001	****
I480V	no	9.00	-0.16	Destabilising	-0.70	Destabilising	0.38	150.51	<0.0001	<0.0001	****
H445C	drug	3.85	-0.44	Destabilising	-0.48	Destabilising	0.42	145.63	<0.0001	<0.0001	****
S441L	no	6.97	0.46	Stabilising	-0.58	Destabilising	0.42	135.87	<0.0001	<0.0001	****
I488V	no	7.18	0.12	Stabilising	-0.92	Destabilising	0.26	111.51	<0.0001	<0.0001	****
H445L	drug	3.85	-0.06	Destabilising	-0.55	Destabilising	1.13	86.80	<0.0001	<0.0001	****
H445Q	drug	3.85	-0.34	Destabilising	-0.94	Destabilising	0.24	82.33	<0.0001	<0.0001	****
S450F	drug	3.32	-0.38	Destabilising	-0.43	Destabilising	0.52	77.69	<0.0001	<0.0001	****
V168A	no	7.19	0.73	Stabilising	-0.79	Destabilising	0.16	67.76	<0.0001	<0.0001	****
Q432L	drug	3.25	-0.47	Destabilising	-0.76	Destabilising	0.23	67.76	<0.0001	<0.0001	****
H445R	drug	3.85	-0.20	Destabilising	-1.02	Destabilising	1.10	57.86	<0.0001	<0.0001	****
S450W	drug	3.32	-0.50	Destabilising	-0.43	Destabilising	1.43	57.07	<0.0001	<0.0001	****
I491V	drug	3.62	-1.15	Destabilising	-0.92	Destabilising	0.14	53.21	<0.0001	<0.0001	****
Q432P	drug	3.25	-0.51	Destabilising	-0.77	Destabilising	0.37	48.45	<0.0001	<0.0001	****
D435G	drug	3.10	-0.30	Destabilising	-0.88	Destabilising	1.31	46.33	<0.0001	<0.0001	****
Q432K	drug	3.25	-0.64	Destabilising	-0.77	Destabilising	0.27	46.02	<0.0001	<0.0001	****
S431G	drug	4.65	-0.11	Destabilising	-1.09	Destabilising	0.11	43.52	<0.0001	<0.001	***

S428R	drug	4.35	-0.87	Destabilising	-0.91	Destabilising	0.10	38.68	<0.0001	<0.001	***
S450Q	drug	3.32	-0.82	Destabilising	-0.61	Destabilising	0.10	38.68	<0.0001	<0.001	***
L430R	drug	4.04	-0.91	Destabilising	-0.94	Destabilising	0.19	36.31	<0.0001	<0.0001	****
Q172R	no	6.26	0.08	Stabilising	-0.78	Destabilising	0.08	33.84	<0.0001	0.01	**
F424V	no	8.77	-0.27	Destabilising	-0.35	Destabilising	0.07	29.00	<0.001	0.01	**
I491L	drug	3.62	-1.12	Destabilising	-0.92	Destabilising	0.07	29.00	<0.001	0.01	**
Q429H	drug	3.44	-1.33	Destabilising	-0.67	Destabilising	0.09	29.00	<0.001	0.01	**
F424L	no	8.77	-0.23	Destabilising	-0.54	Destabilising	0.10	29.00	<0.001	0.01	**
D435A	drug	3.10	-0.31	Destabilising	-0.81	Destabilising	0.19	26.60	<0.0001	<0.001	***
E481A	no	5.39	-0.05	Destabilising	-0.74	Destabilising	0.05	24.16	<0.001	0.02	*
L378R	no	8.71	-0.58	Destabilising	-0.56	Destabilising	0.07	24.16	<0.001	0.02	*
H445P	drug	3.85	-0.19	Destabilising	-0.43	Destabilising	0.08	24.16	<0.001	0.02	*
K446Q	no	8.89	0.52	Stabilising	-0.73	Destabilising	0.11	24.16	<0.001	0.02	*
V170A	no	3.52	-0.31	Destabilising	-0.85	Destabilising	0.04	19.33	0.01	0.05	ns
T427I	no	5.26	-0.38	Destabilising	-0.61	Destabilising	0.05	19.33	0.01	0.05	ns
S428G	drug	4.35	-0.53	Destabilising	-1.09	Destabilising	0.05	19.33	0.01	0.05	ns
E460G	no	6.79	-0.69	Destabilising	-0.88	Destabilising	0.06	19.33	0.01	0.05	ns
D435E	drug	3.10	0.11	Stabilising	-0.92	Destabilising	0.08	19.33	0.01	0.05	ns
H674R	no	4.61	-0.62	Destabilising	-0.62	Destabilising	0.09	19.33	0.01	0.05	ns
D435F	drug	3.10	-0.28	Destabilising	-0.67	Destabilising	0.41	17.77	<0.0001	<0.0001	****
M434I	no	6.61	0.26	Stabilising	-0.92	Destabilising	0.19	15.73	<0.0001	<0.001	***
I491T	drug	3.62	-1.41	Destabilising	-1.06	Destabilising	0.08	14.50	<0.001	0.03	*
S441Q	no	6.97	0.51	Stabilising	-0.72	Destabilising	0.19	14.50	<0.001	0.03	*
S441M	no	6.97	0.29	Stabilising	-0.57	Destabilising	0.03	14.49	0.03	0.15	ns
R448Q	drug	3.14	-0.41	Destabilising	-0.96	Destabilising	0.03	14.49	0.03	0.15	ns
G456S	no	7.41	-0.65	Destabilising	-1.42	Destabilising	0.03	14.49	0.03	0.15	ns

K446R	no	8.89	0.33	Stabilising	-0.73	Destabilising	0.04	14.49	0.03	0.15	ns
I491M	drug	3.62	-1.03	Destabilising	-0.92	Destabilising	0.05	14.49	0.03	0.15	ns
A451V	no	6.18	-0.45	Destabilising	-0.93	Destabilising	0.06	14.49	0.03	0.15	ns
A584D	no	8.97	-0.42	Destabilising	-0.55	Destabilising	0.06	14.49	0.03	0.15	ns
N437D	no	5.34	-0.12	Destabilising	-0.96	Destabilising	0.13	12.09	<0.001	0.01	**
D435N	drug	3.10	-0.28	Destabilising	-0.73	Destabilising	0.06	12.08	0.01	0.07	ns
H445G	drug	3.85	-0.26	Destabilising	-0.40	Destabilising	0.09	12.08	0.01	0.07	ns
L452P	drug	3.46	-1.11	Destabilising	-0.82	Destabilising	3.04	11.41	<0.0001	<0.0001	****
V168G	no	7.19	0.63	Stabilising	-0.92	Destabilising	0.02	9.66	0.09	0.35	ns
Q436P	no	7.38	0.02	Stabilising	-0.53	Destabilising	0.02	9.66	0.09	0.35	ns
T444S	no	5.96	0.68	Stabilising	-0.96	Destabilising	0.02	9.66	0.09	0.35	ns
S450A	drug	3.32	-0.52	Destabilising	-0.94	Destabilising	0.02	9.66	0.09	0.35	ns
A451G	no	6.18	-0.54	Destabilising	-1.06	Destabilising	0.02	9.66	0.09	0.35	ns
P483S	drug	3.56	-1.13	Destabilising	-1.11	Destabilising	0.02	9.66	0.09	0.35	ns
S582A	no	9.26	0.67	Stabilising	-0.72	Destabilising	0.02	9.66	0.09	0.35	ns
N437S	no	5.34	0.18	Stabilising	-0.81	Destabilising	0.03	9.66	0.09	0.35	ns
I480T	no	9.00	-0.57	Destabilising	-0.58	Destabilising	0.03	9.66	0.09	0.35	ns
N437H	no	5.34	0.04	Stabilising	-0.74	Destabilising	0.04	9.66	0.09	0.35	ns
H445T	drug	3.85	-0.37	Destabilising	-0.40	Destabilising	0.04	9.66	0.09	0.35	ns
M434V	no	6.61	0.27	Stabilising	-0.92	Destabilising	0.06	9.66	0.03	0.16	ns
N673S	no	9.91	-0.34	Destabilising	-0.57	Destabilising	0.06	9.66	0.09	0.35	ns
S493L	no	5.80	0.76	Stabilising	-0.86	Destabilising	0.07	9.66	0.03	0.16	ns
Q432E	drug	3.25	-0.64	Destabilising	-0.93	Destabilising	0.09	9.66	0.09	0.35	ns
D435Y	drug	3.10	-0.30	Destabilising	-0.83	Destabilising	2.96	8.92	<0.0001	<0.0001	****
L430P	drug	4.04	-0.28	Destabilising	-1.01	Destabilising	1.78	5.07	<0.0001	<0.0001	****
E166G	no	9.50	-0.08	Destabilising	-0.46	Destabilising	0.01	4.83	0.29	0.58	ns

V170G	no	3.52	-0.42	Destabilising	-0.47	Destabilising	0.01	4.83	0.29	0.58	ns
V170L	no	3.52	-0.13	Destabilising	-0.92	Destabilising	0.01	4.83	0.29	0.58	ns
L173P	no	8.72	-0.02	Destabilising	-1.10	Destabilising	0.01	4.83	0.29	0.58	ns
T427P	no	5.26	-0.36	Destabilising	-0.93	Destabilising	0.01	4.83	0.29	0.58	ns
Q429P	drug	3.44	-1.30	Destabilising	-0.92	Destabilising	0.01	4.83	0.29	0.58	ns
L430V	drug	4.04	-0.31	Destabilising	-0.92	Destabilising	0.01	4.83	0.29	0.58	ns
Q432H	drug	3.25	-0.70	Destabilising	-0.67	Destabilising	0.01	4.83	0.29	0.58	ns
D435H	drug	3.10	-0.39	Destabilising	-0.72	Destabilising	0.01	4.83	0.29	0.58	ns
D435S	drug	3.10	-0.29	Destabilising	-0.78	Destabilising	0.01	4.83	0.29	0.58	ns
Q436L	no	7.38	-0.03	Destabilising	-0.75	Destabilising	0.01	4.83	0.29	0.58	ns
S441V	no	6.97	0.49	Stabilising	-0.97	Destabilising	0.01	4.83	0.29	0.58	ns
G442R	no	8.85	-0.35	Destabilising	-0.94	Destabilising	0.01	4.83	0.29	0.58	ns
H445V	drug	3.85	-0.11	Destabilising	-0.48	Destabilising	0.01	4.83	0.29	0.58	ns
R448K	drug	3.14	-0.39	Destabilising	-0.94	Destabilising	0.01	4.83	0.29	0.58	ns
L449M	no	6.09	-0.24	Destabilising	-0.92	Destabilising	0.01	4.83	0.29	0.58	ns
S450Y	drug	3.32	-0.45	Destabilising	-0.44	Destabilising	0.01	4.83	0.29	0.58	ns
L452S	drug	3.46	-1.34	Destabilising	-0.50	Destabilising	0.01	4.83	0.29	0.58	ns
G453R	drug	4.32	-1.53	Destabilising	-0.82	Destabilising	0.01	4.83	0.29	0.58	ns
P454R	no	6.38	-0.88	Destabilising	-0.69	Destabilising	0.01	4.83	0.29	0.58	ns
N604S	no	4.80	-0.02	Destabilising	-0.81	Destabilising	0.01	4.83	0.29	0.58	ns
N673D	no	9.91	-0.66	Destabilising	-0.72	Destabilising	0.01	4.83	0.29	0.58	ns
H1028D	no	8.77	-0.58	Destabilising	-1.12	Destabilising	0.01	4.83	0.29	0.58	ns
N437Y	no	5.34	0.20	Stabilising	-0.83	Destabilising	0.02	4.83	0.29	0.58	ns
T444P	no	5.96	1.00	Stabilising	-0.50	Destabilising	0.02	4.83	0.29	0.58	ns
S450C	drug	3.32	-0.63	Destabilising	-0.63	Destabilising	0.02	4.83	0.29	0.58	ns
S450G	drug	3.32	-0.61	Destabilising	-1.09	Destabilising	0.02	4.83	0.29	0.58	ns

S450V	drug	3.32	-0.41	Destabilising	-0.94	Destabilising	0.02	4.83	0.29	0.58	ns
I488L	no	7.18	0.11	Stabilising	-0.92	Destabilising	0.02	4.83	0.29	0.58	ns
I491S	drug	3.62	-1.44	Destabilising	-0.47	Destabilising	0.02	4.83	0.29	0.58	ns
S431C	drug	4.65	-0.30	Destabilising	-0.63	Destabilising	0.03	4.83	0.29	0.58	ns
T444I	no	5.96	0.92	Stabilising	-0.97	Destabilising	0.03	4.83	0.29	0.58	ns
H674N	no	4.61	-0.49	Destabilising	-0.59	Destabilising	0.03	4.83	0.29	0.58	ns
H445F	drug	3.85	0.03	Stabilising	-0.97	Destabilising	0.03	4.83	0.21	0.58	ns
M434L	no	6.61	0.26	Stabilising	-0.92	Destabilising	0.04	4.83	0.29	0.58	ns
H674Y	no	4.61	-0.23	Destabilising	-0.80	Destabilising	0.04	4.83	0.21	0.58	ns
R167H	no	5.86	0.31	Stabilising	-0.64	Destabilising	0.07	4.83	0.29	0.58	ns
P454H	no	6.38	-0.67	Destabilising	-0.41	Destabilising	0.07	4.83	0.21	0.58	ns
P454L	no	6.38	-0.27	Destabilising	-0.78	Destabilising	0.08	4.83	0.21	0.58	ns
A584G	no	8.97	0.53	Stabilising	-0.81	Destabilising	0.10	4.03	0.05	0.29	ns
H445S	drug	3.85	-0.41	Destabilising	-0.40	Destabilising	0.16	3.22	0.21	0.58	ns
I491F	drug	3.62	-1.00	Destabilising	-0.88	Destabilising	1.35	3.11	<0.0001	<0.0001	****
H445N	drug	3.85	-0.39	Destabilising	-1.02	Destabilising	1.06	2.66	<0.001	0.01	**
S431T	drug	4.65	-0.27	Destabilising	-0.96	Destabilising	0.02	2.41	0.5	0.86	ns
H674Q	no	4.61	-0.51	Destabilising	-0.60	Destabilising	0.02	2.41	0.5	0.86	ns
N487S	drug	3.33	-1.28	Destabilising	-0.81	Destabilising	0.05	2.41	0.5	0.86	ns
S431R	drug	4.65	-1.06	Destabilising	-0.47	Destabilising	0.06	2.41	0.59	0.94	ns
I491N	drug	3.62	-1.45	Destabilising	-0.72	Destabilising	0.01	1.21	>1	>1	ns
N437T	no	5.34	0.17	Stabilising	-0.87	Destabilising	0.03	1.21	>1	>1	ns
P483L	drug	3.56	-0.87	Destabilising	-0.78	Destabilising	0.03	1.21	>1	>1	ns
H1028L	no	8.77	-0.13	Destabilising	-0.55	Destabilising	0.03	1.21	>1	>1	ns
T427S	no	5.26	-0.11	Destabilising	-0.96	Destabilising	0.04	1.21	>1	>1	ns
S428T	drug	4.35	-0.50	Destabilising	-0.96	Destabilising	0.04	1.21	>1	>1	ns

H1028N	no	8.77	-0.29	Destabilising	-0.53	Destabilising	0.04	1.21	>1	>1		ns
H1028R	no	8.77	-0.28	Destabilising	-0.78	Destabilising	0.15	1.21	>1	>1		ns
A603S	no	7.98	-0.01	Destabilising	-0.99	Destabilising	0.24	0.60	>1	>1		ns
T482S	no	5.58	0.01	Stabilising	-0.96	Destabilising	0.08	0.48	0.68	>1		ns
G442E	no	8.85	-0.38	Destabilising	-0.47	Destabilising	0.24	0.48	0.68	>1		ns
S450N	drug	3.32	-0.92	Destabilising	-1.03	Destabilising	0.03	0.40	0.56	0.91		ns
S458T	drug	4.88	-0.55	Destabilising	-0.96	Destabilising	0.09	0.24	0.33	0.61		ns
S441A	no	6.97	0.52	Stabilising	-0.93	Destabilising	0.95	0.17	<0.0001	<0.001		***
E460D	no	6.79	-0.06	Destabilising	-0.80	Destabilising	0.51	0.11	0.01	0.06		ns
L449Q	no	6.09	-0.69	Destabilising	-1.17	Destabilising	0.53	0.04	<0.001	0.01		**
V168L	no	7.19	0.78	Stabilising	-0.92	Destabilising	0.01	NA	NA	NA		ns
S171T	no	6.10	0.76	Stabilising	-0.96	Destabilising	0.01	NA	NA	NA		ns
E423A	no	9.71	-0.40	Destabilising	-0.50	Destabilising	0.01	NA	NA	NA		ns
Q429R	drug	3.44	-1.43	Destabilising	-0.78	Destabilising	0.01	NA	NA	NA		ns
F433L	drug	4.87	-0.20	Destabilising	-0.89	Destabilising	0.01	NA	NA	NA		ns
M434K	no	6.61	-0.10	Destabilising	-0.94	Destabilising	0.01	NA	NA	NA		ns
D435L	drug	3.10	-0.35	Destabilising	-0.78	Destabilising	0.01	NA	NA	NA		ns
Q436N	no	7.38	0.04	Stabilising	-1.06	Destabilising	0.01	NA	NA	NA		ns
S441P	no	6.97	0.55	Stabilising	-0.73	Destabilising	0.01	NA	NA	NA		ns
S441W	no	6.97	0.13	Stabilising	-0.43	Destabilising	0.01	NA	NA	NA		ns
G442D	no	8.85	-0.37	Destabilising	-1.17	Destabilising	0.01	NA	NA	NA		ns
K446T	no	8.89	0.57	Stabilising	-0.85	Destabilising	0.01	NA	NA	NA		ns
L452M	drug	3.46	-0.99	Destabilising	-0.92	Destabilising	0.01	NA	NA	NA		ns
L452R	drug	3.46	-1.29	Destabilising	-0.50	Destabilising	0.01	NA	NA	NA		ns
R459H	drug	6.34	-0.43	Destabilising	-0.92	Destabilising	0.01	NA	NA	NA		ns
T482A	no	5.58	0.22	Stabilising	-0.96	Destabilising	0.01	NA	NA	NA		ns

P486S	no	5.73	-0.65	Destabilising	-1.09	Destabilising	0.01	NA	NA	NA	ns
S582L	no	9.26	0.59	Stabilising	-0.43	Destabilising	0.01	NA	NA	NA	ns
A584P	no	8.97	0.37	Stabilising	-0.95	Destabilising	0.01	NA	NA	NA	ns
A584S	no	8.97	0.09	Stabilising	-0.75	Destabilising	0.01	NA	NA	NA	ns
N673H	no	9.91	-0.33	Destabilising	-0.52	Destabilising	0.01	NA	NA	NA	ns
H674P	no	4.61	-0.23	Destabilising	-1.02	Destabilising	0.01	NA	NA	NA	ns
H1028Q	no	8.77	-0.33	Destabilising	-0.94	Destabilising	0.01	NA	NA	NA	ns
H1028Y	no	8.77	-0.22	Destabilising	-0.76	Destabilising	0.01	NA	NA	NA	ns
E423G	no	9.71	-0.40	Destabilising	-0.59	Destabilising	0.02	NA	NA	NA	ns
M434R	no	6.61	-0.31	Destabilising	-0.47	Destabilising	0.02	NA	NA	NA	ns
G453A	drug	4.32	-0.91	Destabilising	-1.21	Destabilising	0.02	NA	NA	NA	ns
T482I	no	5.58	0.24	Stabilising	-0.97	Destabilising	0.02	NA	NA	NA	ns
G492S	no	6.19	-0.57	Destabilising	-1.07	Destabilising	0.02	NA	NA	NA	ns
V168M	no	7.19	0.35	Stabilising	-0.92	Destabilising	0.03	NA	NA	NA	ns
S450M	drug	3.32	-0.48	Destabilising	-0.58	Destabilising	0.03	NA	NA	NA	ns
Q429L	drug	3.44	-1.34	Destabilising	-0.74	Destabilising	0.05	NA	NA	NA	ns
S493T	no	5.80	0.48	Stabilising	-0.96	Destabilising	0.06	NA	NA	NA	ns

Table 8.A.1: Mutations close to RFP
One hundred and sixty eight single amino acid variation (SAV) mutations lying within 10Å of RFP and their corresponding ligand affinity changes (log fold change) measured by mCSM-Lig and mmCSM-lig. The estimated effect are categorised as Destabilising (log fold affinity change<0) and Stabilising ($\Delta\Delta G>0$). The genomic measures of minor allele frequency (MAF), Odds Ratio (OR) , OR related P-values, and FDR adjusted P-values are shown. Statistical significance is indicated as: *P $<$ 0.05, **P $<$ 0.01, ***P $<$ 0.001, ****P $<$0.0001, ns: $>$0.05. The table is arranged by Odds Ratio to show mutation with the highest OR at the top for mutations close to RFP. Columns with NA indicate insufficient data to calculate OR. Abbreviations used: FDR: false discovery rate, ns: not significant, RFP: rifampicin.

8.B Mutations close to the nucleic acid

Mutation	Interacting partner	NA-Dist (Å)	mCSM-NA (ΔΔG)	mCSM-NA outcome	MAF (%)	Odds Ratio	P-value	Adjusted P-value	Adjusted P-value significance
H445D	drug	9.32	-3.67	Reduced_affinity	3.23	1,038.14	<0.0001	<0.0001	****
D435V	drug	6.61	0.38	Increased_affinity	6.40	863.35	<0.0001	<0.0001	****
S450L	drug	9.34	-3.56	Reduced_affinity	53.66	573.58	<0.0001	<0.0001	****
H445Y	drug	9.32	2.31	Increased_affinity	4.58	237.23	<0.0001	<0.0001	****
H445C	drug	9.32	0.25	Increased_affinity	0.42	145.63	<0.0001	<0.0001	****
S441L	no	8.98	-3.54	Reduced_affinity	0.42	135.87	<0.0001	<0.0001	****
I488V	no	9.20	0.01	Increased_affinity	0.26	111.51	<0.0001	<0.0001	****
T400A	no	6.27	-3.43	Reduced_affinity	0.27	106.64	<0.0001	<0.0001	****
H445L	drug	9.32	-3.35	Reduced_affinity	1.13	86.80	<0.0001	<0.0001	****
H445Q	drug	9.32	-1.70	Reduced_affinity	0.24	82.33	<0.0001	<0.0001	****
S450F	drug	9.34	2.07	Increased_affinity	0.52	77.69	<0.0001	<0.0001	****
Q432L	drug	8.61	-1.60	Reduced_affinity	0.23	67.76	<0.0001	<0.0001	****
H445R	drug	9.32	-1.04	Reduced_affinity	1.10	57.86	<0.0001	<0.0001	****
S450W	drug	9.34	4.91	Increased_affinity	1.43	57.07	<0.0001	<0.0001	****
I491V	drug	7.59	0.02	Increased_affinity	0.14	53.21	<0.0001	<0.0001	****
Q432P	drug	8.61	-1.62	Reduced_affinity	0.37	48.45	<0.0001	<0.0001	****
D435G	drug	6.61	0.34	Increased_affinity	1.31	46.33	<0.0001	<0.0001	****
Q432K	drug	8.61	1.35	Increased_affinity	0.27	46.02	<0.0001	<0.0001	****
T399I	no	6.28	-3.56	Reduced_affinity	0.09	38.68	<0.0001	<0.001	***
S450Q	drug	9.34	-1.90	Reduced_affinity	0.10	38.68	<0.0001	<0.001	***
Q172R	no	8.12	0.70	Increased_affinity	0.08	33.84	<0.001	0.01	**
I491L	drug	7.59	0.03	Increased_affinity	0.07	29.00	<0.001	0.01	**

D435A	drug	6.61	0.35	Increased_affinity	0.19	26.60	<0.0001	<0.001	***
E481A	no	9.54	0.33	Increased_affinity	0.05	24.16	<0.001	0.02	*
G675D	no	9.51	-0.27	Reduced_affinity	0.06	24.16	<0.001	0.02	*
H445P	drug	9.32	-3.36	Reduced_affinity	0.08	24.16	<0.001	0.02	*
Q401R	no	8.31	0.71	Increased_affinity	0.04	19.33	0.01	0.05	ns
E460G	no	3.33	0.12	Increased_affinity	0.06	19.33	0.01	0.05	ns
D435E	drug	6.61	0.05	Increased_affinity	0.08	19.33	0.01	0.05	ns
H674R	no	6.48	-0.92	Reduced_affinity	0.09	19.33	0.01	0.05	ns
D435F	drug	6.61	6.00	Increased_affinity	0.41	17.77	<0.0001	<0.0001	****
I491T	drug	7.59	3.63	Increased_affinity	0.08	14.50	<0.001	0.03	*
S441Q	no	8.98	-1.91	Reduced_affinity	0.19	14.50	<0.001	0.03	*
S441M	no	8.98	-3.55	Reduced_affinity	0.03	14.49	0.03	0.15	ns
R448Q	drug	6.24	-0.63	Reduced_affinity	0.03	14.49	0.03	0.15	ns
G456S	no	5.29	4.25	Increased_affinity	0.03	14.49	0.03	0.15	ns
I1035V	no	7.66	-0.06	Reduced_affinity	0.04	14.49	0.03	0.15	ns
I491M	drug	7.59	0.03	Increased_affinity	0.05	14.49	0.03	0.15	ns
A451V	no	6.83	-0.03	Reduced_affinity	0.06	14.49	0.03	0.15	ns
N437D	no	5.73	-1.81	Reduced_affinity	0.13	12.09	<0.001	0.01	**
D435N	drug	6.61	2.01	Increased_affinity	0.06	12.08	0.01	0.07	ns
H445G	drug	9.32	-3.40	Reduced_affinity	0.09	12.08	0.01	0.07	ns
L452P	drug	9.24	0.02	Increased_affinity	3.04	11.41	<0.0001	<0.0001	****
R219L	no	3.80	-2.07	Reduced_affinity	0.02	9.66	0.09	0.35	ns
R224L	no	2.59	-1.87	Reduced_affinity	0.02	9.66	0.09	0.35	ns
R225W	no	7.40	6.41	Increased_affinity	0.02	9.66	0.09	0.35	ns
Q436P	no	9.63	-1.58	Reduced_affinity	0.02	9.66	0.09	0.35	ns
S450A	drug	9.34	-3.59	Reduced_affinity	0.02	9.66	0.09	0.35	ns

A451G	no	6.83	-0.07	Reduced_affinity	0.02	9.66	0.09	0.35	ns
P483S	drug	5.79	2.89	Increased_affinity	0.02	9.66	0.09	0.35	ns
R871H	no	8.96	1.08	Increased_affinity	0.02	9.66	0.09	0.35	ns
Q1056R	no	5.94	1.90	Increased_affinity	0.02	9.66	0.09	0.35	ns
T399A	no	6.28	-3.59	Reduced_affinity	0.03	9.66	0.09	0.35	ns
T400N	no	6.27	-1.75	Reduced_affinity	0.03	9.66	0.09	0.35	ns
N437S	no	5.73	2.14	Increased_affinity	0.03	9.66	0.09	0.35	ns
N437H	no	5.73	1.92	Increased_affinity	0.04	9.66	0.09	0.35	ns
H445T	drug	9.32	0.24	Increased_affinity	0.04	9.66	0.09	0.35	ns
N673S	no	8.42	1.98	Increased_affinity	0.06	9.66	0.09	0.35	ns
Q432E	drug	8.61	-1.93	Reduced_affinity	0.09	9.66	0.09	0.35	ns
D435Y	drug	6.61	6.00	Increased_affinity	2.96	8.92	<0.0001	<0.0001	****
P280L	no	5.44	0.04	Increased_affinity	0.06	7.25	0.08	0.35	ns
F93V	no	9.45	-5.61	Reduced_affinity	0.01	4.83	0.29	0.58	ns
L173P	no	5.75	-0.09	Reduced_affinity	0.01	4.83	0.29	0.58	ns
P177A	no	8.91	-0.02	Reduced_affinity	0.01	4.83	0.29	0.58	ns
R219S	no	3.80	1.51	Increased_affinity	0.01	4.83	0.29	0.58	ns
R225G	no	7.40	-2.11	Reduced_affinity	0.01	4.83	0.29	0.58	ns
P280S	no	5.44	3.63	Increased_affinity	0.01	4.83	0.29	0.58	ns
L293M	no	4.03	0.00	Increased_affinity	0.01	4.83	0.29	0.58	ns
F294L	no	8.47	-5.75	Reduced_affinity	0.01	4.83	0.29	0.58	ns
E391G	no	8.41	0.30	Increased_affinity	0.01	4.83	0.29	0.58	ns
T399N	no	6.28	-1.91	Reduced_affinity	0.01	4.83	0.29	0.58	ns
T400I	no	6.27	-3.41	Reduced_affinity	0.01	4.83	0.29	0.58	ns
Q432H	drug	8.61	1.75	Increased_affinity	0.01	4.83	0.29	0.58	ns
D435H	drug	6.61	3.73	Increased_affinity	0.01	4.83	0.29	0.58	ns

D435S	drug	6.61	3.95	Increased_affinity	0.01	4.83	0.29	0.58	ns
Q436L	no	9.63	-1.57	Reduced_affinity	0.01	4.83	0.29	0.58	ns
S441V	no	8.98	-3.55	Reduced_affinity	0.01	4.83	0.29	0.58	ns
H445V	drug	9.32	-3.36	Reduced_affinity	0.01	4.83	0.29	0.58	ns
R448K	drug	6.24	0.68	Increased_affinity	0.01	4.83	0.29	0.58	ns
S450Y	drug	9.34	2.08	Increased_affinity	0.01	4.83	0.29	0.58	ns
L452S	drug	9.24	3.62	Increased_affinity	0.01	4.83	0.29	0.58	ns
G453R	drug	7.91	2.39	Increased_affinity	0.01	4.83	0.29	0.58	ns
P454R	no	6.14	2.37	Increased_affinity	0.01	4.83	0.29	0.58	ns
N604S	no	4.39	1.35	Increased_affinity	0.01	4.83	0.29	0.58	ns
N673D	no	8.42	-1.93	Reduced_affinity	0.01	4.83	0.29	0.58	ns
H1028D	no	3.64	-2.21	Reduced_affinity	0.01	4.83	0.29	0.58	ns
K1034R	no	2.88	0.86	Increased_affinity	0.01	4.83	0.29	0.58	ns
I1035T	no	7.66	3.54	Increased_affinity	0.01	4.83	0.29	0.58	ns
Q1056H	no	5.94	2.94	Increased_affinity	0.01	4.83	0.29	0.58	ns
R219C	no	3.80	1.53	Increased_affinity	0.02	4.83	0.29	0.58	ns
R225P	no	7.40	-2.07	Reduced_affinity	0.02	4.83	0.29	0.58	ns
N437Y	no	5.73	4.19	Increased_affinity	0.02	4.83	0.29	0.58	ns
S450C	drug	9.34	0.05	Increased_affinity	0.02	4.83	0.29	0.58	ns
S450G	drug	9.34	-3.61	Reduced_affinity	0.02	4.83	0.29	0.58	ns
S450V	drug	9.34	-3.57	Reduced_affinity	0.02	4.83	0.29	0.58	ns
I488L	no	9.20	0.03	Increased_affinity	0.02	4.83	0.29	0.58	ns
I491S	drug	7.59	3.62	Increased_affinity	0.02	4.83	0.29	0.58	ns
S188A	no	4.30	-3.53	Reduced_affinity	0.03	4.83	0.29	0.58	ns
H674N	no	6.48	-1.58	Reduced_affinity	0.03	4.83	0.29	0.58	ns
H445F	drug	9.32	2.30	Increased_affinity	0.03	4.83	0.21	0.58	ns

H674Y	no	6.48	2.42	Increased_affinity	0.04	4.83	0.21	0.58	ns
T400S	no	6.27	0.20	Increased_affinity	0.05	4.83	0.29	0.58	ns
R167H	no	9.52	1.07	Increased_affinity	0.07	4.83	0.29	0.58	ns
P454H	no	6.14	3.43	Increased_affinity	0.07	4.83	0.21	0.58	ns
P454L	no	6.14	0.04	Increased_affinity	0.08	4.83	0.21	0.58	ns
H445S	drug	9.32	0.23	Increased_affinity	0.16	3.22	0.21	0.58	ns
I491F	drug	7.59	5.66	Increased_affinity	1.35	3.11	<0.0001	<0.0001	****
H445N	drug	9.32	-1.71	Reduced_affinity	1.06	2.66	<0.001	0.01	**
V403A	no	9.84	0.01	Increased_affinity	0.02	2.41	0.5	0.86	ns
H674Q	no	6.48	-1.57	Reduced_affinity	0.02	2.41	0.5	0.86	ns
V385L	no	7.45	0.04	Increased_affinity	0.03	2.41	0.5	0.86	ns
N487S	drug	6.50	2.00	Increased_affinity	0.05	2.41	0.5	0.86	ns
C1067G	no	8.45	-3.42	Reduced_affinity	0.08	2.41	0.5	0.86	ns
D211V	no	7.47	0.35	Increased_affinity	0.01	1.21	>1	>1	ns
R299C	no	2.82	1.26	Increased_affinity	0.01	1.21	>1	>1	ns
D362H	no	8.96	3.73	Increased_affinity	0.01	1.21	>1	>1	ns
L372P	no	9.60	0.01	Increased_affinity	0.01	1.21	>1	>1	ns
M390I	no	9.55	0.03	Increased_affinity	0.01	1.21	>1	>1	ns
I491N	drug	7.59	1.68	Increased_affinity	0.01	1.21	>1	>1	ns
M601I	no	6.26	-0.48	Reduced_affinity	0.01	1.21	>1	>1	ns
L879I	no	9.25	0.03	Increased_affinity	0.01	1.21	>1	>1	ns
V1031I	no	8.53	0.17	Increased_affinity	0.01	1.21	>1	>1	ns
D190A	no	9.31	0.33	Increased_affinity	0.02	1.21	>1	>1	ns
R225Q	no	7.40	-0.41	Reduced_affinity	0.02	1.21	>1	>1	ns
S388A	no	6.44	-3.58	Reduced_affinity	0.02	1.21	>1	>1	ns
A609S	no	8.70	3.75	Increased_affinity	0.02	1.21	>1	>1	ns

H366N	no	7.05	-1.85	Reduced_affinity	0.03	1.21	>1		>1	ns
V403I	no	9.84	0.05	Increased_affinity	0.03	1.21	>1		>1	ns
N437T	no	5.73	2.15	Increased_affinity	0.03	1.21	>1		>1	ns
P483L	drug	5.79	-0.71	Reduced_affinity	0.03	1.21	>1		>1	ns
H1028L	no	3.64	-1.89	Reduced_affinity	0.03	1.21	>1		>1	ns
F367L	no	7.82	-5.90	Reduced_affinity	0.04	1.21	>1		>1	ns
H1028N	no	3.64	-0.25	Reduced_affinity	0.04	1.21	>1		>1	ns
R225K	no	7.40	0.90	Increased_affinity	0.09	1.21	>1		>1	ns
I1035M	no	7.66	-0.05	Reduced_affinity	0.14	1.21	>1		>1	ns
H1028R	no	3.64	0.41	Increased_affinity	0.15	1.21	>1		>1	ns
M390T	no	9.55	3.62	Increased_affinity	0.10	0.97	>1		>1	ns
D190E	no	9.31	0.05	Increased_affinity	0.59	0.69		0.61	0.98	ns
A880S	no	8.30	3.63	Increased_affinity	0.02	0.60	>1		>1	ns
G463S	no	5.83	3.52	Increased_affinity	0.04	0.60	>1		>1	ns
L372I	no	9.60	0.02	Increased_affinity	0.05	0.60	>1		>1	ns
D1033E	no	5.62	0.93	Increased_affinity	0.07	0.60	>1		>1	ns
L879M	no	9.25	0.02	Increased_affinity	0.13	0.60	>1		>1	ns
A603S	no	8.11	3.52	Increased_affinity	0.24	0.60	>1		>1	ns
T482S	no	7.80	0.07	Increased_affinity	0.08	0.48		0.68	>1	ns
S450N	drug	9.34	-1.91	Reduced_affinity	0.03	0.40		0.56	0.91	ns
S458T	drug	6.48	0.24	Increased_affinity	0.09	0.24		0.33	0.61	ns
C1067V	no	8.45	-3.38	Reduced_affinity	0.27	0.24		0.19	0.58	ns
L464M	no	7.08	0.09	Increased_affinity	0.37	0.22		0.2	0.58	ns
S201G	no	3.72	-4.79	Reduced_affinity	0.19	0.17		0.11	0.45	ns
S441A	no	8.98	-3.57	Reduced_affinity	0.95	0.17	<0.0001		<0.001	***
S388L	no	6.44	-3.55	Reduced_affinity	1.79	0.17	<0.001		0.01	**

E460D	no	3.33	-0.15	Reduced_affinity	0.51	0.11	0.01	0.06	ns
V196A	no	9.99	0.00	Reduced_affinity	0.01	NA	NA	NA	ns
P200S	no	5.11	-0.08	Reduced_affinity	0.01	NA	NA	NA	ns
E207G	no	7.00	0.59	Increased_affinity	0.01	NA	NA	NA	ns
K212E	no	4.26	-3.23	Reduced_affinity	0.01	NA	NA	NA	ns
F294Y	no	8.47	-0.10	Reduced_affinity	0.01	NA	NA	NA	ns
K296N	no	9.92	-1.28	Reduced_affinity	0.01	NA	NA	NA	ns
D365E	no	2.51	0.27	Increased_affinity	0.01	NA	NA	NA	ns
L372M	no	9.60	0.02	Increased_affinity	0.01	NA	NA	NA	ns
R373H	no	6.24	1.07	Increased_affinity	0.01	NA	NA	NA	ns
V385T	no	7.45	3.65	Increased_affinity	0.01	NA	NA	NA	ns
Q401L	no	8.31	-1.61	Reduced_affinity	0.01	NA	NA	NA	ns
D402E	no	7.31	0.05	Increased_affinity	0.01	NA	NA	NA	ns
F433L	drug	7.42	-5.61	Reduced_affinity	0.01	NA	NA	NA	ns
D435L	drug	6.61	0.39	Increased_affinity	0.01	NA	NA	NA	ns
Q436N	no	9.63	0.04	Increased_affinity	0.01	NA	NA	NA	ns
S441P	no	8.98	-3.55	Reduced_affinity	0.01	NA	NA	NA	ns
S441W	no	8.98	4.91	Increased_affinity	0.01	NA	NA	NA	ns
L452M	drug	9.24	0.03	Increased_affinity	0.01	NA	NA	NA	ns
L452R	drug	9.24	2.36	Increased_affinity	0.01	NA	NA	NA	ns
R459H	drug	7.60	1.09	Increased_affinity	0.01	NA	NA	NA	ns
T482A	no	7.80	-3.55	Reduced_affinity	0.01	NA	NA	NA	ns
P486S	no	5.36	3.79	Increased_affinity	0.01	NA	NA	NA	ns
N597S	no	9.75	1.99	Increased_affinity	0.01	NA	NA	NA	ns
R671G	no	8.35	-2.34	Reduced_affinity	0.01	NA	NA	NA	ns
N673H	no	8.42	1.76	Increased_affinity	0.01	NA	NA	NA	ns

H674P	no	6.48	-3.24	Reduced_affinity	0.01	NA	NA	NA	ns
H1028Q	no	3.64	-0.24	Reduced_affinity	0.01	NA	NA	NA	ns
H1028Y	no	3.64	3.77	Increased_affinity	0.01	NA	NA	NA	ns
I1035L	no	7.66	-0.05	Reduced_affinity	0.01	NA	NA	NA	ns
A1037S	no	8.57	3.57	Increased_affinity	0.01	NA	NA	NA	ns
C1067W	no	8.45	5.10	Increased_affinity	0.01	NA	NA	NA	ns
I220V	no	8.15	0.02	Increased_affinity	0.02	NA	NA	NA	ns
G453A	drug	7.91	0.05	Increased_affinity	0.02	NA	NA	NA	ns
T482I	no	7.80	-3.51	Reduced_affinity	0.02	NA	NA	NA	ns
L1027Q	no	6.52	2.67	Increased_affinity	0.02	NA	NA	NA	ns
S450M	drug	9.34	-3.56	Reduced_affinity	0.03	NA	NA	NA	ns
L464V	no	7.08	0.08	Increased_affinity	0.03	NA	NA	NA	ns
K1054R	no	3.12	-0.62	Reduced_affinity	0.04	NA	NA	NA	ns
G463A	no	5.83	-0.10	Reduced_affinity	0.05	NA	NA	NA	ns
I220L	no	8.15	0.03	Increased_affinity	0.06	NA	NA	NA	ns
E207K	no	7.00	3.60	Increased_affinity	0.07	NA	NA	NA	ns
N381H	no	8.96	1.76	Increased_affinity	0.79	NA	NA	NA	ns

Table 8.B.1: Mutations close to nucleic acid in RpoB RNA polymerase β subunit
One hundred and ninety five amino acid variation (SAV) mutations lying within 10Å of the Nucleic Acid (NA) and their corresponding PPI affinity changes ($\Delta\Delta$G) measured by mCSM-NA. The estimated effect are categorised as Destabilising ($\Delta\Delta$G<0) and Stabilising ($\Delta\Delta$G>0). The genomic measures of minor allele frequency (MAF), Odds Ratio (OR), OR related P-values, and FDR adjusted P-values are shown. Statistical significance is indicated as: *P < 0.05, **P < 0.01, ***P < 0.001, ****P <0.0001, ns: >0.05. The table is arranged by OR to show mutation with the highest OR at the top. Columns with NA indicate insufficient data to calculate OR. Abbreviations used: $\Delta\Delta$G: change in Gibbs free energy in Kcal/mol, FDR: false discovery rate, ns: not significant, RFP: rifampicin.

8.C Mutations close to the protein-protein interface

Mutation	Interacting partner	PPI2-Dist (Å)	mCSM-PPI2 (ΔΔG)	mCSM-PPI2 outcome	MAF (%)	Odds Ratio	P-value	Adjusted P-value	Adjusted P-value significance
H445D	drug	7.97	-0.43	Decreasing	3.23	1,038.14	<0.0001	<0.0001	****
D435V	drug	6.61	-0.31	Decreasing	6.40	863.35	<0.0001	<0.0001	****
S450L	drug	7.97	-0.15	Decreasing	53.66	573.58	<0.0001	<0.0001	****
V170F	no	9.21	1.05	Increasing	0.81	297.78	<0.0001	<0.0001	****
R827C	no	9.04	-0.20	Decreasing	0.87	292.85	<0.0001	<0.0001	****
H445Y	drug	7.97	0.19	Increasing	4.58	237.23	<0.0001	<0.0001	****
I480V	no	3.49	-0.67	Decreasing	0.38	150.51	<0.0001	<0.0001	****
H445C	drug	7.97	-0.46	Decreasing	0.42	145.63	<0.0001	<0.0001	****
S441L	no	8.98	-0.73	Decreasing	0.42	135.87	<0.0001	<0.0001	****
E761D	no	7.93	-0.45	Decreasing	2.55	123.88	<0.0001	<0.0001	****
R552L	no	7.55	0.03	Increasing	0.29	121.25	<0.0001	<0.0001	****
A286V	no	9.01	0.25	Increasing	0.38	116.38	<0.0001	<0.0001	****
I488V	no	3.83	0.08	Increasing	0.26	111.51	<0.0001	<0.0001	****
I1106T	no	3.10	-0.51	Decreasing	0.80	111.51	<0.0001	<0.0001	****
T400A	no	4.53	-0.13	Decreasing	0.27	106.64	<0.0001	<0.0001	****
V970A	no	6.43	-0.27	Decreasing	0.21	101.77	<0.0001	<0.0001	****
H445L	drug	7.97	-0.53	Decreasing	1.13	86.80	<0.0001	<0.0001	****
H445Q	drug	7.97	-0.17	Decreasing	0.24	82.33	<0.0001	<0.0001	****
S450F	drug	7.97	0.05	Increasing	0.52	77.69	<0.0001	<0.0001	****
S874Y	no	6.04	-0.10	Decreasing	0.14	67.76	<0.0001	<0.0001	****
Q432L	drug	6.63	-0.02	Decreasing	0.23	67.76	<0.0001	<0.0001	****
Q409R	no	3.14	-0.22	Decreasing	0.35	58.06	<0.0001	<0.0001	****

H445R	drug	7.97	-0.15	Decreasing	1.10	57.86	<0.0001	<0.0001	****
S450W	drug	7.97	0.18	Increasing	1.43	57.07	<0.0001	<0.0001	****
I491V	drug	5.35	0.22	Increasing	0.14	53.21	<0.0001	<0.0001	****
F971L	no	3.55	-1.92	Decreasing	0.15	53.21	<0.0001	<0.0001	****
Q432P	drug	6.63	-0.02	Decreasing	0.37	48.45	<0.0001	<0.0001	****
D435G	drug	6.61	-0.55	Decreasing	1.31	46.33	<0.0001	<0.0001	****
Q432K	drug	6.63	-0.12	Decreasing	0.27	46.02	<0.0001	<0.0001	****
S431G	drug	9.51	0.00	Decreasing	0.11	43.52	<0.0001	<0.001	***
I873F	no	8.98	0.63	Increasing	0.11	43.52	<0.0001	<0.001	***
R827H	no	9.04	-0.03	Decreasing	0.19	43.52	<0.0001	<0.001	***
H835R	no	5.66	-0.10	Decreasing	0.21	38.73	<0.0001	<0.0001	****
T399I	no	4.44	-0.13	Decreasing	0.09	38.68	<0.0001	<0.001	***
S428R	drug	7.24	0.08	Increasing	0.10	38.68	<0.0001	<0.001	***
S450Q	drug	7.97	-0.13	Decreasing	0.10	38.68	<0.0001	<0.001	***
L430R	drug	8.96	-0.07	Decreasing	0.19	36.31	<0.0001	<0.0001	****
N163K	no	9.54	0.09	Increasing	0.07	33.84	<0.001	0.01	**
V562A	no	3.66	-0.35	Decreasing	0.07	33.84	<0.001	0.01	**
Q172R	no	8.12	-0.58	Decreasing	0.08	33.84	<0.001	0.01	**
R552H	no	7.55	-0.11	Decreasing	0.06	29.00	<0.001	0.01	**
I491L	drug	5.35	-0.50	Decreasing	0.07	29.00	<0.001	0.01	**
Q429H	drug	3.39	0.02	Increasing	0.09	29.00	<0.001	0.01	**
D574E	no	5.48	-0.51	Decreasing	0.09	29.00	<0.001	0.01	**
V970M	no	6.43	-0.47	Decreasing	0.18	26.60	<0.0001	<0.001	***
D435A	drug	6.61	-0.54	Decreasing	0.19	26.60	<0.0001	<0.001	***
E481A	no	4.48	-0.72	Decreasing	0.05	24.16	<0.001	0.02	*
G675D	no	9.51	-0.03	Decreasing	0.06	24.16	<0.001	0.02	*

H445P	drug	7.97	-0.60	Decreasing	0.08	24.16	<0.001		0.02	*
R827L	no	9.04	-0.24	Decreasing	0.08	24.16	<0.001		0.02	*
M920V	no	3.46	-1.17	Decreasing	0.08	24.16	<0.001		0.02	*
Y564H	no	6.16	-0.18	Decreasing	0.09	24.16	<0.001		0.02	*
V170A	no	9.21	-0.25	Decreasing	0.04	19.33	0.01		0.05	ns
Q401R	no	7.77	0.04	Increasing	0.04	19.33	0.01		0.05	ns
R557H	no	7.00	0.09	Increasing	0.04	19.33	0.01		0.05	ns
T427I	no	5.91	-0.04	Decreasing	0.05	19.33	0.01		0.05	ns
S428G	drug	7.24	-0.11	Decreasing	0.05	19.33	0.01		0.05	ns
D265G	no	6.03	-0.21	Decreasing	0.06	19.33	0.01		0.05	ns
E460G	no	3.33	-0.80	Decreasing	0.06	19.33	0.01		0.05	ns
D435E	drug	6.61	0.20	Increasing	0.08	19.33	0.01		0.05	ns
H674R	no	6.48	0.07	Increasing	0.09	19.33	0.01		0.05	ns
K799Q	no	8.27	-0.33	Decreasing	0.13	19.33	0.01		0.05	ns
D435F	drug	6.61	-0.48	Decreasing	0.41	17.77	<0.0001	<0.0001		****
I491T	drug	5.35	-0.34	Decreasing	0.08	14.50	<0.001		0.03	*
K891E	no	4.42	-1.24	Decreasing	0.10	14.50	<0.001		0.03	*
S441Q	no	8.98	-0.40	Decreasing	0.19	14.50	<0.001		0.03	*
S441M	no	8.98	-0.71	Decreasing	0.03	14.49	0.03		0.15	ns
R448Q	drug	4.52	-0.34	Decreasing	0.03	14.49	0.03		0.15	ns
G456S	no	5.29	-0.11	Decreasing	0.03	14.49	0.03		0.15	ns
V1096M	no	3.91	0.44	Increasing	0.03	14.49	0.03		0.15	ns
A753V	no	8.76	0.11	Increasing	0.04	14.49	0.03		0.15	ns
I1035V	no	3.41	-0.55	Decreasing	0.04	14.49	0.03		0.15	ns
I491M	drug	5.35	-0.52	Decreasing	0.05	14.49	0.03		0.15	ns
G836S	no	5.25	0.10	Increasing	0.05	14.49	0.03		0.15	ns

A974V	no	7.35	0.14	Increasing	0.05	14.49	0.03	0.15	ns
A451V	no	6.83	-0.03	Decreasing	0.06	14.49	0.03	0.15	ns
V581M	no	7.37	-0.62	Decreasing	0.07	14.49	0.03	0.15	ns
N437D	no	5.73	-0.29	Decreasing	0.13	12.09	<0.001	0.01	**
D435N	drug	6.61	-0.51	Decreasing	0.06	12.08	0.01	0.07	ns
H445G	drug	7.97	-0.50	Decreasing	0.09	12.08	0.01	0.07	ns
L452P	drug	6.98	-0.30	Decreasing	3.04	11.41	<0.0001	<0.0001	****
Q975H	no	4.27	0.17	Increasing	0.43	11.32	<0.0001	<0.0001	****
K37R	no	8.98	0.16	Increasing	0.13	9.67	<0.001	0.01	*
R219L	no	3.80	-0.30	Decreasing	0.02	9.66	0.09	0.35	ns
R224L	no	2.59	-0.36	Decreasing	0.02	9.66	0.09	0.35	ns
R225W	no	7.40	-0.21	Decreasing	0.02	9.66	0.09	0.35	ns
Q436P	no	9.63	-0.40	Decreasing	0.02	9.66	0.09	0.35	ns
S450A	drug	7.97	-0.15	Decreasing	0.02	9.66	0.09	0.35	ns
A451G	no	6.83	-0.30	Decreasing	0.02	9.66	0.09	0.35	ns
P483S	drug	5.60	-0.22	Decreasing	0.02	9.66	0.09	0.35	ns
A538V	no	6.50	0.02	Increasing	0.02	9.66	0.09	0.35	ns
S576L	no	6.51	-0.12	Decreasing	0.02	9.66	0.09	0.35	ns
S582A	no	9.74	-0.49	Decreasing	0.02	9.66	0.09	0.35	ns
A670D	no	8.28	0.34	Increasing	0.02	9.66	0.09	0.35	ns
R754H	no	3.08	0.11	Increasing	0.02	9.66	0.09	0.35	ns
P768L	no	6.89	-0.03	Decreasing	0.02	9.66	0.09	0.35	ns
K832E	no	6.62	-0.58	Decreasing	0.02	9.66	0.09	0.35	ns
L855S	no	6.04	-0.18	Decreasing	0.02	9.66	0.09	0.35	ns
R871H	no	7.28	0.04	Increasing	0.02	9.66	0.09	0.35	ns
E1011G	no	2.93	-1.38	Decreasing	0.02	9.66	0.09	0.35	ns

Q1056R	no	4.16	0.26	Increasing	0.02	9.66	0.09	0.35	ns
T399A	no	4.44	-0.11	Decreasing	0.03	9.66	0.09	0.35	ns
T400N	no	4.53	0.08	Increasing	0.03	9.66	0.09	0.35	ns
N437S	no	5.73	-0.19	Decreasing	0.03	9.66	0.09	0.35	ns
I480T	no	3.49	-1.29	Decreasing	0.03	9.66	0.09	0.35	ns
A599V	no	3.36	0.13	Increasing	0.03	9.66	0.09	0.35	ns
H723D	no	3.26	-1.18	Decreasing	0.03	9.66	0.09	0.35	ns
E812G	no	3.23	-0.78	Decreasing	0.03	9.66	0.09	0.35	ns
K1102T	no	3.82	-0.43	Decreasing	0.03	9.66	0.09	0.35	ns
N437H	no	5.73	-0.41	Decreasing	0.04	9.66	0.09	0.35	ns
H445T	drug	7.97	-0.32	Decreasing	0.04	9.66	0.09	0.35	ns
I588V	no	7.25	-0.20	Decreasing	0.04	9.66	0.09	0.35	ns
H593Y	no	2.83	0.20	Increasing	0.04	9.66	0.09	0.35	ns
R824L	no	6.56	-0.36	Decreasing	0.04	9.66	0.09	0.35	ns
N673S	no	8.42	-0.24	Decreasing	0.06	9.66	0.09	0.35	ns
S493L	no	7.37	-0.47	Decreasing	0.07	9.66	0.03	0.16	ns
D777N	no	4.00	0.18	Increasing	0.07	9.66	0.09	0.35	ns
Q432E	drug	6.63	0.03	Increasing	0.09	9.66	0.09	0.35	ns
Q980A	no	3.65	-0.38	Decreasing	0.14	9.66	0.09	0.35	ns
D435Y	drug	6.61	-0.43	Decreasing	2.96	8.92	<0.0001	<0.0001	****
P280L	no	4.60	-0.43	Decreasing	0.06	7.25	0.08	0.35	ns
L982M	no	7.78	0.10	Increasing	0.20	7.25	0.08	0.35	ns
L430P	drug	8.96	0.03	Increasing	1.78	5.07	<0.0001	<0.0001	****
F93V	no	9.45	-0.44	Decreasing	0.01	4.83	0.29	0.58	ns
D150E	no	3.90	0.67	Increasing	0.01	4.83	0.29	0.58	ns
V170G	no	9.21	-0.41	Decreasing	0.01	4.83	0.29	0.58	ns

V170L	no	9.21	-0.12	Decreasing	0.01	4.83	0.29	0.58	ns
L173P	no	5.75	-1.03	Decreasing	0.01	4.83	0.29	0.58	ns
P177A	no	8.91	-0.15	Decreasing	0.01	4.83	0.29	0.58	ns
R219S	no	3.80	-0.15	Decreasing	0.01	4.83	0.29	0.58	ns
R225G	no	7.40	-0.36	Decreasing	0.01	4.83	0.29	0.58	ns
D265A	no	6.03	-0.15	Decreasing	0.01	4.83	0.29	0.58	ns
K274N	no	8.09	-0.05	Decreasing	0.01	4.83	0.29	0.58	ns
P280S	no	4.60	-0.13	Decreasing	0.01	4.83	0.29	0.58	ns
E284G	no	7.39	-0.06	Decreasing	0.01	4.83	0.29	0.58	ns
L293M	no	4.03	-0.03	Decreasing	0.01	4.83	0.29	0.58	ns
F294L	no	8.47	-0.77	Decreasing	0.01	4.83	0.29	0.58	ns
E391G	no	8.41	-0.21	Decreasing	0.01	4.83	0.29	0.58	ns
T399N	no	4.44	0.09	Increasing	0.01	4.83	0.29	0.58	ns
T400I	no	4.53	0.20	Increasing	0.01	4.83	0.29	0.58	ns
T427P	no	5.91	-0.08	Decreasing	0.01	4.83	0.29	0.58	ns
Q429P	drug	3.39	0.23	Increasing	0.01	4.83	0.29	0.58	ns
L430V	drug	8.96	-0.13	Decreasing	0.01	4.83	0.29	0.58	ns
Q432H	drug	6.63	0.26	Increasing	0.01	4.83	0.29	0.58	ns
D435H	drug	6.61	-0.22	Decreasing	0.01	4.83	0.29	0.58	ns
D435S	drug	6.61	-0.50	Decreasing	0.01	4.83	0.29	0.58	ns
Q436L	no	9.63	-0.31	Decreasing	0.01	4.83	0.29	0.58	ns
S441V	no	8.98	-0.53	Decreasing	0.01	4.83	0.29	0.58	ns
H445V	drug	7.97	-0.51	Decreasing	0.01	4.83	0.29	0.58	ns
R448K	drug	4.52	0.06	Increasing	0.01	4.83	0.29	0.58	ns
L449M	no	8.82	-0.60	Decreasing	0.01	4.83	0.29	0.58	ns
S450Y	drug	7.97	0.13	Increasing	0.01	4.83	0.29	0.58	ns

L452S	drug	6.98	-0.13	Decreasing	0.01	4.83	0.29	0.58	ns
G453R	drug	7.91	-0.30	Decreasing	0.01	4.83	0.29	0.58	ns
P454R	no	6.14	0.05	Increasing	0.01	4.83	0.29	0.58	ns
V469L	no	2.33	-0.46	Decreasing	0.01	4.83	0.29	0.58	ns
Y474H	no	3.68	-1.22	Decreasing	0.01	4.83	0.29	0.58	ns
P479S	no	2.46	-1.30	Decreasing	0.01	4.83	0.29	0.58	ns
N539K	no	5.65	-0.31	Decreasing	0.01	4.83	0.29	0.58	ns
L554Q	no	5.21	-0.62	Decreasing	0.01	4.83	0.29	0.58	ns
A559V	no	7.68	-0.03	Decreasing	0.01	4.83	0.29	0.58	ns
N604S	no	4.39	-0.41	Decreasing	0.01	4.83	0.29	0.58	ns
R667L	no	7.84	-0.05	Decreasing	0.01	4.83	0.29	0.58	ns
N673D	no	8.42	-0.04	Decreasing	0.01	4.83	0.29	0.58	ns
D704Q	no	9.80	-0.18	Decreasing	0.01	4.83	0.29	0.58	ns
N733S	no	4.88	-0.41	Decreasing	0.01	4.83	0.29	0.58	ns
E750Q	no	8.81	-0.51	Decreasing	0.01	4.83	0.29	0.58	ns
T756N	no	7.81	-0.09	Decreasing	0.01	4.83	0.29	0.58	ns
E761A	no	7.93	-0.08	Decreasing	0.01	4.83	0.29	0.58	ns
G793V	no	3.43	-0.52	Decreasing	0.01	4.83	0.29	0.58	ns
E807A	no	3.16	-1.22	Decreasing	0.01	4.83	0.29	0.58	ns
E807G	no	3.16	-1.00	Decreasing	0.01	4.83	0.29	0.58	ns
E821V	no	3.05	-0.36	Decreasing	0.01	4.83	0.29	0.58	ns
T829I	no	8.39	-0.04	Decreasing	0.01	4.83	0.29	0.58	ns
K832Q	no	6.62	-0.33	Decreasing	0.01	4.83	0.29	0.58	ns
P834T	no	8.48	-0.23	Decreasing	0.01	4.83	0.29	0.58	ns
H835Q	no	5.66	-0.59	Decreasing	0.01	4.83	0.29	0.58	ns
V867M	no	6.98	-0.37	Decreasing	0.01	4.83	0.29	0.58	ns

S874F	no	6.04	-0.18	Decreasing	0.01	4.83	0.29	0.58	ns
H883Y	no	8.51	-0.07	Decreasing	0.01	4.83	0.29	0.58	ns
I892F	no	8.20	0.10	Increasing	0.01	4.83	0.29	0.58	ns
V895F	no	4.60	0.08	Increasing	0.01	4.83	0.29	0.58	ns
H929Y	no	3.07	-0.55	Decreasing	0.01	4.83	0.29	0.58	ns
V970G	no	6.43	-0.44	Decreasing	0.01	4.83	0.29	0.58	ns
V970L	no	6.43	-0.37	Decreasing	0.01	4.83	0.29	0.58	ns
G973S	no	9.62	-0.25	Decreasing	0.01	4.83	0.29	0.58	ns
L979R	no	4.42	-0.73	Decreasing	0.01	4.83	0.29	0.58	ns
G981D	no	6.12	-0.14	Decreasing	0.01	4.83	0.29	0.58	ns
V996G	no	9.64	-0.05	Decreasing	0.01	4.83	0.29	0.58	ns
H1028D	no	3.64	-0.23	Decreasing	0.01	4.83	0.29	0.58	ns
K1034R	no	2.88	-0.57	Decreasing	0.01	4.83	0.29	0.58	ns
I1035T	no	3.41	-1.11	Decreasing	0.01	4.83	0.29	0.58	ns
Q1056H	no	4.16	0.22	Increasing	0.01	4.83	0.29	0.58	ns
Q1080R	no	3.16	-0.71	Decreasing	0.01	4.83	0.29	0.58	ns
K1095R	no	6.21	0.05	Increasing	0.01	4.83	0.29	0.58	ns
A124S	no	8.42	-0.23	Decreasing	0.02	4.83	0.29	0.58	ns
R219C	no	3.80	-0.33	Decreasing	0.02	4.83	0.29	0.58	ns
R225P	no	7.40	-0.31	Decreasing	0.02	4.83	0.29	0.58	ns
K274E	no	8.09	0.02	Increasing	0.02	4.83	0.29	0.58	ns
N437Y	no	5.73	-0.35	Decreasing	0.02	4.83	0.29	0.58	ns
S450C	drug	7.97	-0.22	Decreasing	0.02	4.83	0.29	0.58	ns
S450G	drug	7.97	-0.16	Decreasing	0.02	4.83	0.29	0.58	ns
S450V	drug	7.97	0.00	Decreasing	0.02	4.83	0.29	0.58	ns
I488L	no	3.83	-0.94	Decreasing	0.02	4.83	0.29	0.58	ns

I491S	drug	5.35	-0.50	Decreasing	0.02	4.83	0.29	0.58	ns
V562M	no	3.66	-0.26	Decreasing	0.02	4.83	0.29	0.58	ns
V581L	no	7.37	-0.21	Decreasing	0.02	4.83	0.29	0.58	ns
V736L	no	3.56	-0.20	Decreasing	0.02	4.83	0.29	0.58	ns
H745Y	no	6.85	0.18	Increasing	0.02	4.83	0.29	0.58	ns
E750G	no	8.81	-0.10	Decreasing	0.02	4.83	0.29	0.58	ns
K891R	no	4.42	0.22	Increasing	0.02	4.83	0.29	0.58	ns
G992R	no	7.61	-0.21	Decreasing	0.02	4.83	0.29	0.58	ns
R1008C	no	3.21	-1.16	Decreasing	0.02	4.83	0.29	0.58	ns
E1104D	no	3.30	0.51	Increasing	0.02	4.83	0.29	0.58	ns
I1111V	no	3.51	-0.03	Decreasing	0.02	4.83	0.29	0.58	ns
S188A	no	4.30	-0.03	Decreasing	0.03	4.83	0.29	0.58	ns
I271V	no	9.64	-0.19	Decreasing	0.03	4.83	0.29	0.58	ns
S431C	drug	9.51	0.03	Increasing	0.03	4.83	0.29	0.58	ns
L554P	no	5.21	0.08	Increasing	0.03	4.83	0.29	0.58	ns
R578H	no	6.06	0.18	Increasing	0.03	4.83	0.29	0.58	ns
H674N	no	6.48	-0.29	Decreasing	0.03	4.83	0.29	0.58	ns
T756A	no	7.81	-0.22	Decreasing	0.03	4.83	0.29	0.58	ns
L862R	no	9.78	-0.33	Decreasing	0.03	4.83	0.29	0.58	ns
G973D	no	9.62	-0.30	Decreasing	0.03	4.83	0.29	0.58	ns
H445F	drug	7.97	-0.02	Decreasing	0.03	4.83	0.21	0.58	ns
H835P	no	5.66	0.22	Increasing	0.03	4.83	0.21	0.58	ns
H674Y	no	6.48	0.03	Increasing	0.04	4.83	0.21	0.58	ns
C681G	no	3.21	-0.66	Decreasing	0.04	4.83	0.21	0.58	ns
T400S	no	4.53	0.06	Increasing	0.05	4.83	0.29	0.58	ns
I652L	no	7.23	-0.20	Decreasing	0.05	4.83	0.21	0.58	ns

S838T	no	6.34	-0.27	Decreasing	0.05	4.83	0.29		0.58	ns
Q980R	no	3.65	-0.11	Decreasing	0.06	4.83	0.29		0.58	ns
R167H	no	9.02	-0.11	Decreasing	0.07	4.83	0.29		0.58	ns
P454H	no	6.14	0.06	Increasing	0.07	4.83	0.21		0.58	ns
D997N	no	6.79	0.08	Increasing	0.07	4.83	0.29		0.58	ns
P454L	no	6.14	0.21	Increasing	0.08	4.83	0.21		0.58	ns
T1090V	no	3.55	0.34	Increasing	0.11	4.83	0.29		0.58	ns
A977E	no	5.77	1.08	Increasing	0.18	3.62	0.15		0.58	ns
D993E	no	5.65	0.01	Increasing	0.23	3.62	0.15		0.58	ns
L995M	no	8.12	-0.43	Decreasing	0.23	3.62	0.15		0.58	ns
P577A	no	4.16	-0.75	Decreasing	0.08	3.22	0.21		0.58	ns
H445S	drug	7.97	-0.34	Decreasing	0.16	3.22	0.21		0.58	ns
I491F	drug	5.35	0.06	Increasing	1.35	3.11	<0.0001	<0.0001		****
H445N	drug	7.97	-0.35	Decreasing	1.06	2.66	<0.001		0.01	**
V403A	no	9.84	0.16	Increasing	0.02	2.41	0.5		0.86	ns
S431T	drug	9.51	0.05	Increasing	0.02	2.41	0.5		0.86	ns
N539H	no	5.65	-0.11	Decreasing	0.02	2.41	0.5		0.86	ns
H674Q	no	6.48	0.19	Increasing	0.02	2.41	0.5		0.86	ns
D779G	no	8.59	0.06	Increasing	0.02	2.41	0.5		0.86	ns
R824S	no	6.56	0.03	Increasing	0.02	2.41	0.5		0.86	ns
V385L	no	7.45	-0.06	Decreasing	0.03	2.41	0.5		0.86	ns
I696L	no	7.52	-0.46	Decreasing	0.03	2.41	0.5		0.86	ns
V740T	no	8.04	0.03	Increasing	0.03	2.41	0.5		0.86	ns
L741F	no	9.95	0.16	Increasing	0.03	2.41	0.5		0.86	ns
E807Q	no	3.16	-0.60	Decreasing	0.04	2.41	0.5		0.86	ns
T585A	no	6.55	0.05	Increasing	0.04	2.41	0.59		0.94	ns

D270E	no	7.39	-0.16	Decreasing	0.05	2.41	0.5	0.86	ns
N487S	drug	4.11	-0.18	Decreasing	0.05	2.41	0.5	0.86	ns
I925V	no	3.79	-0.06	Decreasing	0.05	2.41	0.59	0.94	ns
S431R	drug	9.51	0.10	Increasing	0.06	2.41	0.59	0.94	ns
C1067G	no	3.63	-0.72	Decreasing	0.08	2.41	0.5	0.86	ns
H723Y	no	3.26	1.66	Increasing	0.12	2.41	0.5	0.86	ns
V109I	no	9.64	-0.25	Decreasing	0.01	1.21	1	1	ns
M121I	no	8.44	-0.50	Decreasing	0.01	1.21	1	1	ns
D211V	no	7.47	-0.23	Decreasing	0.01	1.21	1	1	ns
V262F	no	9.21	0.37	Increasing	0.01	1.21	1	1	ns
G263N	no	9.92	-0.08	Decreasing	0.01	1.21	1	1	ns
R299C	no	2.82	-0.62	Decreasing	0.01	1.21	1	1	ns
D362H	no	8.96	0.16	Increasing	0.01	1.21	1	1	ns
L372P	no	9.60	-0.78	Decreasing	0.01	1.21	1	1	ns
M390I	no	5.71	-0.44	Decreasing	0.01	1.21	1	1	ns
M477V	no	4.90	0.04	Increasing	0.01	1.21	1	1	ns
I491N	drug	5.35	-0.46	Decreasing	0.01	1.21	1	1	ns
V555A	no	4.73	-0.03	Decreasing	0.01	1.21	1	1	ns
M601I	no	6.26	-0.19	Decreasing	0.01	1.21	1	1	ns
S615N	no	8.32	-0.18	Decreasing	0.01	1.21	1	1	ns
V647L	no	3.53	-0.81	Decreasing	0.01	1.21	1	1	ns
A649V	no	2.81	-0.26	Decreasing	0.01	1.21	1	1	ns
M666T	no	7.28	-0.54	Decreasing	0.01	1.21	1	1	ns
F669L	no	3.24	-1.44	Decreasing	0.01	1.21	1	1	ns
P682T	no	4.18	-0.51	Decreasing	0.01	1.21	1	1	ns
I683T	no	3.73	-0.21	Decreasing	0.01	1.21	1	1	ns

A697T	no	8.91	0.45	Increasing	0.01	1.21	1	1	ns
I717Y	no	3.15	2.12	Increasing	0.01	1.21	1	1	ns
S732C	no	8.51	-0.19	Decreasing	0.01	1.21	1	1	ns
N733R	no	4.88	-0.50	Decreasing	0.01	1.21	1	1	ns
E737G	no	3.53	-0.69	Decreasing	0.01	1.21	1	1	ns
I744T	no	3.68	-0.99	Decreasing	0.01	1.21	1	1	ns
P810A	no	0.73	-1.05	Decreasing	0.01	1.21	1	1	ns
P810C	no	0.73	-0.36	Decreasing	0.01	1.21	1	1	ns
P810R	no	0.73	0.48	Increasing	0.01	1.21	1	1	ns
P834L	no	8.48	-0.18	Decreasing	0.01	1.21	1	1	ns
D851G	no	8.43	-0.42	Decreasing	0.01	1.21	1	1	ns
P856L	no	4.64	-0.01	Decreasing	0.01	1.21	1	1	ns
A868V	no	4.05	-0.05	Decreasing	0.01	1.21	1	1	ns
L879I	no	7.31	-0.56	Decreasing	0.01	1.21	1	1	ns
L893R	no	7.59	-0.28	Decreasing	0.01	1.21	1	1	ns
L893V	no	7.59	0.03	Increasing	0.01	1.21	1	1	ns
V895D	no	4.60	-0.10	Decreasing	0.01	1.21	1	1	ns
V895L	no	4.60	-0.24	Decreasing	0.01	1.21	1	1	ns
P899A	no	6.27	-0.63	Decreasing	0.01	1.21	1	1	ns
D903N	no	3.29	0.74	Increasing	0.01	1.21	1	1	ns
P906L	no	4.65	-0.57	Decreasing	0.01	1.21	1	1	ns
I910T	no	8.15	-0.53	Decreasing	0.01	1.21	1	1	ns
S984L	no	8.53	-0.05	Decreasing	0.01	1.21	1	1	ns
G992N	no	7.61	0.04	Increasing	0.01	1.21	1	1	ns
M1003I	no	7.85	-0.28	Decreasing	0.01	1.21	1	1	ns
F1005V	no	3.74	-1.89	Decreasing	0.01	1.21	1	1	ns

D1006G	no	3.79	-1.75	Decreasing	0.01	1.21	1	1	ns
P1012L	no	3.56	-0.51	Decreasing	0.01	1.21	1	1	ns
P1014S	no	3.74	-0.32	Decreasing	0.01	1.21	1	1	ns
V1019T	no	6.36	-0.17	Decreasing	0.01	1.21	1	1	ns
M1022L	no	9.04	-0.14	Decreasing	0.01	1.21	1	1	ns
V1031I	no	3.72	-0.14	Decreasing	0.01	1.21	1	1	ns
P1042S	no	3.61	-0.32	Decreasing	0.01	1.21	1	1	ns
V1091I	no	6.53	-0.11	Decreasing	0.01	1.21	1	1	ns
V1091S	no	6.53	-0.22	Decreasing	0.01	1.21	1	1	ns
V1096A	no	3.91	-1.06	Decreasing	0.01	1.21	1	1	ns
E1108A	no	4.30	-1.08	Decreasing	0.01	1.21	1	1	ns
V1117L	no	3.63	0.52	Increasing	0.01	1.21	1	1	ns
V1129A	no	3.52	-1.41	Decreasing	0.01	1.21	1	1	ns
S1134C	no	3.63	-0.21	Decreasing	0.01	1.21	1	1	ns
E1145G	no	2.87	-0.99	Decreasing	0.01	1.21	1	1	ns
E1147Q	no	2.89	-0.91	Decreasing	0.01	1.21	1	1	ns
D190A	no	3.64	-0.23	Decreasing	0.02	1.21	1	1	ns
R225Q	no	7.40	-0.20	Decreasing	0.02	1.21	1	1	ns
T261A	no	8.15	-0.07	Decreasing	0.02	1.21	1	1	ns
S388A	no	6.44	-0.01	Decreasing	0.02	1.21	1	1	ns
V418I	no	4.66	-0.08	Decreasing	0.02	1.21	1	1	ns
A609S	no	8.70	0.48	Increasing	0.02	1.21	1	1	ns
V613L	no	3.64	0.41	Increasing	0.02	1.21	1	1	ns
A670N	no	8.28	0.17	Increasing	0.02	1.21	1	1	ns
C681K	no	3.21	0.37	Increasing	0.02	1.21	1	1	ns
P682A	no	4.18	-1.20	Decreasing	0.02	1.21	1	1	ns

D688E	no	5.95	0.18	Increasing	0.02	1.21	1	1	ns
L714M	no	8.95	-0.37	Decreasing	0.02	1.21	1	1	ns
E737S	no	3.53	-0.44	Decreasing	0.02	1.21	1	1	ns
R791P	no	2.82	-1.46	Decreasing	0.02	1.21	1	1	ns
D792G	no	3.57	-0.31	Decreasing	0.02	1.21	1	1	ns
R813K	no	3.00	0.92	Increasing	0.02	1.21	1	1	ns
E850D	no	8.98	0.09	Increasing	0.02	1.21	1	1	ns
L901R	no	4.65	-0.60	Decreasing	0.02	1.21	1	1	ns
F1005Y	no	3.74	0.79	Increasing	0.02	1.21	1	1	ns
P1109S	no	4.39	-0.04	Decreasing	0.02	1.21	1	1	ns
S1133A	no	4.05	-0.60	Decreasing	0.02	1.21	1	1	ns
H366N	no	7.05	-0.26	Decreasing	0.03	1.21	1	1	ns
V403I	no	9.84	-0.23	Decreasing	0.03	1.21	1	1	ns
V418A	no	4.66	-0.12	Decreasing	0.03	1.21	1	1	ns
N437T	no	5.73	0.41	Increasing	0.03	1.21	1	1	ns
P483L	drug	5.60	-0.70	Decreasing	0.03	1.21	1	1	ns
L735M	no	7.63	-0.49	Decreasing	0.03	1.21	1	1	ns
R791C	no	2.82	-0.76	Decreasing	0.03	1.21	1	1	ns
I795L	no	6.22	-0.44	Decreasing	0.03	1.21	1	1	ns
I795V	no	6.22	0.33	Increasing	0.03	1.21	1	1	ns
R882S	no	9.03	-0.24	Decreasing	0.03	1.21	1	1	ns
P894A	no	5.32	-0.45	Decreasing	0.03	1.21	1	1	ns
V895I	no	4.60	0.07	Increasing	0.03	1.21	1	1	ns
H1028L	no	3.64	-0.59	Decreasing	0.03	1.21	1	1	ns
P1107Q	no	5.01	0.13	Increasing	0.03	1.21	1	1	ns
K1116N	no	5.38	0.77	Increasing	0.03	1.21	1	1	ns

F367L	no	7.82	-0.18	Decreasing	0.04	1.21	1	1	ns
I414V	no	3.58	-1.02	Decreasing	0.04	1.21	1	1	ns
T427S	no	5.91	0.08	Increasing	0.04	1.21	1	1	ns
S428T	drug	7.24	0.00	Decreasing	0.04	1.21	1	1	ns
A686E	no	4.71	0.54	Increasing	0.04	1.21	1	1	ns
I889V	no	3.63	0.27	Increasing	0.04	1.21	1	1	ns
H1028N	no	3.64	-0.29	Decreasing	0.04	1.21	1	1	ns
D792T	no	3.57	0.05	Increasing	0.05	1.21	1	1	ns
V1091L	no	6.53	-0.17	Decreasing	0.05	1.21	1	1	ns
L831M	no	10.00	-0.28	Decreasing	0.06	1.21	1	1	ns
A1002S	no	7.69	0.38	Increasing	0.06	1.21	1	1	ns
M1003T	no	7.85	-0.18	Decreasing	0.08	1.21	1	1	ns
R225K	no	7.40	0.21	Increasing	0.09	1.21	1	1	ns
S495A	no	7.86	-0.08	Decreasing	0.09	1.21	1	1	ns
A823S	no	9.38	0.24	Increasing	0.10	1.21	1	1	ns
P719S	no	3.30	-0.22	Decreasing	0.12	1.21	1	1	ns
I1035M	no	3.41	0.69	Increasing	0.14	1.21	1	1	ns
E1108D	no	4.30	0.54	Increasing	0.14	1.21	1	1	ns
H1028R	no	3.64	0.40	Increasing	0.15	1.21	1	1	ns
Q975K	no	4.27	0.12	Increasing	0.34	1.21	0.73	1	ns
T1040I	no	3.63	0.75	Increasing	0.36	1.21	1	1	ns
T1018A	no	3.09	-1.03	Decreasing	0.50	1.21	1	1	ns
V262A	no	9.21	0.04	Increasing	0.59	1.01	1	1	ns
M390T	no	5.71	-0.49	Decreasing	0.10	0.97	1	1	ns
V895Q	no	4.60	0.46	Increasing	0.27	0.97	1	1	ns
C681S	no	3.21	-0.22	Decreasing	0.28	0.97	1	1	ns

M1003V	no	7.85	-0.29	Decreasing	0.26	0.80	1	1	ns
D190E	no	3.64	-0.01	Decreasing	0.59	0.69	0.61	0.98	ns
M121V	no	8.44	-0.19	Decreasing	0.02	0.60	1	1	ns
N163D	no	9.54	0.13	Increasing	0.02	0.60	1	1	ns
D265N	no	6.03	-0.28	Decreasing	0.02	0.60	1	1	ns
R661Q	no	8.39	-0.04	Decreasing	0.02	0.60	1	1	ns
E721K	no	2.75	-0.96	Decreasing	0.02	0.60	1	1	ns
I783V	no	9.24	-0.45	Decreasing	0.02	0.60	1	1	ns
I786V	no	3.30	0.07	Increasing	0.02	0.60	1	1	ns
D875V	no	4.18	0.57	Increasing	0.02	0.60	1	1	ns
A880S	no	8.30	0.32	Increasing	0.02	0.60	1	1	ns
Q975R	no	4.27	-0.09	Decreasing	0.02	0.60	1	1	ns
A998V	no	5.22	-0.13	Decreasing	0.02	0.60	1	1	ns
S1124A	no	3.27	-0.14	Decreasing	0.02	0.60	1	1	ns
K711Q	no	8.61	0.01	Increasing	0.03	0.60	1	1	ns
D792S	no	3.57	0.08	Increasing	0.03	0.60	1	1	ns
E854D	no	3.92	0.02	Increasing	0.03	0.60	1	1	ns
G463S	no	5.83	-0.14	Decreasing	0.04	0.60	1	1	ns
A977D	no	5.77	0.56	Increasing	0.04	0.60	1	1	ns
L372I	no	9.60	0.00	Decreasing	0.05	0.60	1	1	ns
A559G	no	7.68	-0.21	Decreasing	0.05	0.60	1	1	ns
D1033E	no	5.62	-0.07	Decreasing	0.07	0.60	1	1	ns
I644V	no	7.49	-0.15	Decreasing	0.09	0.60	1	1	ns
E737Q	no	3.53	-0.08	Decreasing	0.09	0.60	1	1	ns
D688Q	no	5.95	-0.24	Decreasing	0.10	0.60	1	1	ns
C681R	no	3.21	0.83	Increasing	0.11	0.60	1	1	ns

L1119I	no	4.67	0.29	Increasing	0.12	0.60	1	1	ns
L879M	no	7.31	-0.01	Decreasing	0.13	0.60	1	1	ns
M1070L	no	3.50	0.88	Increasing	0.18	0.60	0.73	1	ns
A998G	no	5.22	-0.46	Decreasing	0.20	0.60	1	1	ns
A603S	no	8.11	0.38	Increasing	0.24	0.60	1	1	ns
D704N	no	9.80	-0.21	Decreasing	0.41	0.54	0.53	0.91	ns
T482S	no	3.43	-0.55	Decreasing	0.08	0.48	0.68	1	ns
S450N	drug	7.97	0.04	Increasing	0.03	0.40	0.56	0.91	ns
V613I	no	3.64	0.46	Increasing	0.03	0.40	0.56	0.91	ns
S615P	no	8.32	0.03	Increasing	0.03	0.40	0.56	0.91	ns
E616D	no	9.62	-0.02	Decreasing	0.03	0.40	0.56	0.91	ns
P719T	no	3.30	-0.56	Decreasing	0.03	0.40	0.56	0.91	ns
A857T	no	3.27	0.20	Increasing	0.03	0.40	0.56	0.91	ns
V1091A	no	6.53	-0.08	Decreasing	0.03	0.40	0.56	0.91	ns
Q409N	no	3.14	-0.25	Decreasing	0.04	0.40	0.56	0.91	ns
P471T	no	4.09	-0.61	Decreasing	0.04	0.40	0.56	0.91	ns
M666L	no	7.28	0.36	Increasing	0.04	0.40	0.56	0.91	ns
E737A	no	3.53	-0.73	Decreasing	0.04	0.40	0.56	0.91	ns
V740M	no	8.04	-0.26	Decreasing	0.04	0.40	0.56	0.91	ns
S1009T	no	2.64	1.39	Increasing	0.04	0.40	0.56	0.91	ns
S1134K	no	3.63	0.01	Increasing	0.04	0.40	0.56	0.91	ns
A586S	no	9.70	0.02	Increasing	0.06	0.40	0.56	0.91	ns
I717V	no	3.15	-0.54	Decreasing	0.07	0.40	0.56	0.91	ns
E563D	no	6.88	0.03	Increasing	0.08	0.40	0.56	0.91	ns
L1141M	no	4.42	0.47	Increasing	0.08	0.40	0.56	0.91	ns
E1145D	no	2.87	0.60	Increasing	0.09	0.40	0.56	0.91	ns

E738D	no	5.40	0.36	Increasing	0.11	0.40	0.56	0.91	ns
G890S	no	3.55	0.59	Increasing	0.12	0.40	0.68	1	ns
M1022I	no	9.04	-0.07	Decreasing	0.14	0.40	0.56	0.91	ns
T742S	no	7.46	-0.05	Decreasing	0.16	0.40	0.56	0.91	ns
A670E	no	8.28	0.34	Increasing	0.24	0.40	0.68	1	ns
Q1071E	no	3.42	1.47	Increasing	0.38	0.37	0.26	0.58	ns
N733Q	no	4.88	0.16	Increasing	0.19	0.34	0.45	0.82	ns
D108E	no	8.30	0.06	Increasing	0.55	0.32	0.18	0.58	ns
V774S	no	5.80	-0.11	Decreasing	0.04	0.30	0.33	0.61	ns
L735V	no	7.63	-0.44	Decreasing	0.05	0.30	0.33	0.61	ns
I652V	no	7.23	-0.24	Decreasing	0.07	0.30	0.33	0.61	ns
S874T	no	6.04	-0.26	Decreasing	0.08	0.30	0.33	0.61	ns
Y1015D	no	2.86	-1.14	Decreasing	0.09	0.30	0.33	0.61	ns
E1143D	no	1.71	0.88	Increasing	0.09	0.30	0.33	0.61	ns
D107E	no	6.02	-0.08	Decreasing	0.10	0.30	0.33	0.61	ns
V418T	no	4.66	0.07	Increasing	0.10	0.30	0.33	0.61	ns
R791T	no	2.82	-0.41	Decreasing	0.10	0.30	0.46	0.84	ns
D792A	no	3.57	-0.31	Decreasing	0.15	0.30	0.46	0.84	ns
L1122M	no	4.00	0.02	Increasing	0.15	0.30	0.33	0.61	ns
M1045L	no	3.50	0.45	Increasing	0.48	0.30	0.33	0.61	ns
I1046V	no	2.45	0.74	Increasing	0.48	0.30	0.33	0.61	ns
A902P	no	2.59	0.41	Increasing	0.52	0.30	0.12	0.46	ns
T1018S	no	3.09	-0.18	Decreasing	0.43	0.28	0.08	0.35	ns
T261I	no	8.15	0.09	Increasing	0.08	0.24	0.33	0.61	ns
S458T	drug	6.48	0.64	Increasing	0.09	0.24	0.33	0.61	ns
C1067V	no	3.63	1.15	Increasing	0.27	0.24	0.19	0.58	ns

E852D	no	5.35	0.10	Increasing	0.40	0.24	0.19		0.58	ns
L464M	no	3.19	-0.23	Decreasing	0.37	0.22	0.2		0.58	ns
H723L	no	3.26	-1.00	Decreasing	0.09	0.20	0.19		0.58	ns
M1022T	no	9.04	-0.01	Decreasing	0.16	0.20	0.13		0.48	ns
V784I	no	7.29	-0.11	Decreasing	0.20	0.20	0.13		0.48	ns
I717F	no	3.15	2.06	Increasing	0.21	0.20	0.19		0.58	ns
I770V	no	6.00	0.00	Decreasing	0.88	0.20	<0.001		0.02	*
S201G	no	3.72	-0.21	Decreasing	0.19	0.17	0.11		0.45	ns
S441A	no	8.98	-0.38	Decreasing	0.95	0.17	<0.0001	<0.001		***
S388L	no	6.44	0.03	Increasing	1.79	0.17	<0.001		0.01	**
G890D	no	3.55	-1.18	Decreasing	0.08	0.15	0.12		0.45	ns
V1017I	no	4.27	0.77	Increasing	0.32	0.15	0.03		0.19	ns
T1078A	no	2.40	-1.54	Decreasing	0.34	0.13	0.01		0.09	ns
V1091T	no	6.53	-0.22	Decreasing	0.29	0.12	0.01		0.1	ns
V740I	no	8.04	-0.42	Decreasing	0.16	0.11	0.04		0.23	ns
N1105D	no	2.53	-0.20	Decreasing	0.30	0.11	0.01		0.06	ns
E460D	no	3.33	0.52	Increasing	0.51	0.11	0.01		0.06	ns
D755E	no	7.14	0.01	Increasing	0.13	0.10	0.02		0.15	ns
A596S	no	3.49	0.64	Increasing	0.14	0.10	0.02		0.15	ns
I751V	no	8.70	-0.27	Decreasing	0.62	0.10	<0.001		0.03	*
P810S	no	0.73	-0.29	Decreasing	0.33	0.09	<0.001		0.02	*
L449Q	no	8.82	-0.44	Decreasing	0.53	0.04	<0.001		0.01	**
D108G	no	8.30	-0.11	Decreasing	0.01	NA	NA	NA		ns
V129I	no	8.13	0.16	Increasing	0.01	NA	NA	NA		ns
N163I	no	9.54	-0.07	Decreasing	0.01	NA	NA	NA		ns
V196A	no	9.99	-0.20	Decreasing	0.01	NA	NA	NA		ns

P200S	no	5.11	0.22	Increasing	0.01	NA	NA	NA	ns
E207G	no	7.00	-0.28	Decreasing	0.01	NA	NA	NA	ns
K212E	no	4.26	-0.21	Decreasing	0.01	NA	NA	NA	ns
G263S	no	9.92	-0.06	Decreasing	0.01	NA	NA	NA	ns
L269M	no	4.33	-0.01	Decreasing	0.01	NA	NA	NA	ns
S285A	no	8.19	-0.02	Decreasing	0.01	NA	NA	NA	ns
F294Y	no	8.47	-0.06	Decreasing	0.01	NA	NA	NA	ns
K296N	no	9.92	0.05	Increasing	0.01	NA	NA	NA	ns
D365E	no	2.51	-0.33	Decreasing	0.01	NA	NA	NA	ns
L372M	no	9.60	-0.42	Decreasing	0.01	NA	NA	NA	ns
R373H	no	6.24	-0.03	Decreasing	0.01	NA	NA	NA	ns
V385T	no	7.45	-0.08	Decreasing	0.01	NA	NA	NA	ns
Q401L	no	7.77	-0.08	Decreasing	0.01	NA	NA	NA	ns
D402E	no	7.31	0.13	Increasing	0.01	NA	NA	NA	ns
R415L	no	3.60	-0.55	Decreasing	0.01	NA	NA	NA	ns
A419T	no	8.59	0.08	Increasing	0.01	NA	NA	NA	ns
I421L	no	7.81	-0.21	Decreasing	0.01	NA	NA	NA	ns
E423A	no	5.21	-0.15	Decreasing	0.01	NA	NA	NA	ns
Q429R	drug	3.39	0.13	Increasing	0.01	NA	NA	NA	ns
F433L	drug	7.42	-0.49	Decreasing	0.01	NA	NA	NA	ns
D435L	drug	6.61	-0.33	Decreasing	0.01	NA	NA	NA	ns
Q436N	no	9.63	-0.25	Decreasing	0.01	NA	NA	NA	ns
S441P	no	8.98	-0.47	Decreasing	0.01	NA	NA	NA	ns
S441W	no	8.98	-0.64	Decreasing	0.01	NA	NA	NA	ns
L452M	drug	6.98	-0.29	Decreasing	0.01	NA	NA	NA	ns
L452R	drug	6.98	-0.15	Decreasing	0.01	NA	NA	NA	ns

R459H	drug	3.96	-0.07	Decreasing	0.01	NA	NA	NA	ns
V469M	no	2.33	-1.25	Decreasing	0.01	NA	NA	NA	ns
P471H	no	4.09	-0.28	Decreasing	0.01	NA	NA	NA	ns
S472C	no	5.62	-0.39	Decreasing	0.01	NA	NA	NA	ns
P479T	no	2.46	-0.87	Decreasing	0.01	NA	NA	NA	ns
T482A	no	3.43	-1.00	Decreasing	0.01	NA	NA	NA	ns
P486S	no	5.36	0.06	Increasing	0.01	NA	NA	NA	ns
Q537K	no	3.37	-0.54	Decreasing	0.01	NA	NA	NA	ns
S540A	no	8.43	-0.04	Decreasing	0.01	NA	NA	NA	ns
S540P	no	8.43	0.17	Increasing	0.01	NA	NA	NA	ns
V553G	no	8.26	-0.13	Decreasing	0.01	NA	NA	NA	ns
L554V	no	5.21	-0.18	Decreasing	0.01	NA	NA	NA	ns
G560D	no	4.83	-0.01	Decreasing	0.01	NA	NA	NA	ns
R578C	no	6.06	0.16	Increasing	0.01	NA	NA	NA	ns
S582L	no	9.74	-0.80	Decreasing	0.01	NA	NA	NA	ns
F590C	no	6.58	-0.84	Decreasing	0.01	NA	NA	NA	ns
F590Y	no	6.58	-0.32	Decreasing	0.01	NA	NA	NA	ns
D594A	no	2.83	-1.41	Decreasing	0.01	NA	NA	NA	ns
D595E	no	3.47	-1.06	Decreasing	0.01	NA	NA	NA	ns
N597S	no	3.14	-0.16	Decreasing	0.01	NA	NA	NA	ns
S615A	no	8.32	-0.23	Decreasing	0.01	NA	NA	NA	ns
S615R	no	8.32	0.01	Increasing	0.01	NA	NA	NA	ns
R671G	no	8.35	-0.32	Decreasing	0.01	NA	NA	NA	ns
N673H	no	8.42	-0.02	Decreasing	0.01	NA	NA	NA	ns
H674P	no	6.48	-0.35	Decreasing	0.01	NA	NA	NA	ns
A686T	no	4.71	0.20	Increasing	0.01	NA	NA	NA	ns

D704E	no	9.80	-0.05	Decreasing	0.01	NA	NA	NA	ns
V715A	no	8.64	0.17	Increasing	0.01	NA	NA	NA	ns
A728G	no	4.05	-0.81	Decreasing	0.01	NA	NA	NA	ns
S732N	no	8.51	-0.15	Decreasing	0.01	NA	NA	NA	ns
S732T	no	8.51	-0.32	Decreasing	0.01	NA	NA	NA	ns
N733H	no	4.88	0.08	Increasing	0.01	NA	NA	NA	ns
N733K	no	4.88	-0.35	Decreasing	0.01	NA	NA	NA	ns
E737K	no	3.53	0.04	Increasing	0.01	NA	NA	NA	ns
E738A	no	5.40	-0.80	Decreasing	0.01	NA	NA	NA	ns
E738Q	no	5.40	-0.17	Decreasing	0.01	NA	NA	NA	ns
V740E	no	8.04	-0.09	Decreasing	0.01	NA	NA	NA	ns
V740L	no	8.04	-0.16	Decreasing	0.01	NA	NA	NA	ns
D755A	no	7.14	-0.11	Decreasing	0.01	NA	NA	NA	ns
K757E	no	4.20	-0.28	Decreasing	0.01	NA	NA	NA	ns
V774A	no	5.80	0.18	Increasing	0.01	NA	NA	NA	ns
V774T	no	5.80	-0.33	Decreasing	0.01	NA	NA	NA	ns
L775I	no	9.04	-0.30	Decreasing	0.01	NA	NA	NA	ns
K799N	no	8.27	-0.33	Decreasing	0.01	NA	NA	NA	ns
T806S	no	3.55	0.06	Increasing	0.01	NA	NA	NA	ns
G836C	no	5.25	-0.33	Decreasing	0.01	NA	NA	NA	ns
G843D	no	4.88	0.59	Increasing	0.01	NA	NA	NA	ns
I844V	no	7.87	-0.27	Decreasing	0.01	NA	NA	NA	ns
P856A	no	4.64	-0.02	Decreasing	0.01	NA	NA	NA	ns
I892V	no	8.20	-0.21	Decreasing	0.01	NA	NA	NA	ns
L893N	no	7.59	-0.31	Decreasing	0.01	NA	NA	NA	ns
V895A	no	4.60	-0.30	Decreasing	0.01	NA	NA	NA	ns

V895H	no	4.60	0.25	Increasing	0.01	NA	NA	NA	ns
P899L	no	6.27	-0.56	Decreasing	0.01	NA	NA	NA	ns
L901M	no	4.65	0.47	Increasing	0.01	NA	NA	NA	ns
A902T	no	2.59	0.20	Increasing	0.01	NA	NA	NA	ns
P906Q	no	4.65	0.32	Increasing	0.01	NA	NA	NA	ns
I922T	no	3.45	-1.58	Decreasing	0.01	NA	NA	NA	ns
L926F	no	4.22	1.57	Increasing	0.01	NA	NA	NA	ns
H935R	no	8.27	-0.12	Decreasing	0.01	NA	NA	NA	ns
P969S	no	9.26	-0.14	Decreasing	0.01	NA	NA	NA	ns
A977V	no	5.77	0.01	Increasing	0.01	NA	NA	NA	ns
Q980P	no	3.65	-0.20	Decreasing	0.01	NA	NA	NA	ns
R990C	no	2.38	-0.77	Decreasing	0.01	NA	NA	NA	ns
D993G	no	5.65	-0.30	Decreasing	0.01	NA	NA	NA	ns
D993H	no	5.65	0.11	Increasing	0.01	NA	NA	NA	ns
L995V	no	8.12	-0.17	Decreasing	0.01	NA	NA	NA	ns
A998T	no	5.22	0.29	Increasing	0.01	NA	NA	NA	ns
G1000S	no	4.87	-1.40	Decreasing	0.01	NA	NA	NA	ns
A1002T	no	7.69	0.46	Increasing	0.01	NA	NA	NA	ns
M1003Q	no	7.85	-0.13	Decreasing	0.01	NA	NA	NA	ns
E1011A	no	2.93	-1.24	Decreasing	0.01	NA	NA	NA	ns
E1011Q	no	2.93	-1.07	Decreasing	0.01	NA	NA	NA	ns
P1012K	no	3.56	-0.06	Decreasing	0.01	NA	NA	NA	ns
Y1015F	no	2.86	-0.87	Decreasing	0.01	NA	NA	NA	ns
Y1015S	no	2.86	-0.74	Decreasing	0.01	NA	NA	NA	ns
Y1021F	no	8.55	-0.26	Decreasing	0.01	NA	NA	NA	ns
M1022Q	no	9.04	-0.10	Decreasing	0.01	NA	NA	NA	ns

H1028Q	no	3.64	-0.25	Decreasing	0.01	NA	NA	NA	ns
H1028Y	no	3.64	0.17	Increasing	0.01	NA	NA	NA	ns
I1035L	no	3.41	-0.62	Decreasing	0.01	NA	NA	NA	ns
A1037S	no	3.52	0.46	Increasing	0.01	NA	NA	NA	ns
C1067W	no	3.63	0.19	Increasing	0.01	NA	NA	NA	ns
A1072G	no	3.20	-0.23	Decreasing	0.01	NA	NA	NA	ns
Y1073S	no	2.98	-2.45	Decreasing	0.01	NA	NA	NA	ns
Y1077H	no	2.79	0.31	Increasing	0.01	NA	NA	NA	ns
Y1077W	no	2.79	-0.15	Decreasing	0.01	NA	NA	NA	ns
T1078I	no	2.40	-0.87	Decreasing	0.01	NA	NA	NA	ns
Q1080H	no	3.16	0.16	Increasing	0.01	NA	NA	NA	ns
L1082I	no	3.19	-0.84	Decreasing	0.01	NA	NA	NA	ns
L1082M	no	3.19	0.36	Increasing	0.01	NA	NA	NA	ns
L1083M	no	3.40	-1.25	Decreasing	0.01	NA	NA	NA	ns
I1085T	no	5.07	-1.16	Decreasing	0.01	NA	NA	NA	ns
I1085V	no	5.07	0.87	Increasing	0.01	NA	NA	NA	ns
V1091H	no	6.53	0.15	Increasing	0.01	NA	NA	NA	ns
V1091P	no	6.53	-0.02	Decreasing	0.01	NA	NA	NA	ns
R1093S	no	4.66	-0.75	Decreasing	0.01	NA	NA	NA	ns
V1096L	no	3.91	-0.72	Decreasing	0.01	NA	NA	NA	ns
P1107D	no	5.01	0.43	Increasing	0.01	NA	NA	NA	ns
P1107S	no	5.01	-0.03	Decreasing	0.01	NA	NA	NA	ns
G1110A	no	3.88	-0.37	Decreasing	0.01	NA	NA	NA	ns
G1110S	no	3.88	-0.20	Decreasing	0.01	NA	NA	NA	ns
Q1123R	no	3.15	1.55	Increasing	0.01	NA	NA	NA	ns
I1139V	no	3.50	0.50	Increasing	0.01	NA	NA	NA	ns

E1147D	no	2.89	0.53	Increasing	0.01	NA	NA	NA	ns
K37T	no	8.98	-0.04	Decreasing	0.02	NA	NA	NA	ns
R105H	no	9.80	0.00	Decreasing	0.02	NA	NA	NA	ns
M148I	no	5.27	0.40	Increasing	0.02	NA	NA	NA	ns
I220V	no	8.15	-0.05	Decreasing	0.02	NA	NA	NA	ns
E423G	no	5.21	-0.16	Decreasing	0.02	NA	NA	NA	ns
G453A	drug	7.91	-0.23	Decreasing	0.02	NA	NA	NA	ns
T482I	no	3.43	-0.81	Decreasing	0.02	NA	NA	NA	ns
G492S	no	6.34	-0.34	Decreasing	0.02	NA	NA	NA	ns
R552C	no	7.55	-0.29	Decreasing	0.02	NA	NA	NA	ns
E592D	no	3.46	-0.15	Decreasing	0.02	NA	NA	NA	ns
E721D	no	2.75	-0.28	Decreasing	0.02	NA	NA	NA	ns
D752Y	no	5.09	0.34	Increasing	0.02	NA	NA	NA	ns
N769T	no	3.64	-0.23	Decreasing	0.02	NA	NA	NA	ns
V774E	no	5.80	-0.07	Decreasing	0.02	NA	NA	NA	ns
A776V	no	5.83	0.03	Increasing	0.02	NA	NA	NA	ns
L778Q	no	6.91	-0.45	Decreasing	0.02	NA	NA	NA	ns
L831R	no	10.00	-0.30	Decreasing	0.02	NA	NA	NA	ns
K840T	no	3.92	0.30	Increasing	0.02	NA	NA	NA	ns
V841A	no	3.80	-1.04	Decreasing	0.02	NA	NA	NA	ns
Q869N	no	5.14	-0.48	Decreasing	0.02	NA	NA	NA	ns
S874Q	no	6.04	0.64	Increasing	0.02	NA	NA	NA	ns
M1003R	no	7.85	0.10	Increasing	0.02	NA	NA	NA	ns
L1027Q	no	6.52	-0.31	Decreasing	0.02	NA	NA	NA	ns
V1094T	no	4.16	0.06	Increasing	0.02	NA	NA	NA	ns
D265V	no	6.03	-0.09	Decreasing	0.03	NA	NA	NA	ns

S450M	drug	7.97	-0.12	Decreasing	0.03	NA	NA	NA	ns
L464V	no	3.19	-0.62	Decreasing	0.03	NA	NA	NA	ns
L815V	no	3.90	-0.86	Decreasing	0.03	NA	NA	NA	ns
A857P	no	3.27	0.30	Increasing	0.03	NA	NA	NA	ns
S1009C	no	2.64	-1.31	Decreasing	0.03	NA	NA	NA	ns
I767M	no	8.87	-0.37	Decreasing	0.04	NA	NA	NA	ns
G890A	no	3.55	-0.28	Decreasing	0.04	NA	NA	NA	ns
K1054R	no	3.12	0.61	Increasing	0.04	NA	NA	NA	ns
Q429L	drug	3.39	-0.11	Decreasing	0.05	NA	NA	NA	ns
G463A	no	5.83	-0.40	Decreasing	0.05	NA	NA	NA	ns
I696V	no	7.52	-0.37	Decreasing	0.05	NA	NA	NA	ns
E773D	no	6.52	-0.01	Decreasing	0.05	NA	NA	NA	ns
E789D	no	0.68	-0.29	Decreasing	0.05	NA	NA	NA	ns
R791G	no	2.82	-0.64	Decreasing	0.05	NA	NA	NA	ns
I220L	no	8.15	-0.31	Decreasing	0.06	NA	NA	NA	ns
P471S	no	4.09	-0.56	Decreasing	0.06	NA	NA	NA	ns
S493T	no	7.37	-0.13	Decreasing	0.06	NA	NA	NA	ns
V774M	no	5.80	-0.24	Decreasing	0.06	NA	NA	NA	ns
E207K	no	7.00	-0.38	Decreasing	0.07	NA	NA	NA	ns
Y725F	no	3.28	0.56	Increasing	0.07	NA	NA	NA	ns
D875I	no	4.18	0.92	Increasing	0.07	NA	NA	NA	ns
T1090I	no	3.55	0.34	Increasing	0.07	NA	NA	NA	ns
I1106L	no	3.10	-1.01	Decreasing	0.07	NA	NA	NA	ns
L1118V	no	3.85	-1.68	Decreasing	0.10	NA	NA	NA	ns
I1139T	no	3.50	-1.02	Decreasing	0.12	NA	NA	NA	ns
H593N	no	2.83	0.82	Increasing	0.15	NA	NA	NA	ns

L796Q	no	9.58	-0.43	Decreasing	0.22	NA	NA	NA	ns
N381H	no	8.96	0.26	Increasing	0.79	NA	NA	NA	ns

Table 8.C.1: Mutations close to RpoB RNA polymerase β subunit PPI
Six hundred and seventy four single amino acid variation (SAV) mutations lying within 10Å of the Protein-Protein interface (PPI) and their corresponding PPI affinity changes ($\Delta\Delta$G) measured by mCSM-PPI2. The estimated effect are categorised as Destabilising ($\Delta\Delta$G<0) and Stabilising ($\Delta\Delta$G>0). The genomic measures of minor allele frequency (MAF), Odds Ratio (OR), OR related P-values, and FDR adjusted P-values are shown. Statistical significance is indicated as: *P $<$ 0.05, **P $<$ 0.01, ***P $<$ 0.001, ****P $<$ 0.0001, ns: >0.05. The table is arranged by OR to show mutation with the highest OR at the top. Columns with NA indicate insufficient data to calculate OR. Abbreviations used: $\Delta\Delta$G: change in Gibbs free energy in Kcal/mol, FDR: false discovery rate, ns: not significant, RFP: rifampicin.

8.D Average stability comparisons for lineages

Lineage comparisons	Samples (n)	Adjusted P-values	*Adjusted P-values* Significance
L1 vs L2	L1 (737), L2 (5121)	<2.2e-16	****
L1 vs L3	L1 (737), L3 (1341)	<0.0001	****
L1 vs L4	L1 (737), L4 (6861)	<0.0001	****
L2 vs L3	L2 (5121), L3 (1341)	<2.2e-16	****
L2 vs L4	L2 (5121), L4 (6861)	<2.2e-16	****
L3 vs L4	L3 (1341), L4 (6861)	<0.0001	****
	Within Lineage comparisons		
L1: R vs S	R (n=366), S (n=371)	<0.0001	****
L2: R vs S	R (n=4487), S (n=634)	<2.2e-16	****
L3: R vs S	R (n=796), S (n=545)	<2.2e-16	****

Table 8.D.1: Lineage comparisons for *rpoB* mutations
Kolmogorov-Smirnoff (KS) test reporting the statistical differences in distributions between *M. tuberculosis* lineages when assessed based on average stability changes ($\Delta\Delta$G) measured by mCSM-DUET, FoldX, DeepDDG, and Dynamut2. Lineage comparisons were performed for samples containing mutations associated with sensitivity (R: Resistant, S: Sensitive). These comparisons were performed for R and S samples between and within lineages. Statistical significance thresholds used are *P<0.05, **P<0.01, ***P<0.001, ****P<0.0001. Abbreviations used: $\Delta\Delta$G: change in Gibbs free energy in Kcal/mol, Adj. P-values: Bonferroni adjusted P-values, n: number of samples, ns: not significant.

Chapter 9

Combined summary of all six gene-targets

In this chapter, I aim to integrate and summarise the findings from the individual exploration of the six gene-targets in **Chapters 3-8** to present an overview of mutational effects across the genes, their association with resistance and how these affect the protein structure with respect to stability and various (drug, NA, PPI) binding affinities. An attempt is made to understand the SAV driven resistance landscape in *M. tuberculosis* across the six structural genes in the following manner: 1) relating the findings to gene-targets classed as essential (direct targets for drug-binding) and non-essential (indirect targets for drug binding), 2) relating mutational hotspots and active sites, 3) connecting the findings to discuss the potential role of compensatory mutations, and 4) understanding the mutational effects in light of the biological unit of the gene-target (monomer, dimer, etc.).

While Alr, EmbB, and RpoB RNAP are considered essential proteins as they are the direct targets for the drugs DCS, EMB, and RFP respectively, the ancillary protein GidB along with PncA and KatG are the indirect targets for STR, PZA, and INH respectively. When considering the biological unit, GidB and PncA are monomers (single chain proteins), Alr and KatG are a homo-dimers (2 identical chains), EmbB is a hetero-trimer (3 non-identical chains), and RpoB RNAP is a hetero-hexamer (6 non-identical chains).

9.1 Direct and indirect targets

Most (≥55%) mutational consequences resulted in electrostatic changes across all gene targets, as previously observed for *alr*, *katG*, and *rpoB*,[1] and observed here for the first time for *embB* and *gidB*. These observations affirm the understanding that SAV mutations affect protein folding, binding, and other biological functions as electrostatic interactions are known to play a crucial role in protein structures.[2,3]

Similarly, the majority (≥80%) of SAVs resulted in destabilising the overall protein, as well as reducing the binding affinity to their respective drugs. This may be expected since mutations are known to alter protein stability and drug binding affinity, potentially affecting protein and bacterial fitness.[4,5] The relationship between protein stability and activity (protein function, drug-binding, etc.) is largely context dependent.[6,7] Mutational impact that results in a less stable protein increases its entropy. This less stable protein, when bound to the ligand in a particular conformation, decreases the overall entropy (i.e. increases global stability), but in turn may bind less effectively to the ligand, resulting in a deleterious effect on the ΔG of binding (i.e. reduced binding affinity). Thus, there is a clear trade-off between global stability, binding affinity, and the normal function of the protein,[8] compounding the difficulties associated with understanding the mutational impact on protein function.

The mutational consequences on drug binding affinity were largely destabilising as mentioned above, however predictions from mmCSM-lig (David Ascher, personal communication) were *always* destabilising for all mutations across all gene targets. It would be prudent to consider the methodological differences between mCSM-lig and mmCSM-lig to rule out any technical discrepancies in the resulting estimates.

The predicted functional consequences based on conservation trends from protein sequence variations, however, were different for essential and non-essential proteins. Most mutations occurred in the highly variable regions of the essential structural genes (Alr, EmbB, RpoB RNAP), with neutral functional consequences predicted on protein despite the destabilising mutational impact on all three protomers. In contrast, for non-essential structural genes (GidB, PncA, and KatG), most mutations occurred in the highly conserved regions, with predicted detrimental protein functional effects based on sequence variations,[9,10] likely associated with destabilising protomer stability. This supports the understanding that the large fitness penalty associated with detrimental functional consequences is alleviated for non-essential genes, but not necessarily for essential ones. This has been observed for essential genes *alr* and *rpoB*, and non-essential genes *katG*[1] and *pncA* (published paper as part of this project), but being reported for the first time in this manner for *embB* and *gidB*.

9.2 Active site, hotspots, and compensatory effects

SAV heterogeneity at active site residues (including co-factors and binding partners) was prominent for *gidB*, *pncA* and *katG*, owing to low fitness penalty due to their non-essential role in the TB bacillus. The reverse was observed for essential genes *alr* and *embB*, where mutational promiscuity at active site residues is thought to be accompanied by large fitness penalties reflected in their low SAV diversity.

Mutation hotspots in *alr* did not involve DCS and co-factor PLP residues, highlighting the large fitness burden associated with these sites. An interesting observation made regarding the SAV L113R which occurred most frequently and was strongly associated with DCS resistance: it occurred at a site with only a single mutation and was present only in lineages 2 and 4, which are located on different branches on the *M. tuberculosis* phylogenetic tree.[11] Together, these findings suggest this is convergent evolution for mutation L113R, occurring at a highly conserved site involved in DCS resistance. The next most frequent and prominent SAV in *alr* associated with DCS resistance was Y388D, an active residue involved in DCS and PLP binding. It is known that the active site residue Y388 forms a 2.7Å gate with residue Y295 of the opposite chain,[12] where the smaller mutant residue aspartic acid

is thought to destabilise this gate reducing PPI affinity, and the loss of van der Waals interaction with PLP is thought to reduce DCS binding affinity,[1] resulting in the resistant phenotype as observed in this analysis. Further, mutation M343T involved with DCS binding is thought to affect DCS inhibition by disrupting interactions with nearby active site residue Y388. Other prominent hotspots beyond the active site such as R397 and T399 were present at the dimer interface, and are thought to impact DCS binding by disrupting the nearby active site residues M343 and D344. All SAVs mentioned above either occur close to the drug and/or the PPI, implicating the synergistic effects of high frequency and high resistance conferring mutations. It is reasonable to consider the involvement of compensatory mutations restoring any fitness deficit associated with resistant phenotypes and those required to maintain the functioning Alr homo-dimer. With the limited DST data available for DCS, it is difficult to reason effectively on the role of selective pressure exerted by DCS and the outcome of its other SAVs. With only a small minority (<4%) of samples displaying SAVs in *alr* lent support to the slow evolution of Alr's resistance profile,[13] it was interesting to note that multiple SAVs in Alr extended beyond the active site and were associated with resistance, suggesting resistance acquisition by mechanisms other than direct DCS binding inhibition. Further, it would appear that compensatory mutations in both DCS targets (*alr*, *Ddl*), as well as concomitant mutations in clinical isolates need to be considered systematically to gain better insights into the resistance evolution of Alr.

Similarly for *embB*, hotspots were prominent for non-active site residues, with limited involvement of active site residues (M306, F330, A439, and V456). PPI sites involving CDL were largely budding resistance hotspots. The crucial active site residue M306 contained multiple SAVs (M306 →T/V/L/I) and retained the resistant phenotype, akin to observations made by other studies.[14–18] Despite this the resistance-fitness landscape is seldom as straightforward. Plinke, *et. al.*[15] found no association of SAVs at M306 with EMB resistance in MDR-TB strains, highlighting its limited use as a resistance marker only in pan-susceptible *M. tuberculosis* strains. With most (85%, n=731) mutations classed as sensitive according to DST, and the mutational ubiquity in *embB*, it appears that resistance development in *embB* involves compensatory interactions between SAVs in *embB*, as well as in the *embBAC* operon (*embB*, *embA*, and *embC* genes). This perhaps explains the occurrence of most (95%, n=18) CDL interacting residues at the PPI with single mutations or as budding resistance hotspots. Compensatory effects have been noted to occur between distant sites (>20Å),[19] and with EmbB being part of a larger hetero-trimer complex, it is logical to consider the potential role of compensatory mutations in allaying detrimental effects of resistance associated SAVs. Thus, it appears that the mutational ubiquity owing to selective pressure exerted by EMB, and the biophysical consequences of SAVs on the EmbB protein, would be better understood by considering all mutations (and their effects) in the *embBAC* operon.

More recent reports, however, present additional views on the acquisition and evolution of high-level resistance in EMB. Additional gene loci G406 and Q497, associated with EMB resistance with evidence for mutations accumulating in a stepwise manner, have been elucidated by Safi, *et. al.*.[20] Further, Pawar, *et. al.*[21] suggested glutamate racemase, which is involved in the peptidoglycan synthesis of the bacterial cell wall, as an additional target for *embB*. Extending this further, resistance to EMB (used alongside other first-line drugs) is likely to be acquired as part of a multi-step process[20] with sites beyond the active site (A201, D328), including those interacting with EmbB substrate DPA (F330, K511) displaying similar mutational heterogeneity, as observed in this study.

A similar adaptation in *rpoB* involving compensatory effects across the *rpoBAC* operon have been widely reported by others.[22–25] In this analysis, *rpoB* exhibited the greatest mutational diversity in the rifampicin resistance determining region (RRDR), as well as overall, compared with all other genes. As *rpoB* is part of the larger hetero-hexamer complex, this observation is strongly reflective of the known putative compensatory effects from *rpoC* and *rpoA* genes[22–24] in the RRDR region, and those extending beyond it.[25] Mutations in *rpoB* loci 435, 445, and 450 are among the most frequently reported and associated with RFP resistance[26,27] owing to fitness penalties being mitigated by compensatory mutations as mentioned. This was similarly observed in my current analyses, where *rpoB* S450L was the most frequent mutation associated with RFP resistance, and H445D showed the strongest association with RFP resistance, while R448E, a mutation with a severe fitness penalty,[28] was not observed in this analysis.

Additionally, *rpoB* and *katG* exhibited a much higher peak SAV count (9 for *katG* and 13 for *rpoB*) than other targets. An explanation for this is perhaps best achieved by considering the prevalence of MDR-TB strains,[29] with selection pressures exerted by multiple drugs at the same time. For example, the combinations of low-fitness cost resistance mutations: *rpoB* S450L and *katG* S315T in MDR strains, which frequently occur in clinical isolates,[30] contributes to the widespread MDR-TB burden. It stands to reason, then, that mutations becoming 'fixed' in this manner pervade, and acquire additional mutations, manifesting as mutational heterogeneity at active sites and beyond. This subsequently allows mutations to work in tandem with certain combination of mutations in resistant isolates retaining their fitness just as much and possibly more than their drug sensitive counterparts.[31]

9.3 Mutational impact around protein-protein interface and nucleic acid sites

Mutational impact on the PPI and NA affinity presented some interesting observations. Irrespective of the protein size, the prominent effects at the PPI for Alr and RpoB RNA polymerase β subunit proteins were largely destabilising. In contrast, the mutational effects on the PPI for EmbB was largely stabilising, likely due to the molecular interactions involving co-factor CDL present at the interface. For NA binding affinity in GidB and RpoB RNA polymerase β subunit proteins, sites around the NA are highly conserved, with prominent mutational impact on NA affinity being largely stabilising. Together, these findings suggest that residues involved with NA interactions play a crucial role in complex stability and protein function accompanied by a high fitness burden of resistance inducing SAV mutations in GidB,[32] and RpoB RNA polymerase β subunit. When linked to the respective mutational phenotype, these sites mostly consisted of sensitive mutations for the essential gene *rpoB* while for the non-essential gene *gidB*, these predominantly consisted of resistant mutations. For the essential *rpoB* gene, sensitive mutations around the NA site are thought to confer local fitness advantages helping to compensate for resistant mutational phenotype beyond the NA sites. Whereas for the ancillary protein GidB, fitness penalties for resistant SAVs around the conserved NA site are mitigated by its non-essential role, as well as the possibility of compensatory mutations reported in clinical aminoglycoside resistant strains, contributing to the spread of resistance.[33]

9.4 Overview of the resistance landscape

Differences in genomic and biophysical properties of resistant and sensitive SAV mutations were compared for all gene targets except *alr*, due to only two SAVs being present in the resistant group. In all other gene targets except GidB, resistant mutations occurred less frequently, evolved at a slower rate and were more conserved. Similarly, resistant mutations tend to be destabilising for protomer stability and occur closer to the drug without affecting binding affinity. For RpoB RNA polymerase β subunit and KatG proteins which include a PPI, resistant mutations were closer to their interface resulting in marginal reduced binding affinity of the complex. The same, however, was not observed in the EmbB PPI. As mentioned previously this is likely due to the prominent stabilising mutational impact in the presence of co-factor CDL at the interface. For GidB and RpoB RNA polymerase β subunit, resistant mutations were also closer to the NA without affecting its binding affinity, consistent with other findings in these analyses. Similar differences have been previously reported in PncA,[34] and,

RpoB RNA polymerase β subunit,[35] and can be confirmed here for the first time in EmbB, GidB and KatG.

It has been suggested that frequently occurring mutations do not confer extreme changes to protomer stability or ligand affinity, with mildly stabilising mutations conferring fitness advantages, allowing them to become fixed in a population.[1] Empirical assessment based on a much larger 35,000 clinical isolate dataset supports this. While these are prominent for the non-essential gene *katG*, where mutational impact on protomer stability changes are mild, the same is not observed for similar non-essential genes like *pncA* and *gidB*, where the resistant mutations have extreme impact. The fitness landscape in terms of protomer stability appears to be a gene-specific phenomenon rather than the common non-essential functionality of these genes. An additional observation is that both PncA and GidB are monomeric proteins while KatG, RpoB RNA polymerase β subunit, and EmbB have multiple chains. Therefore, it could be that these non-monomeric complexes follow a different adaptation paradigm due to the balanced interplay of molecular interactions required to maintain these functional protein complexes.

Effects related to protein flexibility at mutational sites in essential proteins were mainly low-to-mild, compared with the mild-to-moderate effects observed in non-essential proteins. For the latter, GidB displayed the highest flexibility overall, as well as around all binding partner sites. It would appear that this ancillary protein, indirectly involved with STR binding, offers greater susceptibility towards mutational tolerance as well as resistance, confirmed by observed mutational heterogeneity at these sites and with most (93%, n=493) SAVs being sensitive to STR according to DST. For KatG and PncA, mutational sites around binding partners were similarly associated with mild-to-moderate flexibility, allowing for mutational diversity at these sites without affecting drug binding affinity. When comparing mutational association with resistance at these flexible sites, sites with exclusively resistant mutations were associated with moderate flexibility in KatG, but not for PncA. It is thought that PPI interactions in the KatG homo-dimer play an important role in these dynamics, while for the monomeric protein PncA, protein flexibility allows for mutational diversity without affecting resistance. For the essential proteins, visual inspection highlighted that exclusively resistant mutation sites were associated with low flexibility, as compared with those with sensitive mutations. This suggests that for essential proteins, sites susceptible to flexibility tend to consist of sensitive mutations.

Of the four *M. tuberculosis* lineages, the largest proportion of samples across most targets came from the geographically widespread lineages 2 and 4, which are associated with greater virulence and increased transmission.[36] The exception to this was EmbB, which had a slightly greater number of

samples from lineage 1 compared with lineage 2, though retaining the highest number of samples from lineage 4. Lineages 2 and 4 also presented with limited SAV diversity compared with the more geographically restricted lineages 1 and 3, which showed higher mutation diversity except for KatG which showed low diversity overall. While lineage 1 is less virulent in relation to clinical severity than lineages 2 and 4, the relationship between lineage 3 and virulence with respect to clinical severity is less clear.[36] The lack of mutational heterogeneity in the more virulent strains potentially indicates the optimised fitness of SAVs driving resistance. On the other hand, mutational diversity in lineages 1 and 3 indicates potential adaptive processes for functional innovation in the underlying *M. tuberculosis* genome. External factors like drug concentration and host immune response could influence the number of mutations available to certain lineages, which in turn influences resistance evolution.[30]

Overall, the minimal SAV diversity across lineages displayed by KatG suggests a few selected mutations that optimise the fitness landscape to promote INH resistance development. This becomes more apparent when looking at the distribution of resistant and sensitive mutations with respect to changes in protomer stability. The distribution of resistance mutations, with marginal stability consequences is dominated by the most frequent S315T mutation associated with INH resistance with little-to-no fitness cost,[30] along with the equivalent highly frequent sensitive mutation R463L showing a marginal stabilising impact. In a similar trend, the protomer stability impact for RpoB RNA polymerase β subunit is marginal, with the highly frequent S450L mutation associated with RFP resistance stabilising the protomer. This observation is suggestive of any fitness deficits being corrected by multiple compensatory mutations from its *rpoA/B/C* genes.[30] Subsequently, the protomer stability impact of resistant mutations was marginal for essential genes like *embB* and *rpoB* (excluding *alr* due to lack of DST data). Together these findings support the fitness advantages conferred by a lack of extreme mutational effects reported previously for *katG* and *rpoB*[1] and observed currently in *embB*.

Resistance mutations showed a more extreme effect for non-essential genes like *gidB* and *pncA*. The mutational impact on protomer stability was extremely destabilising across all four lineages in GidB, but only restricted to lineage 1 for PncA. Where the fitness cost for highly destabilising resistance mutations is abated in GidB due to STR not directly binding to GidB. For PncA, it appears that there are lineage 1-specific SAV mutations driving this unique pattern for resistance mutations. Furthermore, lineage 1 is associated with less clinical severity, with certain sub-lineages having limited transmission capability.[37] Altogether, it appears that this 'ancient' lineage 1, with PncA specific resistant mutations, is reaching a distinct protomer stability equilibrium, making way for other 'modern' lineages, which would be worth investigating further.

The systematic gene-target explorations were undertaken to help improve our understanding of the interrelationship of factors associated with SAV mutational effects: local and global impact on protein structures, mutational frequency and heterogeneity, resistance hotspots, and lineage information. This investigation highlighted that resistance development involves interaction between genes including co-occurring and compensatory mutations in individual or across multiple genes, which are important to consider in understanding resistance evolution in *M. tuberculosis*. This in-depth investigation laid the groundwork for downstream ML analyses described in Chapter 11. Additionally, the analysis also afforded the opportunity to explore the SAV associated resistance landscape with respect to lineage, in part attempted in the following chapter **(Chapter 10)** as a pilot study.

References

[1] Stephanie Portelli et al. "Understanding Molecular Consequences of Putative Drug Resistant Mutations in Mycobacterium Tuberculosis". In: *Scientific Reports* (2018). ISSN: 20452322. DOI: 10.1038/s41598-018-33370-6.

[2] Lijun Quan, Qiang Lv, and Yang Zhang. "STRUM: Structure-Based Prediction of Protein Stability Changes upon Single-Point Mutation". In: *Bioinformatics* 32.19 (Oct. 2016), pp. 2936–2946. ISSN: 1367-4803. DOI: 10.1093/bioinformatics/btw361.

[3] Huan-Xiang Zhou and Xiaodong Pang. "Electrostatic Interactions in Protein Structure, Folding, Binding, and Condensation". In: *Chemical reviews* 118.4 (Feb. 28, 2018), pp. 1691–1741. ISSN: 0009-2665. DOI: 10.1021/acs.chemrev.7b00305.

[4] Nobuhiko Tokuriki et al. "The Stability Effects of Protein Mutations Appear to Be Universally Distributed". In: *Journal of Molecular Biology* 369.5 (June 2007), pp. 1318–1332. ISSN: 00222836. DOI: 10.1016/j.jmb.2007.03.069.

[5] Nobuhiko Tokuriki and Dan S Tawfik. "Stability Effects of Mutations and Protein Evolvability". In: *Current Opinion in Structural Biology* 19.5 (Oct. 2009), pp. 596–604. ISSN: 0959440X. DOI: 10.1016/j.sbi.2009.08.003.

[6] Brian K. Shoichet et al. "A Relationship Between Protein Stability and Protein Function". In: *Proceedings of the National Academy of Sciences of the United States of America* 92.2 (1995), pp. 452–456. ISSN: 0027-8424.

[7] Romain A. Studer, Benoit H. Dessailly, and Christine A. Orengo. "Residue Mutations and Their Impact on Protein Structure and Function: Detecting Beneficial and Pathogenic Changes". In: *Biochemical Journal* 449.3 (Feb. 1, 2013), pp. 581–594. ISSN: 0264-6021, 1470-8728. DOI: 10.1042/BJ20121221.

[8] Nobuhiko Tokuriki et al. "How Protein Stability and New Functions Trade Off". In: *PLoS computational biology* 4.2 (Feb. 29, 2008), e1000002. ISSN: 1553-7358. DOI: 10.1371/journal.pcbi.1000002.

[9] Yongwook Choi et al. "Predicting the Functional Effect of Amino Acid Substitutions and Indels". In: *PLoS ONE* 7.10 (Oct. 8, 2012). Ed. by Alexandre G. de Brevern, e46688. ISSN: 1932-6203. DOI: 10.1371/journal.pone.0046688.

[10] Maximilian Hecht, Yana Bromberg, and Burkhard Rost. "Better Prediction of Functional Effects for Sequence Variants". In: *BMC genomics* 16 Suppl 8 (2015), S1. ISSN: 1471-2164. DOI: 10.1186/1471-2164-16-S8-S1.

[11] Daniela Brites and Sebastien Gagneux. "Co-Evolution of Mycobacterium Tuberculosis and Homo Sapiens". In: *Immunological Reviews* 264.1 (Mar. 2015), pp. 6–24. ISSN: 1600-065X. DOI: 10.1111/imr.12264.

[12] Pierre LeMagueres et al. “The 1.9 A Crystal Structure of Alanine Racemase from Mycobacterium Tuberculosis Contains a Conserved Entryway into the Active Site”. In: *Biochemistry* 44.5 (Feb. 8, 2005), pp. 1471–1481. ISSN: 0006-2960. DOI: 10.1021/bi0486583.

[13] Dimitrios Evangelopoulos et al. “Comparative Fitness Analysis of D-cycloserine Resistant Mutants Reveals Both Fitness-Neutral and High-Fitness Cost Genotypes”. In: *Nature Communications* 10.1 (1 Sept. 13, 2019), p. 4177. ISSN: 2041-1723. DOI: 10.1038/s41467-019-12074-z.

[14] S Sreevatsan et al. “Ethambutol Resistance in Mycobacterium Tuberculosis: Critical Role of embB Mutations”. In: *Antimicrobial Agents and Chemotherapy* 41.8 (Aug. 1997), pp. 1677–1681. ISSN: 0066-4804, 1098-6596. DOI: 10.1128/AAC.41.8.1677.

[15] Claudia Plinke, Sabine Rüsch-Gerdes, and Stefan Niemann. “Significance of Mutations in embB Codon 306 for Prediction of Ethambutol Resistance in Clinical Mycobacterium Tuberculosis Isolates”. In: *Antimicrobial Agents and Chemotherapy* 50.5 (May 2006), pp. 1900–1902. ISSN: 0066-4804. DOI: 10.1128/AAC.50.5.1900-1902.2006.

[16] Betzaida Cuevas-Córdoba et al. “Mutation at embB Codon 306, a Potential Marker for the Identification of Multidrug Resistance Associated with Ethambutol in Mycobacterium Tuberculosis”. In: *Antimicrobial Agents and Chemotherapy* 59.9 (Aug. 14, 2015), pp. 5455–5462. DOI: 10.1128/AAC.00117-15.

[17] Precious Bwalya et al. “Characterization of embB Mutations Involved in Ethambutol Resistance in Multi-Drug Resistant Mycobacterium Tuberculosis Isolates in Zambia”. In: *Tuberculosis* 133 (Mar. 1, 2022), p. 102184. ISSN: 1472-9792. DOI: 10.1016/j.tube.2022.102184.

[18] S. Ramaswamy and J. M. Musser. “Molecular Genetic Basis of Antimicrobial Agent Resistance in Mycobacterium Tuberculosis: 1998 Update”. In: *Tubercle and Lung Disease: The Official Journal of the International Union Against Tuberculosis and Lung Disease* 79.1 (1998), pp. 3–29. ISSN: 0962-8479. DOI: 10.1054/tuld.1998.0002.

[19] Mark Lunzer, G. Brian Golding, and Antony M. Dean. “Pervasive Cryptic Epistasis in Molecular Evolution”. In: *PLoS genetics* 6.10 (Oct. 21, 2010), e1001162. ISSN: 1553-7404. DOI: 10.1371/journal.pgen.1001162.

[20] Hassan Safi et al. “Evolution of High-Level Ethambutol-Resistant Tuberculosis through Interacting Mutations in Decaprenylphosphoryl--d-Arabinose Biosynthetic and Utilization Pathway Genes”. In: *Nature genetics* 45.10 (Oct. 2013), pp. 1190–1197. ISSN: 1061-4036. DOI: 10.1038/ng.2743.

[21] Alka Pawar et al. “Ethambutol Targets the Glutamate Racemase of Mycobacterium Tuberculosis-an Enzyme Involved in Peptidoglycan Biosynthesis”. In: *Applied Microbiology and Biotechnology* 103.2 (Jan. 2019), pp. 843–851. ISSN: 1432-0614. DOI: 10.1007/s00253-018-9518-z.

[22] Iñaki Comas et al. “Whole-Genome Sequencing of Rifampicin-Resistant Mycobacterium Tuberculosis Strains Identifies Compensatory Mutations in RNA Polymerase Genes”. In: *Nature Genetics* 44.1 (2012), pp. 106–110. ISSN: 10614036. DOI: 10.1038/ng.1038.

[23] M. de Vos et al. “Putative Compensatory Mutations in the rpoC Gene of Rifampin-Resistant Mycobacterium Tuberculosis Are Associated with Ongoing Transmission”. In: *Antimicrobial Agents and Chemotherapy* 57.2 (Feb. 2013), pp. 827–832. ISSN: 0066-4804. DOI: 10.1128/AAC.01541-12.

[24] João Perdigão et al. “Unraveling Mycobacterium Tuberculosis Genomic Diversity and Evolution in Lisbon, Portugal, a Highly Drug Resistant Setting”. In: *BMC genomics* 15 (Nov. 18, 2014), p. 991. ISSN: 1471-2164. DOI: 10.1186/1471-2164-15-991.

[25] Nicola Casali et al. “Evolution and Transmission of Drug-Resistant Tuberculosis in a Russian Population”. In: *Nature Genetics* 46.3 (3 Mar. 2014), pp. 279–286. ISSN: 1546-1718. DOI: 10.1038/ng.2878.

[26] M. Heep et al. “Frequency of rpoB Mutations inside and Outside the Cluster I Region in Rifampin-Resistant Clinical Mycobacterium Tuberculosis Isolates”. In: *Journal of Clinical Microbiology* 39.1 (Jan. 2001), pp. 107–110. ISSN: 0095-1137. DOI: 10.1128/JCM.39.1.107-110.2001.

[27] Erik C. Böttger and Burkhard Springer. “Tuberculosis: Drug Resistance, Fitness, and Strategies for Global Control”. In: *European Journal of Pediatrics* 167.2 (Feb. 1, 2008), pp. 141–148. ISSN: 1432-1076. DOI: 10.1007/s00431-007-0606-9.

[28] Sebastien Gagneux et al. “The Competitive Cost of Antibiotic Resistance in Mycobacterium Tuberculosis”. In: *Science* 312.5782 (June 30, 2006), pp. 1944–1946. DOI: 10.1126/science.1124410.

[29] WHO Global Tuberculosis Report. *Global Tuberculosis Report 2021*. Geneva: World Health Organization, 2021.

[30] Sebastian M. Gygli et al. “Antimicrobial Resistance in Mycobacterium Tuberculosis: Mechanistic and Evolutionary Perspectives”. In: *FEMS microbiology reviews* 41.3 (May 1, 2017), pp. 354–373. ISSN: 1574-6976. DOI: 10.1093/femsre/fux011.

[31] Fernanda S. Spies et al. “Biological Cost in Mycobacterium Tuberculosis with Mutations in the rpsL, Rrs, rpoB, and katG Genes”. In: *Tuberculosis (Edinburgh, Scotland)* 93.2 (Mar. 2013), pp. 150–154. ISSN: 1873-281X. DOI: 10.1016/j.tube.2012.11.004.

[32] Daniel C. Shippy and Amin A. Fadl. “RNA Modification Enzymes Encoded by the Gid Operon: Implications in Biology and Virulence of Bacteria”. In: *Microbial Pathogenesis* 89 (Dec. 2015), pp. 100–107. ISSN: 08824010. DOI: 10.1016/j.micpath.2015.09.008.

[33] Dmitri Shcherbakov et al. “Directed Mutagenesis of Mycobacterium Smegmatis 16S rRNA to Reconstruct the in Vivo Evolution of Aminoglycoside Resistance in Mycobacterium Tuberculosis”. In: *Molecular Microbiology* 77.4 (Aug. 2010), pp. 830–840. ISSN: 1365-2958. DOI: 10.1111/j.1365-2958.2010.07218.x.

[34] Malancha Karmakar et al. *SUSPECT-PZA | Home*. 2018.

[35] Stephanie Portelli et al. “Prediction of Rifampicin Resistance beyond the RRDR Using Structure-Based Machine Learning Approaches”. In: *Scientific Reports* 10.1 (1 Oct. 22, 2020), p. 18120. ISSN: 2045-2322. DOI: 10.1038/s41598-020-74648-y.

[36] Mireia Coscolla and Sebastien Gagneux. “Consequences of Genomic Diversity in Mycobacterium Tuberculosis”. In: *Seminars in Immunology* 26.6 (Dec. 2014), pp. 431–444. ISSN: 1096-3618. DOI: 10.1016/j.smim.2014.09.012.

[37] Nguyen Thi Le Hang et al. “Phenotypic and Genotypic Features of the Mycobacterium Tuberculosis Lineage 1 Subgroup in Central Vietnam”. In: *Scientific Reports* 11.1 (1 June 30, 2021), p. 13609. ISSN: 2045-2322. DOI: 10.1038/s41598-021-92984-5.

Chapter 10

Sensitivity by lineage results

An exploratory analysis

10.1 Background

The relationships between *M. tuberculosis* lineages and virulence are being increasingly reported[1–4] including lineage specific associations with drug resistance.[5–7] Consequently, the genetic background in which mutations develop becomes important, and has been shown to play a vital role in resistance development in *M. tuberculosis*.[8–11]

Consequences of the genetic background on mutational effects can influence the level of resistance conferred for a given mutation, or include the involvement of multiple interacting genes influencing the resulting phenotype. In the context of resistance, phenotype refers to the ability of a given mutation to confer drug resistance. For example, lineage 2 strains with mutations in *katG* and *inhA* confer high and low level INH resistance respectively, compared with lineage 1 stains which predominantly show mutations in *inhA*.[10] This phenomenon of genetic influences on a mutation's phenotypic effects is known as epistasis or genetic interactions.[12]

Epistasis has important consequences on organismal fitness, which is defined as the ability of an organism to survive and reproduce in a given environment. Epistatic interactions can result in beneficial or deleterious phenotypes, and are classed into positive, negative, sign, and reciprocal sign epistasis. Epistasis for fitness is defined in relation to the difference between fitness of a double (or multiple) mutant bacterial isolate compared with the fitness of the constituent single mutants.

Positive epistasis results in net higher fitness, while negative epistasis results in a net lower fitness, while for sign epistasis, the genetic background of the mutations plays a central role with the resulting effects being deleterious, beneficial, or neutral. Interaction between compensatory and drug resistance mutations serves as an example of sign epistasis.[13,14] Furthermore, reciprocal sign epistasis highlights an extreme form of sign epistasis, where individually deleterious mutations become beneficial in combination, while individually beneficial mutations are rendered deleterious in combination.[12,15]

Although it is recognised that epistasis plays an important role in resistance development in *M. tuberculosis*, studies investigating this systematically have been mainly limited to compensatory mutations in RFP[16–18] and STR resistance,[19,20] with a recent study demonstrating the role of epistasis in four important anti-TB drugs used for MDR-TB treatment.[11]

Motivated by this growing interest in *M. tuberculosis* strain diversity, and its consequences on drug resistance development,[8,21] the project sought to conduct a pilot analysis to investigate mutations with differing drug sensitivities (resistant/sensitive) across lineages. The aim was to highlight if, and which, mutations confer resistance in one lineage but not in a different lineage. Only mutations in

lineages 1-4 were considered, as other lineages had a low number of samples (<20). This was an exploratory, qualitative analysis carried out in the interest of exploring the resistance landscape and to inform future work related to AMR in TB.

10.2 Methods

Mutations with differing sensitivities, defined as a given mutation showing both resistant and sensitive phenotypes, across the four *M. tuberculosis* lineages, were extracted from the six antibiotic target genes analysed in the project. There was no minimum threshold used to determine these mutations, and as such the list of mutations displaying these differing sensitivities is exhaustive.

For each mutation, proportions of resistant and sensitive samples across the four lineages were compared using the Fisher's exact test to identify statistically significant mutations. Statistical significance thresholds used were: .P<0.10, *P<0.05, **P<0.01, ***P<0.001, ****P<0.0001. No correction for multiple testing was performed due to the exploratory nature of this work. Irrespective of the statistical significance, and in line with the qualitative nature of this analysis, emphasis was placed upon identifying mutations that were represented equivalently with respect to their drug sensitivity across the lineages. This would highlight the importance of considering strain diversity, i.e. lineage, to assess mutational effects on drug resistance.

10.3 Results

A summary of the number of mutations with differing drug sensitivity across lineages in each gene-drug combination is shown in **Table 1**.

Mutations with differing drug sensitivity in *pncA*: Nearly 30% (125/419) of mutations were found to be different with respect to PZA sensitivity across the four lineages **(Figure 1A** and **1B)**, which included all active site residues but were not limited to them. Eleven mutations were statistically significant with respect to differing PZA sensitivity across the lineages (P<0.05), and these were predominantly resistant **(Figure 1)** in contrast to what is seen in sections below (*embB*, *gidB*, *katG*, and *rpoB*) where mutations are seen that lead to resistance in one strain (lineage), but the identical mutation doesn't lead to resistance in another strain.

***gene*-drug**	Total SAVs	**L1-L4 (n)** SAVs with different sensitivities	Statistically significant (P<0.05)
alr-cycloserine (DCS)	240	3	1
embB-ethambutol (EMB)	744	67	11
gidB-streptomycin (STR)	520	94	5
katG-isoniazid (INH)	772	68	3
pncA-pyrazinamide (PZA)	419	125	11
rpoB-rifampicin (RFP)	1046	138	9
Total	3741	495	40

Table 1: Number of SAVs in *M. tuberculosis* lineages 1-4
Number of single amino acid variations (SAV) with differing respective drug sensitivity (resistant/sensitive). Abbreviations used: L1-L4: Lineage 1-Lineage 4.

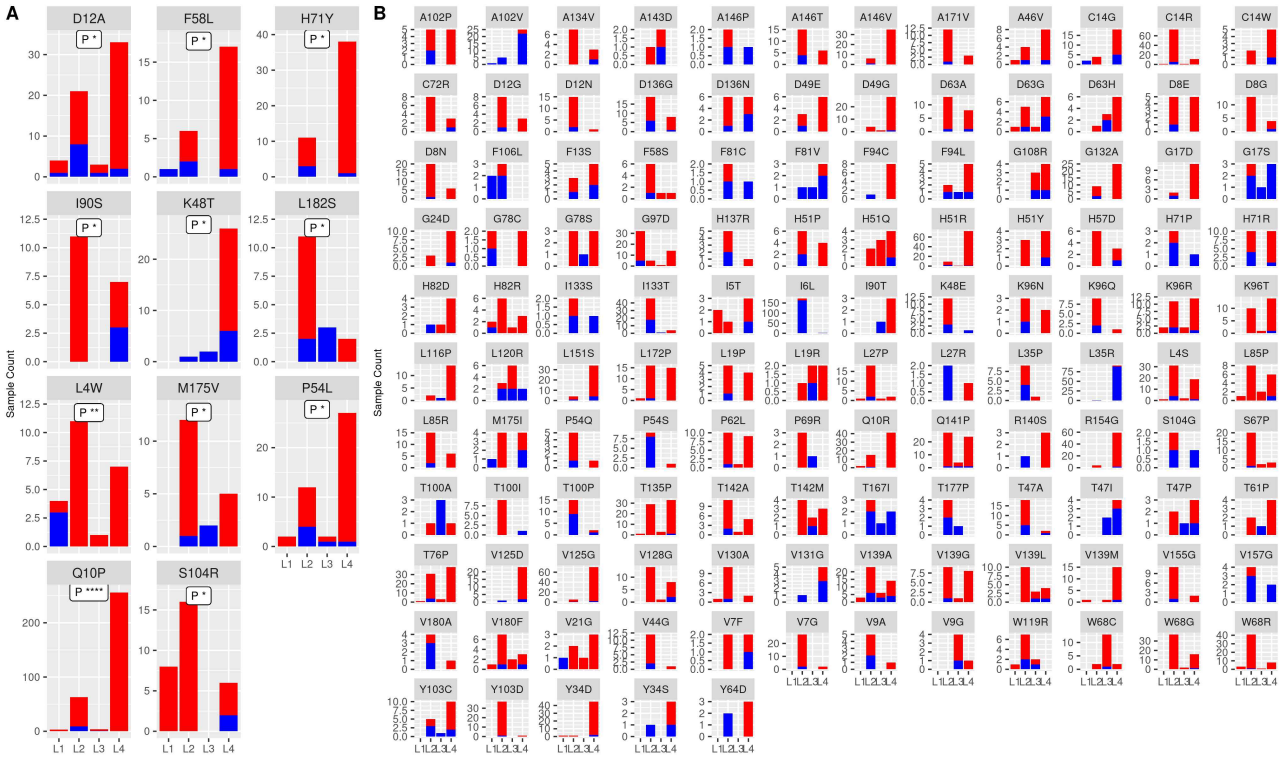


Figure 1: Mutations in *pncA* with differing sensitivities
Mutations in *M. tuberculosis* gene *pncA*, with different sensitivities across lineages 1-4. Red denotes resistant and blue denotes sensitive samples. Fisherss exact test was used to compare mutation proportions across lineages, and statistical significance was assessed according to the thresholds: .P<0.10, *P<0.05, **P<0.01, ***P<0.001, ****P<0.0001 **A)** Statistically significant mutations, **B)** Mutations that did not meet statistical significance.

Mutations with differing drug sensitivity in *alr*: A very small percentage (1.3%, 3/240) of mutations were found to be different with respect to DCS sensitivity across the four lineages **(Figure 2A** and **2B)**. These were mutations: M343T, L113R and Y388D, where M343 and Y388 wild-type residues are involved in the active site, while L113R though not an active site residue **(Figure 2B)**, is the most frequent mutation with the highest association with DCS resistance. Only a single mutation M343T was statistically significant with respect to DCS sensitivity across the lineages (P<0.05), and it was predominantly resistant **(Figure 2)**.

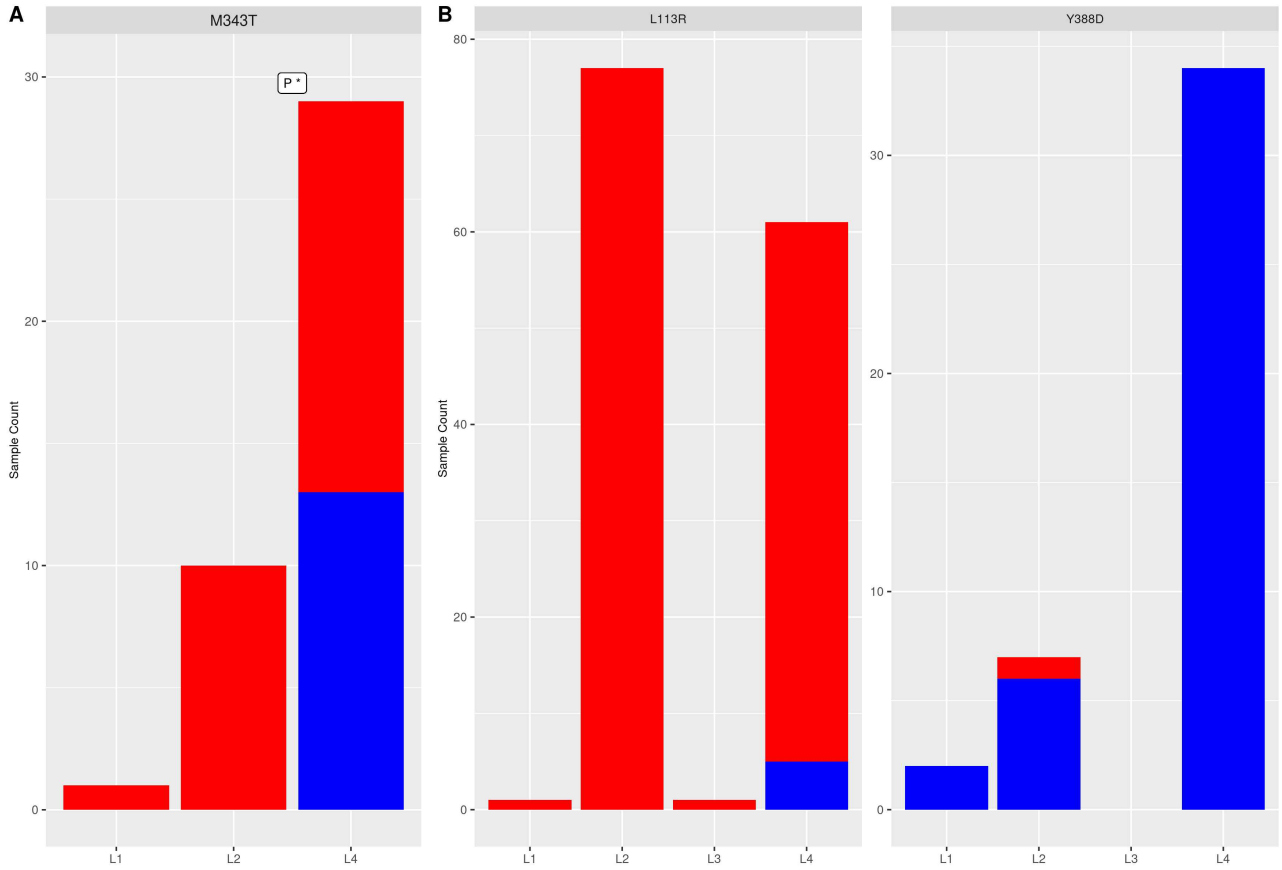


Figure 2: Mutations in *alr* with differing sensitivities
Mutations in *M. tuberculosis* gene *alr*, with different sensitivities across lineages 1-4. Fisherss exact test was used to compare mutation proportions across lineages, and statistical significance was assessed according to the thresholds: .P<0.10, *P<0.05, **P<0.01, ***P<0.001, ****P<0.0001 **A)** Statistically significant mutations, **B)** Mutations that did not meet statistical significance.

Mutations with differing drug sensitivity in *embB*: For *embB*, only 9% (67/744) of mutations were found to be different with respect to EMB sensitivity across the four lineages **(Figure 3A** and **3B)** which included EMB binding residues M306 and Y334, but others were mainly residues not involved with the active site. There were 11 mutations that were statistically significant ($P<0.05$) with respect to differing EMB sensitivity across the lineages, and these were predominantly resistant **(Figure 3A)**. Irrespective of the statistical significance, and despite very low numbers, it was interesting to note that mutations Y319D, A962P, D354N, S651N, and A201S were represented approximately equally in their differing drug sensitivity to EMB. This highlights that mutational effects differ depending on the genetic background of the bacterial stain **(Figure 3C)**.

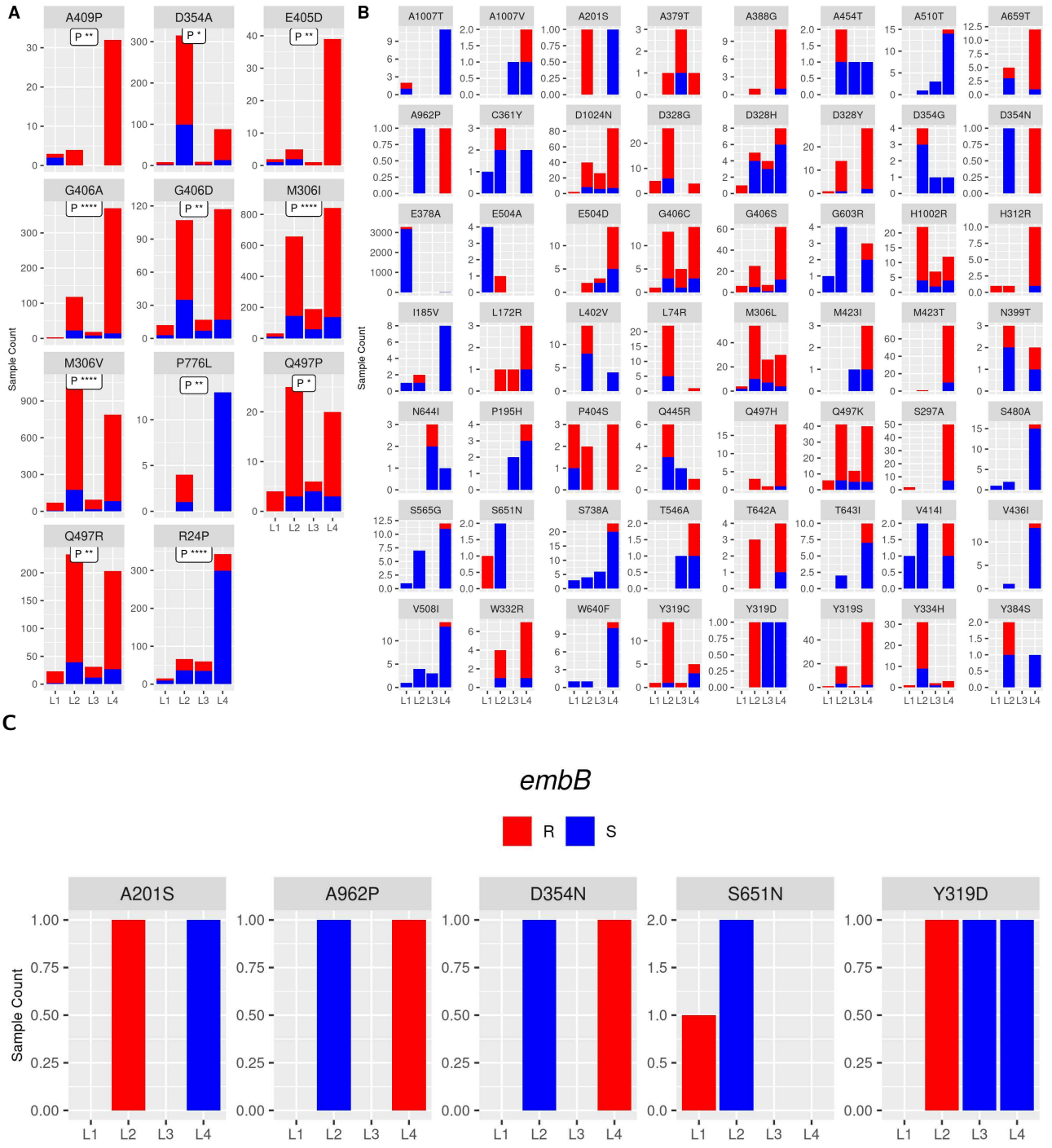


Figure 3: Mutations in *embB* with differing sensitivities
Mutations in *M. tuberculosis* gene *embB*, with different sensitivities across lineages 1-4. Fisherss exact test was used to compare mutation proportions across lineages, and statistical significance was assessed according to the thresholds: .P<0.10, *P<0.05, **P<0.01, ***P<0.001, ****P<0.0001 **A)** Statistically significant mutations, **B)** Mutations that did not meet statistical significance, **C)** Selected mutations showing prominent differing sensitivities across lineages.

Mutations with differing drug sensitivity in *gidB*: Nearly 18% (94/520) of mutations were found to be different with respect to STR sensitivity across the four lineages **(Figures 4A** and **4B)**. Though some of these were gidB interacting residues, a majority (n=81) did not interact with any of the gidB binding partners. There were 5 mutations that reached statistical significance ($P<0.05$) with respect to differing STR sensitivity across the lineages, none of which were involved with gidB binding partners **(Figure 4A)**, though mutation P75R ($P<0.01$) showed comparable numbers of resistant and sensitive samples **(Figure 4A)**. Further, irrespective of the statistical significance, and despite very low numbers, it was interesting to note that mutations A133P, A19G, R118L, and R154W were represented equally in their differing drug sensitivity to STR, highlighting the need to consider lineage effects on mutations associated with resistance **(Figure 4C)**.

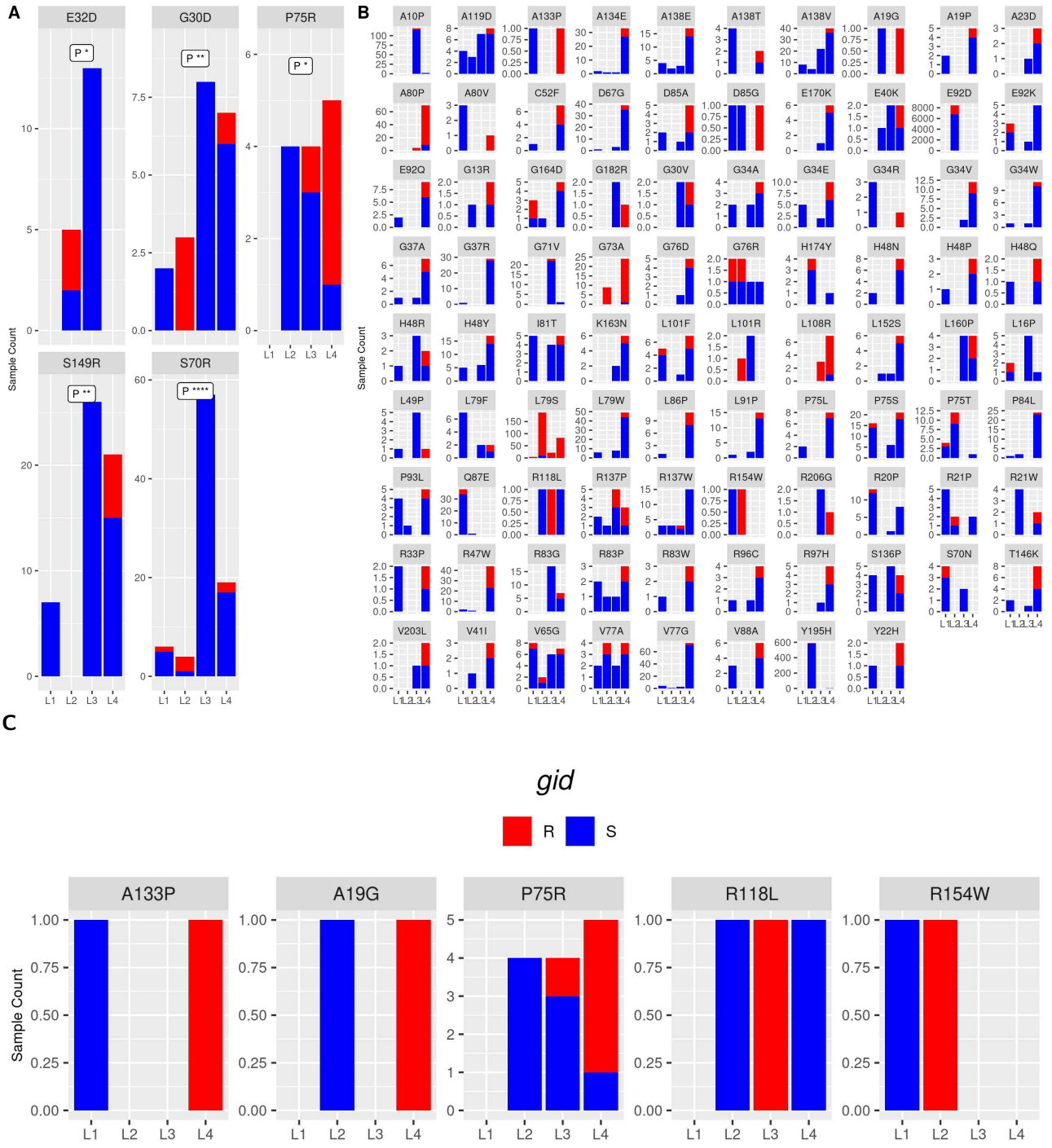


Figure 4: Mutations in *gidB* with differing sensitivities
Mutations in *M. tuberculosis* gene *gidB*, with different sensitivities across lineages 1-4. Fisherss exact test was used to compare mutation proportions across lineages, and statistical significance was assessed according to the thresholds: .P<0.10, *P<0.05, **P<0.01, ***P<0.001, ****P<0.0001 **A)** Statistically significant mutations, **B)** Mutations that did not meet statistical significance, **C)** Selected mutations showing prominent differing sensitivities across lineages.

Mutations with differing drug sensitivity in *katG*: Approximately 9% (68/772) of mutations were found to be different with respect to INH sensitivity across the four lineages **(Figure 5A** and **5B)** which included only two active site residues, with a majority being residues extending beyond the active site. Among these, only 3 mutations (D714E, M257I, R463L) reached statistical significance ($P<0.05$) with respect to differing INH sensitivity across the lineages, none of which were active site residues **(Figure 5A)** and all three mutations were predominantly sensitive across the lineages. Despite very low numbers, and irrespective of the statistical significance, it was interesting to note that mutations G490S, H116F, H97N, M257V, R249C, S383A, V320L, and W300R were represented equally in their differing drug sensitivity to INH, underscoring the importance of accounting for strain diversity in understanding mutational effects linked to resistance development **(Figure 5C)**. In a recent (2023) analysis by Napier, *et. al.*,[22] a list of co-occurring mutations in *katG* was published to help inform genotypic drug-resistance profiling. This was based on the rationale that compensatory mutational effects of the putative resistance markers will help accurately estimate INH resistance. Compensatory effects are known to play a role in mitigating fitness costs associated with resistant mutations, allowing mutations to become fixed in a population. This can result from concomitant mutational effects in the same or multiple genes like *katG*, *inhA*, *ahpC*,[22,23] and positive epistasis of low-cost resistance-conferring mutations influenced by *M. tuberculosis* strain diversity.[24] To the best of our knowledge, the mutations identified in this analysis have not been reported on previously. These mutations were also not identified in the analysis conducted by Napier, *et. al.*,[22] which is based on a subset of the data used in this thesis. It would appear that since strain diversity was only considered indirectly in the 2023 Napier, *et. al.* study, with mutational analysis being undertaken at the genome level with multiple genes, rather than investigating individual SAVs in a given gene across the lineages, these mutations were not observed in their study. The said mutations also did not appear to cluster on the protein, and did not involve any INH binding residues. A systematic approach starting with extracting mutations from the same isolate, and then proceeding to investigate lineage effects, may offer better insights accounting for the combination of compensatory mutations and strain diversity.[12,20,25]

It was noted that the R463L mutation, a known lineage-specific mutation, was found in all non-lineage 4 isolates. As this mutation is present in all lineage 1, 2 and 3 isolates the statistical test is effectively comparing the rates of isoniazid resistance between the three lineages. Many of the lineage 2 samples are sourced from countries where resistance levels are high and, as such, the Fisher's exact test picks up a significant association **(Figure 5A)**.

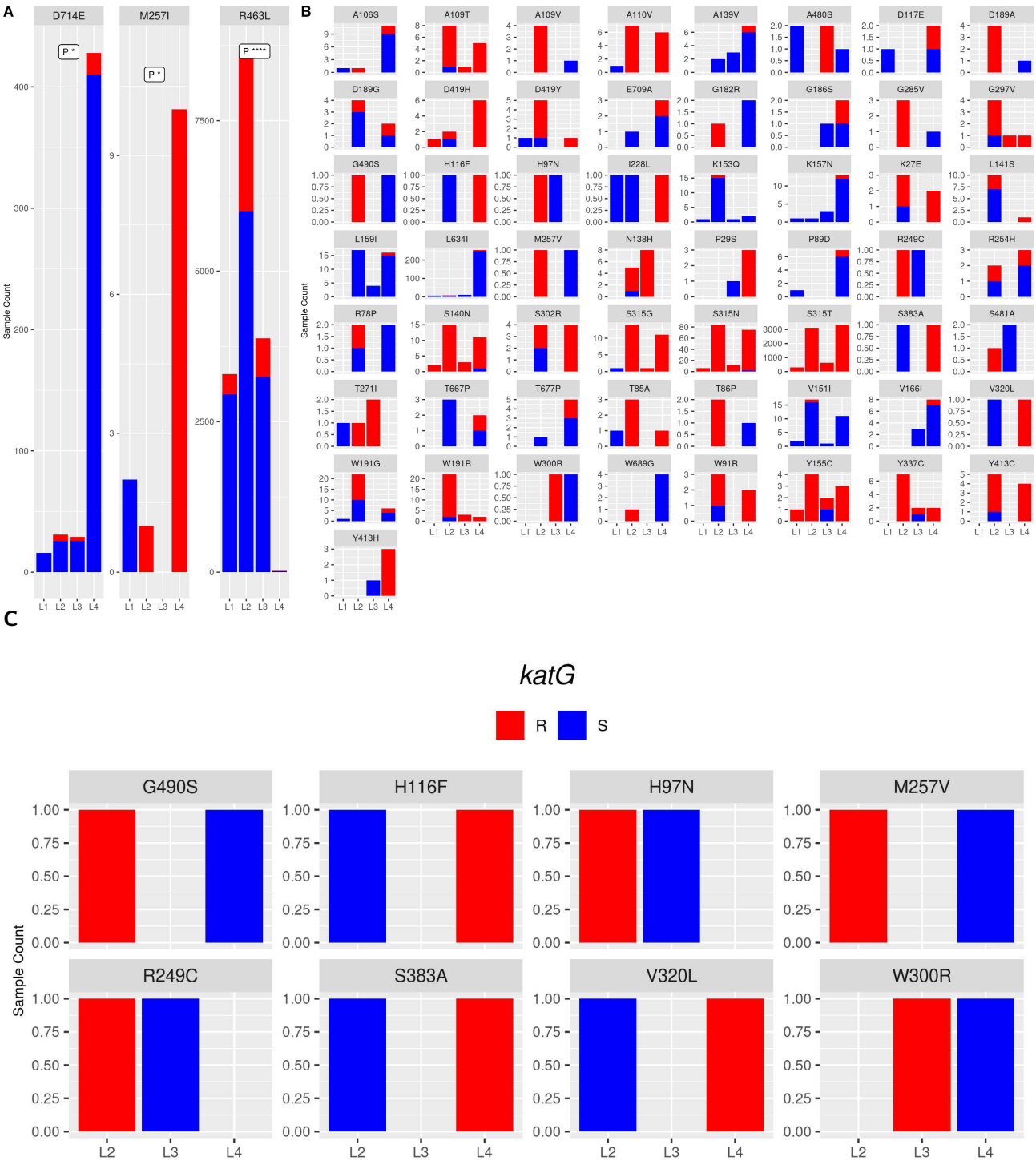


Figure 5: Mutations in *katG* with differing sensitivities
Mutations in *M. tuberculosis* gene *katG*, with different sensitivities across lineages 1-4. Fisherss exact test was used to compare mutation proportions across lineages, and statistical significance was assessed according to the thresholds: .P<0.10, *P<0.05, **P<0.01, ***P<0.001, ****P<0.0001 **A)** Statistically significant mutations, **B)** Mutations that did not meet statistical significance, **C)** Selected mutations showing prominent differing sensitivities across lineages.

Mutations with differing drug sensitivity in *rpoB*: For *rpoB*, 13% (138/1047) of mutations were found to be different with respect to RFP sensitivity across the four lineages **(Figure 6A** and **6B)**, which included residues in and beyond the active site. Among these, 9 mutations reached statistical significance (P<0.05) with respect to differing RFP sensitivity across the lineages, none of which were active site residues **(Figure 6A)**, with mutations D435Y, H445N, H445R, H445S, and S450L being predominantly resistant and mutations D777N, H749Y, and K799Q being predominantly sensitive. The mutation I491F showed comparable resistant and sensitive sample numbers **(Figure 6A)**. Further, despite very low numbers and irrespective of the statistical significance, it was interesting to note that mutations E807Q, H674Q, I696L, P439S, Q287E, S188A, S195R, S239R, S431T (active site residue), T756A, V243I, V385L, V403A, and V517K represented equally in their differing drug sensitivity, bringing to the fore the influence of the genetic background in shaping mutational effects on resistance development **(Figure 6C)**.

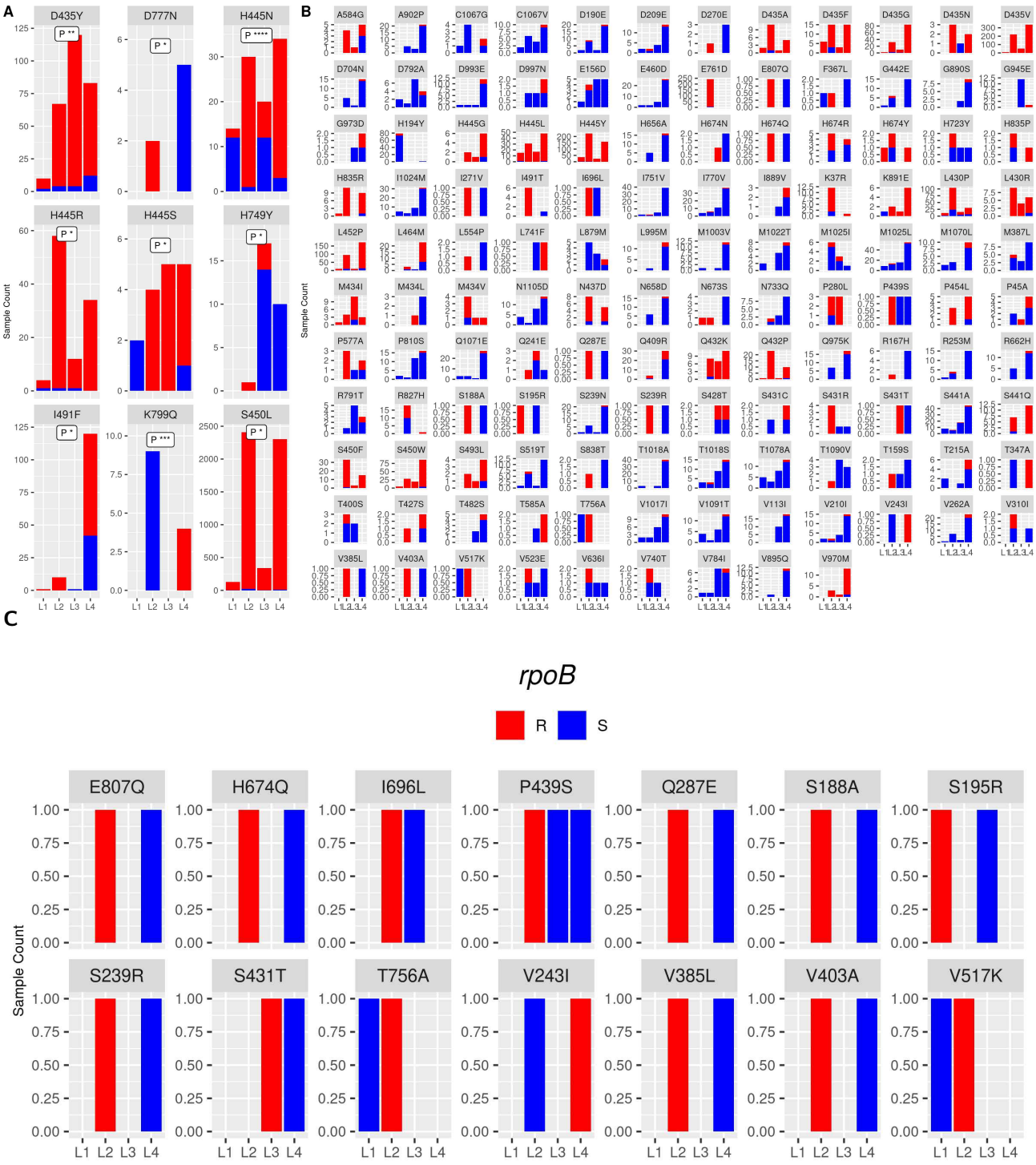


Figure 6: Mutations in *rpoB* with differing sensitivities
Mutations in *M. tuberculosis* gene*rpoB*, with different sensitivities across lineages 1-4. Fisherss exact test was used to compare mutation proportions across lineages, and statistical significance was assessed according to the thresholds: .P<0.10, *P<0.05, **P<0.01, ***P<0.001, ****P<0.0001 **A)** Statistically significant mutations, **B)** Mutations that did not meet statistical significance, **C)** Selected mutations showing prominent differing sensitivities across lineages.

10.3.1 Summary of sensitivity effects across lineages

As an example, for rifampicin resistance, many of the mutations classified as statistically significant are located in codon 445. Mutations in this codon have previously been classed as disputed resistance mutations due to their ability to produce low-level resistance. A study in 2018 by Miotto, *et. al.*[26] found that these mutations are associated with a delayed growth on Mycobacteria Growth Indicator Tube (MGIT). They also considered the structural effects of disputed vs. undisputed mutations, but did not find any difference, concluding that all mutations affect drug binding. A point that was not considered was the lineage of the strains, which showed a differential resistance threshold. These results show that lineage could potentially be a factor and should be taken into account in future studies.

These findings offer insight into the potential existence of lineage effects on mutational sensitivity. To the best of our knowledge, such an analysis has not yet been performed, with data shown in its raw form with respect to drug sensitivity across lineages. With 13% (40/495) of mutations, over six gene-drug targets displaying differing sensitivities, these findings support the occurrence and role of epistasis in *M. tuberculosis*, where the sequence background in which mutations develop affects the resulting phenotype. Epistatic interactions involving compensatory mutations are recognised to influence mutational evolutionary trajectory towards resistance development,[16,18,24,27–29] whereas some lineages have been shown to display intrinsic resistance to specific drugs due to mutations that have become fixed in the lineage populations.[30] Despite these interactions, efforts to investigate resistance development systematically in an epistasis informed manner have been limited.

A consensus and concerted effort into pursuing this route of investigation through genetic and phenotypic testing, as well as computational approaches including mathematical modelling will allow us to apply a systems biology approach towards predicting resistance more accurately and in a targeted manner. Accumulation of mutations considered in a step-wise manner opens up further avenues of understanding resistance development in terms of protein evolution, and will allow development of new phylogenetic models to leverage epistatic interactions.[31]

The implication for such considerations may have a direct impact on therapy and treatment management,[8,32] especially with personalised medicine for TB patients.[33,34] For example, since the *M. tuberculosis* isolates in this project represent clinical samples, the implication for lineage specific effects on mutation sensitivity can be of great benefit in tailoring therapy for personalised medicine.

A further point is that these differential resistance effects were stratified by lineage, while there is a

possibility that epistatic effects could be seen at the sub-lineage level. For example, lineage 4 is a diverse lineage that contains many subclades (labelled as 4.1, 4.2 etc.). The statistical significance of epistatic effects present due to mutations that are fixed in only one of these subclades would be dampened by only analysing the main lineage. Future work could look into these relationships using sub-lineage stratification.

Despite the low numbers, these insights warrant further investigation and future work to consider strain diversity in understanding the evolution of drug resistance.[12,35] In the context of relating the mutational impact on biophysical measures (e.g. stability, drug binding affinity) these insights may call for the use of lineage specific protein structure models. The pursuit of such an approach is necessitated in the interest of informing therapy and improving patient compliance (personalised medicine). Laboratory assessment of the different drugs for DST can be further differentiated by lineage to help determine drug sensitivity and resistance profile in *M. tuberculosis* strains to inform clinical management.

There is also potential in the use of such approaches to predict the future trajectories of DR-TB (MDR/XDR-TB). The widespread use of sequencing technology can rapidly help identify co-occurring mutations in clinical isolates. Computational and statistical methods exploiting genomics, protein structure, and lineage data can be used for assessing mutational impact of concomitant mutations and compensatory mutations, including their association with drug resistance. Utilising the learnings from these systematic assessments, mathematical models attempting to quantify information related to the different measures of resistance can be developed. Together with the help of lineage specific mutations and putative resistance markers identified from such analyses, existing or new computational models can be exploited to predict resistance for multiple mutations in a given sequence background. In this manner, potential resistance routes (combination of mutations) in light of the phylogenetic background (strain) of *M. tuberculosis* can be identified. This will help predict future trajectories of resistance mutations in single or multiple genes, and ultimately help to mitigate the emergence of drug resistance in TB.

10.4 Results dashboard

An interactive dashboard, available at https://thesis.tunstall.in, was built as part of this thesis to explore each of the six gene-drug target pairs. The dashboard has three parts: 'Gene-Drug Target Explorer' is the main dashboard and provides several views of the data for my thesis.

The 'MSA' explorer was constructed to explore a high-level overview of sequence alignment and

provides an easy way to see which positions experience high mutational frequency, and uses the enrichment depletion (ED) score to visualise MSA information.[36] The ED score is an improvement over standard Logo plots as it includes empirical Bayes methods to stabilise estimates of enrichment and depletion, and to highlight the most significant patterns in data. Screenshots are in appendices 10.A.1 and 10.B.1.

References

[1] Rajesh Sarkar et al. "Modern Lineages of Mycobacterium Tuberculosis Exhibit Lineage-Specific Patterns of Growth and Cytokine Induction in Human Monocyte-Derived Macrophages". In: *PLOS ONE* 7.8 (Aug. 16, 2012), e43170. ISSN: 1932-6203. DOI: 10.1371/journal.pone.0043170.

[2] Norbert Reiling et al. "Clade-Specific Virulence Patterns of Mycobacterium Tuberculosis Complex Strains in Human Primary Macrophages and Aerogenically Infected Mice". In: *mBio* 4.4 (July 30, 2013), e00250–13. ISSN: 2150-7511. DOI: 10.1128/mBio.00250-13.

[3] Mireia Coscolla and Sebastien Gagneux. "Consequences of Genomic Diversity in Mycobacterium Tuberculosis". In: *Seminars in Immunology* 26.6 (Dec. 2014), pp. 431–444. ISSN: 1096-3618. DOI: 10.1016/j.smim.2014.09.012.

[4] Eddie M. Wampande et al. "Genetic Variability and Consequence of Mycobacterium Tuberculosis Lineage 3 in Kampala-Uganda". In: *PLOS ONE* 14.9 (Sept. 9, 2019), e0221644. ISSN: 1932-6203. DOI: 10.1371/journal.pone.0221644.

[5] Jitendra Singh et al. "Genetic Diversity and Drug Susceptibility Profile of Mycobacterium Tuberculosis Isolated from Different Regions of India". In: *Journal of Infection* 71.2 (Aug. 2015), pp. 207–219. ISSN: 01634453. DOI: 10.1016/j.jinf.2015.04.028.

[6] Yaa E.A. Oppong et al. "Genome-Wide Analysis of Mycobacterium Tuberculosis Polymorphisms Reveals Lineage-Specific Associations with Drug Resistance". In: *BMC Genomics* (2019). ISSN: 14712164. DOI: 10.1186/s12864-019-5615-3.

[7] Siva Kumar Shanmugam et al. "Mycobacterium Tuberculosis Lineages Associated with Mutations and Drug Resistance in Isolates from India". In: *Microbiology Spectrum* 10.3 (June 29, 2022), e0159421. ISSN: 2165-0497. DOI: 10.1128/spectrum.01594-21.

[8] Sònia Borrell and Sebastien Gagneux. "Strain Diversity, Epistasis and the Evolution of Drug Resistance in Mycobacterium Tuberculosis". In: *Clinical microbiology and infection : the official publication of the European Society of Clinical Microbiology and Infectious Diseases* 17.6 (June 2011), pp. 815–820. ISSN: 1198-743X. DOI: 10.1111/j.1469-0691.2011.03556.x.

[9] Borna Müller et al. "The Heterogeneous Evolution of Multidrug-Resistant Mycobacterium Tuberculosis". In: *Trends in genetics: TIG* 29.3 (Mar. 2013), pp. 160–169. ISSN: 0168-9525. DOI: 10.1016/j.tig.2012.11.005.

[10] Fenner L et al. "Effect of Mutation and Genetic Background on Drug Resistance in Mycobacterium Tuberculosis." In: *Antimicrobial Agents and Chemotherapy* 56.6 (Apr. 2, 2012), pp. 3047–3053. ISSN: 0066-4804, 1098-6596. DOI: 10.1128/aac.06460-11.

[11] Roger Vargas et al. "Role of Epistasis in Amikacin, Kanamycin, Bedaquiline, and Clofazimine Resistance in Mycobacterium Tuberculosis Complex". In: *Antimicrobial Agents and Chemotherapy* 65.11 (Oct. 18, 2021), e01164–21. ISSN: 0066-4804. DOI: 10.1128/AAC.01164-21.

[12] Alex Wong. "Epistasis and the Evolution of Antimicrobial Resistance". In: *Frontiers in Microbiology* 8 (2017), p. 246. ISSN: 1664-302X. DOI: 10.3389/fmicb.2017.00246.

[13] Patrick C. Phillips. "Epistasis the Essential Role of Gene Interactions in the Structure and Evolution of Genetic Systems". In: *Nature Reviews Genetics* 9.11 (11 Nov. 2008), pp. 855–867. ISSN: 1471-0064. DOI: 10.1038/nrg2452.

[14] Quang Huy Nguyen et al. “Insights into the Processes That Drive the Evolution of Drug Resistance in Mycobacterium Tuberculosis”. In: *Evolutionary Applications* 11.9 (2018), pp. 1498–1511. ISSN: 1752-4571. DOI: 10.1111/eva.12654.

[15] Mashael Al-Saeedi and Sahal Al-Hajoj. “Diversity and Evolution of Drug Resistance Mechanisms in Mycobacterium Tuberculosis”. In: *Infection and Drug Resistance* Volume 10 (Oct. 2017), pp. 333–342. ISSN: 1178-6973. DOI: 10.2147/IDR.S144446.

[16] Iñaki Comas et al. “Whole-Genome Sequencing of Rifampicin-Resistant Mycobacterium Tuberculosis Strains Identifies Compensatory Mutations in RNA Polymerase Genes”. In: *Nature Genetics* 44.1 (2012), pp. 106–110. ISSN: 10614036. DOI: 10.1038/ng.1038.

[17] Diarmaid Hughes and Gerrit Brandis. “Rifampicin Resistance: Fitness Costs and the Significance of Compensatory Evolution”. In: *Antibiotics* 2.2 (2 June 2013), pp. 206–216. ISSN: 2079-6382. DOI: 10.3390/antibiotics2020206.

[18] M. de Vos et al. “Putative Compensatory Mutations in the rpoC Gene of Rifampin-Resistant Mycobacterium Tuberculosis Are Associated with Ongoing Transmission”. In: *Antimicrobial Agents and Chemotherapy* 57.2 (Feb. 2013), pp. 827–832. ISSN: 0066-4804. DOI: 10.1128/AAC.01541-12.

[19] Claudio U. Köser et al. “Consequences of whiB7 (Rv3197A) Mutations in Beijing Genotype Isolates of the Mycobacterium Tuberculosis Complex”. In: *Antimicrobial Agents and Chemotherapy* 57.7 (July 2013), p. 3461. ISSN: 1098-6596. DOI: 10.1128/AAC.00626-13.

[20] Matthias Merker et al. “Phylogenetically Informative Mutations in Genes Implicated in Antibiotic Resistance in Mycobacterium Tuberculosis Complex”. In: *Genome Medicine* 12.1 (Mar. 6, 2020), p. 27. ISSN: 1756-994X. DOI: 10.1186/s13073-020-00726-5.

[21] Julian S. Peters et al. “Genetic Diversity in *Mycobacterium Tuberculosis* Clinical Isolates and Resulting Outcomes of Tuberculosis Infection and Disease”. In: *Annual Review of Genetics* 54.1 (Nov. 23, 2020), pp. 511–537. ISSN: 0066-4197, 1545-2948. DOI: 10.1146/annurev-genet-022820-085940.

[22] Gary Napier et al. “Large-Scale Genomic Analysis of Mycobacterium Tuberculosis Reveals Extent of Target and Compensatory Mutations Linked to Multi-Drug Resistant Tuberculosis”. In: *Scientific Reports* 13.1 (Jan. 12, 2023), p. 623. ISSN: 2045-2322. DOI: 10.1038/s41598-023-27516-4.

[23] Pauline Lempens et al. “Isoniazid Resistance Levels of Mycobacterium Tuberculosis Can Largely Be Predicted by High-Confidence Resistance-Conferring Mutations”. In: *Scientific Reports* 8.1 (Feb. 19, 2018), p. 3246. ISSN: 2045-2322. DOI: 10.1038/s41598-018-21378-x.

[24] Qin-Jing Li et al. “Positive Epistasis of Major Low-Cost Drug Resistance Mutations rpoB531-TTG and katG315-ACC Depends on the Phylogenetic Background of Mycobacterium Tuberculosis Strains”. In: *International journal of antimicrobial agents* 49.6 (June 2017), pp. 757–762. ISSN: 0924-8579. DOI: 10.1016/j.ijantimicag.2017.02.009.

[25] Keira A. Cohen et al. “Evolution of Extensively Drug-Resistant Tuberculosis over Four Decades: Whole Genome Sequencing and Dating Analysis of Mycobacterium tuberculosis Isolates from KwaZulu-Natal”. In: *PLoS medicine* 12.9 (Sept. 2015), e1001880. ISSN: 1549-1277. DOI: 10.1371/journal.pmed.1001880.

[26] Paolo Miotto et al. “Role of Disputed Mutations in the rpoB Gene in Interpretation of Automated Liquid MGIT Culture Results for Rifampin Susceptibility Testing of Mycobacterium Tuberculosis”. In: *Journal of Clinical Microbiology* 56.5 (May 2018), e01599–17. ISSN: 1098-660X. DOI: 10.1128/JCM.01599-17.

[27] Amel Kevin Alame Emane et al. “Drug Resistance, Fitness and Compensatory Mutations in Mycobacterium Tuberculosis”. In: *Tuberculosis* 129 (July 1, 2021), p. 102091. ISSN: 1472-9792. DOI: 10.1016/j.tube.2021.102091.

[28] Qingyun Liu et al. “Have Compensatory Mutations Facilitated the Current Epidemic of Multidrug-Resistant Tuberculosis?” In: *Emerging Microbes & Infections* 7.1 (2018), pp. 1–8. DOI: 10.1038/s41426-018-0101-6.

[29] Shengfen Wang et al. “Characteristics of Compensatory Mutations in the rpoC Gene and Their Association with Compensated Transmission of Mycobacterium Tuberculosis”. In: *Frontiers of Medicine* 14.1 (Feb. 1, 2020), pp. 51–59. ISSN: 2095-0225. DOI: 10.1007/s11684-019-0720-x.

[30] Bouke C. de Jong et al. “Does Resistance to Pyrazinamide Accurately Indicate the Presence of Mycobacterium Bovis?” In: *Journal of Clinical Microbiology* 43.7 (July 2005), pp. 3530–3532. ISSN: 0095-1137. DOI: 10.1128/JCM.43.7.3530-3532.2005.

[31] Premal Shah, David M. McCandlish, and Joshua B. Plotkin. “Contingency and Entrenchment in Protein Evolution under Purifying Selection”. In: *Proceedings of the National Academy of Sciences* 112.25 (June 23, 2015), E3226–E3235. DOI: 10.1073/pnas.1412933112.

[32] Mireilla Coscolla and Sebastien Gagneux. “Does M. Tuberculosis Genomic Diversity Explain Disease Diversity?” In: *Drug discovery today. Disease mechanisms* 7.1 (2010), e43–e59. ISSN: 1740-6765. DOI: 10.1016/j.ddmec.2010.09.004.

[33] Kartik Kumar and Onn Min Kon. “Personalised Medicine for Tuberculosis and Non-Tuberculous Mycobacterial Pulmonary Disease”. In: *Microorganisms* 9.11 (Oct. 26, 2021), p. 2220. ISSN: 2076-2607. DOI: 10.3390/microorganisms9112220.

[34] Ioana D. Olaru, Christoph Lange, and Jan Heyckendorf. “Personalized Medicine for Patients with MDR-TB”. In: *Journal of Antimicrobial Chemotherapy* 71.4 (Apr. 1, 2016), pp. 852–855. ISSN: 0305-7453. DOI: 10.1093/jac/dkv354.

[35] Matthias Merker et al. “Compensatory Evolution Drives Multidrug-Resistant Tuberculosis in Central Asia”. In: *eLife* 7 (Oct. 2018). Ed. by Gisela Storz, e38200. ISSN: 2050-084X. DOI: 10.7554/eLife.38200.

[36] Kushal K. Dey, Dongyue Xie, and Matthew Stephens. “A New Sequence Logo Plot to Highlight Enrichment and Depletion”. In: *BMC Bioinformatics* 19.1 (Dec. 10, 2018), p. 473. ISSN: 1471-2105. DOI: 10.1186/s12859-018-2489-3.

Appendix for Chapter 10

10.A Gene-drug targets dashboard

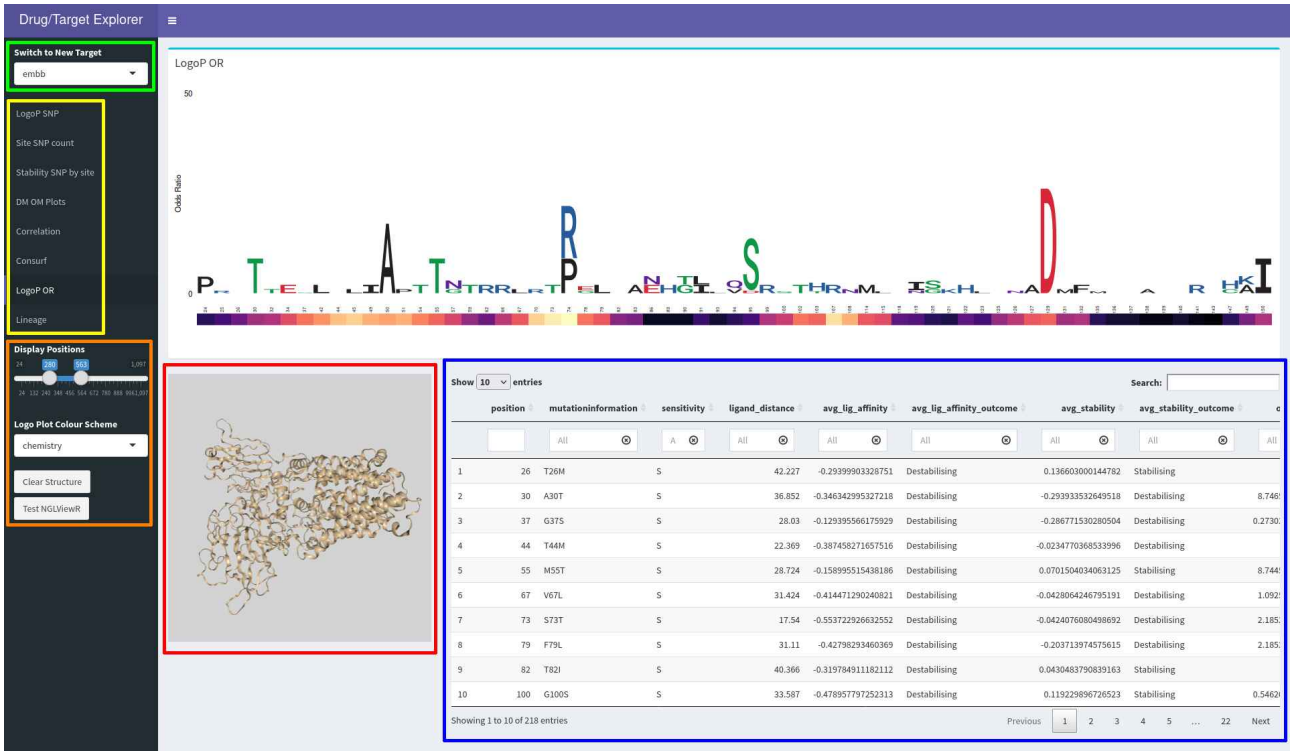


Figure 10.A.1: Gene-drug targets web interface
The web interface for gene-drug targets allows the user to switch between the multiple drug/gene combinations (highlighted in green), view multiple plot types as used in this thesis (highlighted in yellow) and dynamically adjust various parameters for the current plot (highlighted in orange). A 3D view of the current target molecule is also visible (highlighted in red). Mutation positions can be added to the 3D view by searching for and highlighting them by clicking on the entries in the data table (highlighted in blue).

10.B Multiple Sequence Alignment dashboard

Figure 10.B.1: Multiple Sequence Alignment interface
The web interface for Multiple Sequence Alignment allows the user to dynamically explore a Logo Plot for each of the gene-drug targets (highlighted in green). The bottom part of the plot indicates the wild-type residues by position. The upper part indicates the Enrichment Depletion score.[36] Due to the number of gene positions, it is necessary to adjust the range of visible positions using the slider (highlighted in magenta). As per the figures in other chapters, the ligand distance and presence of drug and other interacting partners (where applicable) are reflected beneath each position (highlighted in blue). The colour scheme can also be changed to reflect different residue properties: Chemistry, hydrophobicity, "clustalx" and "taylor" (highlighted in cyan). The key (highlighted in red) changes according to the colour scheme selected.

Chapter 11

Machine learning results

11.1 Machine Learning

11.1.1 Methods

The scikit-learn package version 1.1.1[1] was used to perform all Machine Learning (ML)-based tasks. Estimates related to SAVs from the computational tools (biophysical measures) and calculated measures (genomics and residue level properties) mentioned above together with numerical measures from AAindex[2–6] formed the set of comprehensive features used for all ML tasks. AAindex calculations return numerical values corresponding to various physicochemical and biochemical properties of amino acids and pairs of amino acids from a database curated from published literature. AAindex consists of three sections: AAindex1 for the amino acid index of 20 numerical values, AAindex2 for the amino acid mutation matrix, and AAindex3 for the statistical protein contact potentials. All three indices were used in our analyses.[2–6] AAindex was run using a combination of Bash and Python scripts on all six gene-drug targets.

A supervised learning approach to classify mutational effect as 'Sensitive' or 'Resistant' was employed. A total of 23 representative classification models (classifiers) were used. These employ distinct strategies: Linear Classifiers: Gaussian Process, Linear Discriminant Analysis (LDA), Logistic Regression, Logistic Regression CV, Passive Aggressive, Ridge Classifier, Ridge Classifier CV, Stochastic Gradient Descent (SGD), Support Vector Classifier (SVC); Tree based classifiers: Decision Tree, Extra Trees, Extra Tree; Bagging based classifiers: Bagging Classifier, Random Forest; Boosting based classifiers: Adaptive Boosting Classifier (AdaBoost), Gradient Boosting Classifier, Extreme Gradient Boosting (XGBoost); Naive Bayes (NB) classifiers: Gaussian NB, Multinomial NB, Complement NB; Other classifiers: Quadratic Discriminant Analysis (QDA), K-Nearest Neighbours (KNN), and the artificial neural network based Multi Layer Perceptron (MLP). A summary description of the models used is provided in Appendix Table 11.A.1. The outcome of the aggregate DST for each mutation (designated as 0: Sensitive, 1: Resistant) was used as the 'target' variable, which is the outcome feature a given model attempts to predict.

Results obtained from all *in silico* predictors were used as input features. Oversampling methods (random oversampling,[7] random undersampling,[8] combined random over-and-undersampling,[7,8] and synthetic minority oversampling technique (SMOTE)[9]) were used to address the imbalanced distribution of these mutational classes. SMOTE uses the nearest-neighbour technique to generate oversampled data. Both 70/30 and 80/20 train-test split thresholds were applied, along with the train-test split according to the scaling law method.[10] In the scaling method, the size of the test set is inversely pro-

portional to the square root of the number of input features (excluding the target). All input features were scaled as part of the data pre-processing pipeline before running ML models. A stratified K-fold cross validation (CV) strategy was used. CV is a common technique used in ML to mitigate over-fitting and issues related to a single train-test split. In cross validation, training data is divided into n subsets, where the n-1 set is used for training and the remaining set is used for testing. This procedure is repeated n times to ensure each subset is used for both training and testing. Stratification with CV ensures that the distribution of the target variable is preserved during training. All models were run using a 10-fold stratified cross validation (CV) on the training data, and validated on the test data. Models were optimised using the Matthews Correlation Coefficient (MCC) metric. MCC is a single value that summarises a confusion matrix (summary table for ML classification predictions), and as such is currently the most balanced metric in use when evaluating model performance. Evaluation metrics for ML models are summarised below. Since there was no independent blind test data in our project, model performance was assessed using the validation set, and different CV folds of 3, 5, and 10. Initial ML models were built for each gene-drug target, followed by a combined approach to build a gene-agnostic ML model to classify mutations. The latter was done with the intention of identifying the underlying common molecular mechanisms in resistance development. As a first attempt to generate a combined model predicting resistance, a test MCC score of at least 0.4, along with an absolute difference of $\leq$0.1 between train and test MCC, would indicate the potential of pursuing such a gene-agnostic approach in predicting resistance. The latter criteria was also applied to individual gene-target model predictions.

11.1.1.1 Feature selection

Second only to data cleaning, feature selection is one of the most important and time consuming parts of the machine learning process. The two feature selection methods used in the ML workflow were the Recursive Feature Elimination CV (RFECV) and the Boruta feature selection method.[11] As RFECV is a greedy algorithm which iteratively add or drops features, it can become computationally expensive when used with large datasets. It is therefore usually preceded by a pre feature selection step, where a knowledge-based approach is used to remove any redundant features, or statistical tests are carried out to include only significant features for use in RFECV. On the other hand, Boruta is an algorithm designed to take the "all-relevant" approach to feature selection, where it attempts to find all features that contribute to model performance, rather than finding the minimal subset of features that are important in a model. Boruta was originally written as an R package,[11] and has subsequently been been ported to Python and made available as part of the scikit-learn (BorutaPy[12]) with some

additional improvements to allow additional user control. It is an efficient algorithm with faster run times, and offers a neat and easy way to interpret the model output. It displays the number of features selected and rejected at each iteration, and provides feature rankings for all features used as the process. Boruta works by creating copies of the original features and randomly shuffles them to create a number of randomly shuffled shadow attributes, while removing correlated features with the target variable. This adds a further sanity check to the ML process by ensuring that no highly correlated features dominate the model performance. This helps establish a baseline performance for the model. This is followed by testing the hypothesis at a default significance threshold (0.05) to determine the statistical significance (i.e. importance) of the feature with respect to the target. This is achieved using the random forest classifier. This iterative process is then able to accept and reject features according to the significance threshold. As Boruta iteratively removes uninformative variables, the noise introduced by other features is removed leading to improvements in model performance.

11.1.1.2 Evaluation metrics

Performance for classification models in ML is evaluated using the common metrics listed below:

1. Confusion Matrix: Tabular visualisation of actual values (ground truth) versus model predictions. Rows and columns represent instances of actual and predicted classes. Though not an evaluation metric as such, the confusion matrix is the basis from which all other metrics are derived.

Confusion Matrix		**Predicted**	
		False	True
Actual	False	True Negative (TN)	False Positive (FP)
	True	False Negative (FN)	True Positive (TP)

TN: Observation is negative, and is predicted negative.

FP: Observation is negative, and is predicted positive.

FN: Observation is positive, and is predicted negative.

TP: Observation is positive, and is predicted positive.

2. Accuracy: Ratio of number of correct predictions to the total number predictions

$$Accuracy = \frac{(TP + TN)}{(TP + TN + FP + FN)}$$

3. Precision: Ratio of true positives to total predicted positives. Also known as the positive

predictive value.

$$Precision = \frac{TP}{(TP + FP)}$$

4. Recall: Ratio of predicted positives to actual number of positives. Also known as Sensitivity or the true positive rate.

$$Recall = \frac{TP}{(TP + FN)}$$

5. F1-score: The F1-score or the F-score is the harmonic mean of two metrics: Precision and Recall.

$$F1 = \frac{2}{(Recall^{-}1) + (Precision^{-}1)}$$

6. Matthews Correlation Coefficient (MCC): MCC accounts for true and false positives and negatives, and is regarded as a balanced measure irrespective of class size.

$$MCC = \frac{(TP \times TN)(FP \times FN)}{\sqrt{(TP + FP)(TP + FN)(TN + FP)(TN + FN)}}$$

7. Jaccard Index or Jaccard score (JCC): A similarity metric used to compare the set of predicted labels for a given class to the corresponding complete set of labels. It may be a poor metric if there are no positives for some samples or classes.

$$Jaccard(A, B) = \frac{A \cap B}{A \cup B}$$

A and B denote data classes

11.1.2 Results

Each of the six gene targets was tested individually using several ML models to predict resistance for the corresponding drug, before proceeding to combine data across the genes to predict resistance in a gene-agnostic manner. A panel of ML models was trained and tested to predict resistance for each gene-drug target individually, using several train-test split strategies in the absence of independent blind test sets. The same models were used with the combined data across all genes to predict resistance for a given drug used as the 'test' set for gene agnostic predictions.

Table 1 summarises the data collected and the number of mutations in the sensitive and resistant groups as per aggregate DST. More than 150 features were used to train several ML models (see Appendix Table 11.B.1), followed by a feature selection step (See Methods section above for details), to identify which features were making the largest contribution towards the model performance in predicting resistance. The MCC scoring function was chosen to evaluate model performance as it is the most balanced metric accounting for all categories of prediction in its score (True Positive, True Negative, False Positive and False Negative) (See Methods section above for details). Models which showed the least difference ($\leq$0.1) between the train and test MCC scores were considered to be reliable, as big differences between train and test scores imply that the 'learning' is more influenced by a given train-test split rather than any discernible patterns in the data. While there is no consensus regarding a minimum MCC threshold to be a good/strong predictor, the higher the absolute value of MCC (ranges between -1 and 1), the better the model is considered at making predictions. It is generally considered that an MCC>0.6 is considered a good result, MCC of 0.5 is considered a moderate result, 0.3<MCC<0.5 is considered to be an acceptable result, while an MCC<0.3 is considered to be a poor result.

gene-drug	Sensitive (%)	Resistant (%)	Total
Individual models			
alr-DCS	269 (99.26)	2 (0.74)	271
embB-EMB	731 (85.20)	127 (14.80)	858
gidB-STR	493 (92.84)	38 (7.16)	531
katG-INH	448 (54.83)	369 (45.17)	817
pncA-PZA	174 (41.04)	250 (58.96)	424
rpoB-RFP	803 (70.94)	329 (29.06)	1133
Combined model			
Data from five genes (excludes *alr*)	2649 (70.94)	1133 (29.06)	3762

Table 1: Summary of data used for machine learning
Numbers of sensitive and resistant mutations used for machine learning analysis. Abbreviations used: DCS: cycloserine, EMB: ethambutol, STR: streptomycin, INH: isoniazid, PZA: pyrazinamide, RFP: rifampicin.

11.1.3 Individual gene-drug model

Due to the presence of only two resistant mutations in *alr*, data from *alr* could not be used for individual gene-target ML analysis **(Table 1)**. Further, running ML models on *gidB*-STR revealed inconclusive results due to most predictors (ML models) returning training scores ∼0 (Appendix Table 11.C.2) irrespective of how the data was split, highlighting the heavily imbalanced mutation class distribution with less than 10% of resistant mutations in the data available. The imbalanced distribution may suggest the difficulty associated with using DST data to classify drug sensitivity for GidB.

The train-test split achieved by the scaling law principle (see Methods above for details) was consistently the best across all gene targets. Further, the choice of resampling type (random oversampling, random undersampling, over and undersampling, and SMOTE) improved the training MCC scores (up to 0.89 for *embB*-EMB with SMOTE resampling) but did not make any significant improvements to reduce the difference between train-test MCC scores. Thus, data without any resampling consistently performed the best across the gene targets analysed. There is no consensus in the literature with respect to the need to consider train-test MCC score differences. However, in these analyses, a train-test MCC score difference (absolute value $\leq$0.1) was considered important in ensuring a conservative approach to the interpretation of results, and to allow stochastic differences between train and test splits during ML model runs.

Predicting EMB resistance

For *embB*-EMB when considering all features, the Bagging Classifier achieved a comparable train and test MCC score of 0.40 and 0.42 respectively. With feature selection included, a similar performance was achieved though by a different model, XGBoost with MCC scores of 0.4 and 0.42 for train and test sets respectively **(Table 2)**. When the test scores are higher than the train scores, this usually suggests

inaccuracies with respect to train-test split where a random test set chosen during the model run can perform better than the training data by chance. However, in this case the difference of -0.02 between the train-test MCC scores is considered acceptable as per our criteria established in the section above. The nine contributing features were: ConSurf, SNAP2, PROVEAN, DeepDDG, distance from EMB, residue depth, MAF, SAV frequency, as well as site frequency in the dataset (Appendix Table 11.C.6). The relatively modest MCC score of 0.4 indicates effects due to either imbalanced data (since MCC accounts for true and false positives and negatives), with only 15% resistant mutations **(Table 1)**, or potentially a lack of consensus among different data sources for classifying EMB sensitivity as per DST.

A summary table of all classification model predictions for EMB with all features appears in Appendix Table 11.C.1, and results after feature selection are available in Appendix Table 11.C.6. Features used in ML analyses are described in Appendix Table 11.B.1.

Test *gene*-drug target	Best model post feature selection	Train MCC	Test MCC	Difference (Train and Test MCC)	Features selected
A) Individual models					
embB-EMB	XGBoost	0.40	0.42	-0.02	9
katG-INH	LDA and Ridge	0.37	0.42	-0.05	11
pncA-PZA	MLP	0.50	0.52	-0.02	14
rpoB-RFP	XGBoost and Extra Trees	0.45 0.44	0.49 0.48	-0.04	22
B) Combined model					
embB-EMB	Random Forest	0.47	0.34	0.13	32
katG-INH	Stochastic Descent	0.41	0.31	0.10	31
pncA-PZA	Extra Trees	0.40	0.46	-0.06	24
rpoB-RFP	MLP	0.46	0.39	0.07	30

Table 2: Summary of ML predictions post feature selection
Best performing models using a 10-fold stratified cross validation. The higher the MCC score, the greater the confidence associated with the model prediction. The smaller the difference between the train and test MCC scores, the greater the consistency in model performance. A negative value indicates that the test MCC score is higher than the training MCC. **A)** Individual gene-target model using a scaling law train/test split, **B)** Combined model consisting of data from genes *embB*, *gidB*, *katG*, *pncA*, and *rpoB* following a 'leave-one-gene-out' approach to test predictions on the 'left out' (i.e. test) gene. Only *pncA*-PZA prediction achieved a higher test MCC score compared with training. Abbreviations used: DCS: cycloserine, EMB: ethambutol, INH: isoniazid, PZA: pyrazinamide, RFP: rifampicin.

Predicting INH resistance

Performance of XGBoost (an extreme form of Gradient Boosting, see Appendix Table 11.A.1) and

Gradient Boosting models achieved similar MCC scores for train (≥0.32) and test (≥0.35) sets with all features present in the model. This was slightly lower than the *embB*-EMB MCC scores (∼0.4). With feature selection included, however, LDA and Ridge classifier models were the best with train MCC score of 0.37 and test MCC score of 0.42 for both **(Table 2)**. The 11 contributing features were: ConSurf, SNAP2, PROVEAN, DeepDDG, FoldX, average stability, distance from INH and PPI, residue depth, and AAindex properties (BENS940104 and GIAG010101: defined in AAindex2 related to amino acid mutation matrices) (Appendix Table 11.C.8). BENS940104 assigns a numerical value to an amino acid using an amino acid substitution matrix based on functionally constrained divergent evolution of protein sequences,[13] and GIAG010101 assigns a value to amino acids using an amino acid substitution matrix calculated on the temperature dependant structural adaptation of enzymes.[14] Similar to EMB prediction, the MCC score of 0.4 in the final model(s) are modest, despite the more balanced DST categorised dataset with 54% sensitive and 45% resistant mutations **(Table 1)**. It may be that interactions between mutations in *katG*, or the presence of a few mutations with prominent effects, not reflected in the features captured here, may be stronger drivers of INH resistance.

A summary table of all classification model predictions for INH with all features appears in Appendix Table 11.C.3, and results after feature selection are available in Appendix Table 11.C.8. Features used in ML analyses are described in Appendix Table 11.B.1.

Predicting PZA resistance

Random Forest was the best model when considering all features, and achieved a train MCC score of 0.49 and a test MCC score of 0.58. With feature selection included, however, the MLP model performed best with comparable MCC score of 0.50 and 0.52 between train and test sets respectively **(Table 2)**. Other models with comparable performance were Logistic Regression, Gaussian Process, LDA, and Ridge Classifier with a difference of 0.07 between their train and test MCC scores of ∼0.45 and ∼0.52 respectively. The 14 contributing features were: ConSurf, SNAP2, PROVEAN, DeepDDG, FoldX, distance from PZA, residue depth, relative surface area, AAindex properties (DOSZ010103 and RISJ880101: defined in AAindex2 related to amino acid mutation matrices), MAF, SAV frequency and site frequency in the dataset, as well as the frequency of mutational sites in the gene (Appendix Table 11.C.9). AAindex2 DOSZ010103 assigns a numerical value to each amino acid using a similarity matrix based on the THREADER force field,[15] and AAindex2 RISJ880101 derives its numerical value based on a scoring matrix of amino acid substitutions in structurally related proteins.[16]

A summary table of all classification model predictions for PZA with all features appears in Appendix

Table 11.C.4, and results after feature selection are available in Appendix Table 11.C.9. A consistent and comparable performance between MCC scores with all features and post feature selection delivers high confidence in the use of ML approach to predict PZA resistance. Features used in ML analyses are described in Appendix Table 11.B.1.

Predicting RFP resistance

Logistic Regression was the best model when considering all features, and achieved a train MCC score of 0.40 and a test MCC score of 0.47. With feature selection included, however, XGBoost and Extra Trees showed equivalent performance with an MCC difference of 0.04 between train and test sets. XGBoost achieved a train MCC score of 0.45 and a test MCC score of 0.49, while Extra Trees achieved a train MCC score of 0.44 and a test MCC score of 0.48 **(Table 2)**. The 22 contributing features were: ConSurf, SNAP2, PROVEAN, mCSM-DUET stability change, DeepDDG, FoldX intermolecular interactions, distance from RFP and RFP binding affinity changes, distance from NA and NA binding affinity changes, distance to PPI, relative surface area, AAindex property (MIYT790101: defined in AAindex2 related to amino acid mutation matrices), site frequency in the dataset, as well as the frequency of mutational sites in the gene, lineage and active site residue contribution (Appendix Table 11.C.10). AAindex2 MIYT790101 uses two types of amino acid substitution matrices to calculate amino acid pair distance based on evolution to account for physicochemical differences.[17]

A summary table of all classification model predictions for RFP with all features appears in Appendix Table 11.C.5, and results after feature selection are available in Appendix Table 11.C.10 Features used in ML analyses are described in Appendix Table 11.B.1.

11.1.3.1 Individual gene-drug model ML analysis summary

Irrespective of the target, all measures of evolutionary conservation (ConSurf, SNAP2, PROVEAN), protomer stability changes from DeepDDG, and distance from the drug were the common features chosen as part of the feature selection process, underscoring the importance of these measures in predicting drug resistance. PZA and RFP resistance predictions were the best among the individual gene-target models with test MCC score of ∼0.5. Drug binding affinity changes was among the features selected for RFP resistance prediction, but not for PZA. Since *pncA* is a non-essential gene, changes in PncA-PZA binding affinity are unlikely to contribute to resistance. For RFP, as it binds to RpoB RNA polymerase β subunit, part of the large RNA polymerase complex, distance to RNA, and PPI were among the features chosen. Mutational change in NA binding affinity in RpoB RNA polymerase β subunit was among the chosen features that highlight the importance of these sites, while changes in the PPI was not. Another interesting observation was that the active site residues (RRDR region)

of RpoB RNA polymerase β subunit, as well as those beyond the active site, were together chosen as part of the feature selection process, stressing the importance of mutational effects extending beyond the RRDR, impacting RFP binding and contributing to resistance development. Together, these findings suggest that the features selected by the ML models for predicting individual drug resistance are meaningful as they offer biological insights that link with observations made in the individual exploratory analyses. The poor performance of INH resistance prediction could perhaps be explained by the dominant role played by the single mutation S315T which masks all other mutational effects, which are given equal weight in the ML analysis. For *embB*, the low reliability of DST data (owing to a lack of consensus to classify EMB drug sensitivity) is likely to be a bigger contributor compared with the imbalanced data for its low performance, since RFP also has imbalanced data, but is able to achieve an MCC of 0.5.

11.1.4 Combined model

The combined model, with a 'leave-one-gene-out' approach, was adopted to investigate the feasibility of developing a gene-agnostic ML predictor. Data for each gene was iteratively excluded, and training was performed on data from all other genes. In this way, performance of multiple ML models was evaluated. Considering the limited resistant mutations (n=2) in the *alr* dataset **(Table 1)**, *alr* was excluded as a test gene.

When *alr* data was included in the training set, all models performed consistently worse, on individual train and test MCC scores, as well as at minimising the difference between train and test MCC scores. This is indicative of the poor quality of aggregate DST data used for classifying *alr* mutations. Therefore, *alr* was excluded from training data, and ML models were developed from the remaining five genes (*embB*, *gidB*, *katG*, *pncA*, and *rpoB)*. Further, like the individual gene-drug target model, the choice of resampling type (random oversampling, random undersampling, over and undersampling, and SMOTE) improved the training MCC scores, but did not make any improvements to reduce the difference between train-test MCC scores. As resampling is applied only to training data, naturally the training scores are expected to improve as observed in our analyses. The performance on the test set revealed that the differences between the train-test MCC scores were much greater for all genes than initially specified (train-test difference $\leq$0.1). Hence, data without any resampling consistently performed the best for the combined model. Full results for these analyses showing the comparative performance for all models and all genes are best viewed via the interactive ML dashboard which can be accessed via https://thesis.tunstall.in, while the individual drug prediction results from a combined approach are indicated below. A description of the the ML dashboard is available

in Appendix 11.D.1.

As a first attempt to generate a combined model predicting resistance, a test MCC score of at least 0.4, along with a ≤0.1 difference between train and test MCC, would indicate the potential of pursuing such a gene-agnostic approach in predicting resistance.

Model evaluation with all features: Most models achieved a training MCC score of at least 0.4 across all the test genes. The test MCC scores, however, were mostly lower (≤0.3) compared with the respective train MCC score for all genes except for the single model Stochastic Descent, used in predicting PZA resistance. The model achieved train and test MCC scores of 0.34 and 0.39 respectively. There were others models, namely: Bagging Classifier, Gradient Boosting, and XGBoost, with train and test MCC scores ∼0.4, with the test MCC score being marginally lower than the train MCC score as expected. Similar to the individual gene models, GidB results were inconclusive due to most test MCC scores being close to zero (Appendix Table 11.C.12). Summary tables of all classification models used in the combined approach to predict resistance are available in Appendix tables: EMB (Appendix Table 11.C.11), INH (Appendix Table 11.C.13), PZA (Appendix Table 11.C.14), and RFP (Appendix Table 11.C.15). Features used in ML analyses are described in Appendix Table 11.B.1.

Model evaluation with feature selection: Model performance generally improved with feature selection, though this did not occur consistently across all gene-drug targets (https://thesis.tunstall.in). Prediction of INH resistance was the poorest (most test MCC scores <0.25), followed by prediction of EMB resistance (most test MCC scores <0.3). The best resistance prediction was for PZA, with RFP resistance prediction also showing potential (test MCC score approaching 0.4) in this combined approach **(Table 2)**. Summary tables of all classification models used in the combined approach to predict resistance are available in Appendix tables: EMB (Appendix Table 11.C.16), INH (Appendix Table 11.C.18), PZA (Appendix Table 11.C.19), and RFP (Appendix Table 11.C.20).

The combined model approach performed the best for PZA resistance prediction. Extra Trees was the best model with the highest test MCC score of 0.46 (train MCC score of 0.4) **(Table 2)**, followed by Random Forest test MCC score of 0.44 (train MCC score of 0.42) and XGBoost train and test sets with MCC scores of 0.4. There were a total of 24 features selected: ConSurf, PROVEAN, SNAP2, mCSM-DUET, FoldX, DeepDDG, Dynamut2, average stability changes, FoldX interactions, distance and binding affinity changes related to drugs (EMB, INH, RFP and STR), relative surface area, residue depth and hydrophobicity, AAindex property (OVEJ920102: defined in AAindex2 related to amino acid mutation matrices), SAV and site frequency in the dataset, frequency of mutational sites in the gene, and lineage contribution. Features used in ML analyses including the selected AAindex

properties are described in Appendix Table 11.B.1.

11.1.4.1 Combined model ML analysis summary

The combined model performance achieved the minimum threshold (MCC of 0.4) defined at the start of the analysis, and the training performance was consistently on par with this threshold for most models across the five genes, highlighting the combined model's potential to learn across different gene-drug targets. The test performance, however, was variable across the five genes with EMB and INH resistance predictions among the poorest for most models: test MCC <0.3, a difference of >0.1 between train and test MCC scores. The drug with the best predictions was PZA, with close agreement between the train and test MCC scores (<0.1) for most models. This highlights that prediction of PZA resistance by a gene-agnostic ML approach offers promise. RFP prediction followed closely with test MCC scores for most models approaching 0.4 but with test scores being lower than training. Considering that EMB and RFP directly bind to their respective proteins while INH and PZA are prodrugs, the inconsistency in model predictions is independent of the underlying fundamental functional roles associated with the genes (i.e. essential vs. non-essential genes). It is more likely that the data integrity for EMB (low reliability of EMB DST data) and a lack of adequate weighting strategy to compensate for the dominant effect of *katG* S315T, a highly frequent and resistant mutation, contribute to their poor predictions. Similar to the individual gene models, *gidB*-STR ML analyses were inconclusive due to most predictors returning test scores ∼0 suggesting low reliability of DST data leading to a heavily imbalanced mutation class distribution with less than 10% resistant mutations available in the data. Ultimately, a combined model that is able to 'learn' across targets in order to predict resistance is of great utility as evinced by these analyses. However, further improvements with inclusion of additional gene-target data requires careful consideration of DST data quality and knowledge-based approaches for model optimisations.

11.1.5 Chapter Summary

PZA resistance prediction using ML models performed consistently in the individual gene-drug target (PncA-PZA) model, as well as in a combined model. Measures of evolutionary conservation, residue depth and relative surface area, distance from drug, mutational frequency and mutational position frequency, along with stability effects from DeepDDG were shared between the two ML approaches indicating the strong association of these features with resistance prediction. Naturally, additional features that are applicable for individual genes like PPI and NA distances, along with changes in molecular interactions and lineage contribution to a given mutation (measured by how many distinct

lineages contribute to a given mutation), are some of the more general features chosen during the feature selection process. This reveals some of the shared molecular features and their utility in predicting resistance. An advantage of using a gene-agnostic ML approach for building a general AMR predictor is that it allows smaller diverse datasets to be exploited in a more targeted manner provided that data integrity is maintained. For example, if the aggregate DST data across the targets were consistent, predicting DCS resistance in *alr* would have been possible.

11.2 ML results dashboard

An interactive dashboard, available at https://thesis.tunstall.in, was built as part of this thesis to explore each of the six gene-drug target pairs as well as the ML analysis results. A screenshot of this dashboard is available in Appendix 11.D.1.

References

[1] F. Pedregosa et al. "Scikit-Learn: Machine Learning in Python". In: *Journal of Machine Learning Research* 12 (2011), pp. 2825–2830.

[2] K. Nakai, A. Kidera, and M. Kanehisa. "Cluster Analysis of Amino Acid Indices for Prediction of Protein Structure and Function". In: *Protein Engineering* 2.2 (July 1988), pp. 93–100. ISSN: 0269-2139. DOI: 10.1093/protein/2.2.93.

[3] K. Tomii and M. Kanehisa. "Analysis of Amino Acid Indices and Mutation Matrices for Sequence Comparison and Structure Prediction of Proteins." In: *Protein engineering* (1996). DOI: 10.1093/PROTEIN/9.1.27.

[4] S. Kawashima, H. Ogata, and M. Kanehisa. "AAindex: Amino Acid Index Database". In: *Nucleic Acids Research* 27.1 (Jan. 1, 1999), pp. 368–369. ISSN: 0305-1048. DOI: 10.1093/nar/27.1.368.

[5] S. Kawashima and M. Kanehisa. "AAindex: Amino Acid Index Database". In: *Nucleic Acids Research* 28.1 (Jan. 1, 2000), p. 374. ISSN: 0305-1048. DOI: 10.1093/nar/28.1.374.

[6] Shuichi Kawashima et al. "AAindex: Amino Acid Index Database, Progress Report 2008". In: *Nucleic Acids Research* 36 (Database issue Jan. 2008), pp. D202–205. ISSN: 1362-4962. DOI: 10.1093/nar/gkm998.

[7] Giovanna Menardi and Nicola Torelli. "Training and Assessing Classification Rules with Imbalanced Data". In: *Data Mining and Knowledge Discovery* 28.1 (Jan. 1, 2014), pp. 92–122. ISSN: 1573-756X. DOI: 10.1007/s10618-012-0295-5.

[8] Guillaume Lemaître, Fernando Nogueira, and Christos K. Aridas. "Imbalanced-Learn: A Python Toolbox to Tackle the Curse of Imbalanced Datasets in Machine Learning". In: *Journal of Machine Learning Research* 18.17 (2017), pp. 1–5.

[9] Nitesh V Chawla et al. "SMOTE: Synthetic Minority over-Sampling Technique". In: *Journal of artificial intelligence research* 16 (2002), pp. 321–357.

[10] Isabelle Guyon. "A Scaling Law for the Validation-Set Training-Set Size Ratio". In: 1997.

[11] Miron B. Kursa, Aleksander Jankowski, and Witold R. Rudnicki. "Boruta A System for Feature Selection". In: *Fundamenta Informaticae* 101.4 (Jan. 1, 2010), pp. 271–285. ISSN: 0169-2968. DOI: 10.3233/FI-2010-288.

[12] *BorutaPy*. Daniel Homola. May 8, 2015. URL: https://danielhomola.com/feature%20selection/phd/borutapy-an-all-relevant-feature-selection-method/ (visited on 09/20/2022).

[13] S. A. Benner, M. A. Cohen, and G. H. Gonnet. "Amino Acid Substitution during Functionally Constrained Divergent Evolution of Protein Sequences". In: *Protein Engineering* 7.11 (Nov. 1994), pp. 1323–1332. ISSN: 0269-2139. DOI: 10.1093/protein/7.11.1323.

[14] G. Gianese, P. Argos, and S. Pascarella. "Structural Adaptation of Enzymes to Low Temperatures". In: *Protein Engineering* 14.3 (Mar. 2001), pp. 141–148. ISSN: 0269-2139. DOI: 10.1093/protein/14.3.141.

[15] Z. Dosztányi and A. E. Torda. "Amino Acid Similarity Matrices Based on Force Fields". In: *Bioinformatics (Oxford, England)* 17.8 (Aug. 2001), pp. 686–699. ISSN: 1367-4803. DOI: 10.1093/bioinformatics/17.8.686.

[16] J. L. Risler et al. "Amino Acid Substitutions in Structurally Related Proteins. A Pattern Recognition Approach. Determination of a New and Efficient Scoring Matrix". In: *Journal of Molecular Biology* 204.4 (Dec. 20, 1988), pp. 1019–1029. ISSN: 0022-2836. DOI: 10.1016/0022-2836(88)90058-7.

[17] T. Miyata, S. Miyazawa, and T. Yasunaga. "Two Types of Amino Acid Substitutions in Protein Evolution". In: *Journal of Molecular Evolution* 12.3 (Mar. 15, 1979), pp. 219–236. ISSN: 0022-2844. DOI: 10.1007/BF01732340.

Appendix for Chapter 11

11.A Classification models used for supervised machine learning

Linear classifiers
Models using a linear combination of its input features to make a classification decision.

Logistic Regression scikit-learn-LR **Logistic Regression CV** scikit-learn-LRCV	Despite the name, it is a classification algorithm. Instead of fitting a straight line like in a simple linear regression model, an S shaped logistic function is fitted to predict a binary outcome. The curve indicates the likelihood of the occurrence of a given event. While less prone to over-fitting, it is weak in learning complex patterns. An extended version with cross validation called Logistic Regression CV is also available in scikit-learn.
Stochastic Gradient Descent (SGD) scikit-learn-SGD	Gradient descent refers to descending the slope (gradient) to reach the lowest/minimum point where stochastic refers to a process linked with a random probability. It is an iterative algorithm which starts from a random point on a function and descends down its slope in steps until it reaches the lowest point of that function. The step size is an important parameter in this process. With SGD, a few samples are randomly selected instead of the entire input data in each iteration. SGD requires a number of iterations, and is sensitive to feature scaling (an important data-preprocessing step).
Gaussian Processes Classifier scikit-learn-GPC	A non-parametric algorithm applied to classification tasks. Stochastic in nature, Gaussian Processes are a type of kernel model (methods using a linear classifier to solve a non-linear problem) which are a generalisation of the Gaussian probability distribution. Rather than summarising the distribution of random variables like in a Gaussian probability distribution, Gaussian Processes summarise the properties of the parameters of the functions, which jointly have a Gaussian distribution.
Passive Aggressive Classifier scikit-learn-PAC	A type of online-learning algorithm where input data arrives sequentially and the model is updated step-by-step, contrary to batch learning where the entire input data is processed at once. This is particularly useful in scenarios where training on an entire dataset is computationally infeasible. The name of the algorithm is based on the models behaviour where *Passive* implies keeping the model without any changes if the prediction is correct, while *Aggressive* refers to making some changes to the model if the prediction is incorrect.
Ridge Classifier scikit-learn-RC **Ridge Classifier CV** scikit-learn-RCCV	Adapted from linear regression with an added penalty term in its loss function. Where linear regression is not penalised for its choice of weights assigned to features, Ridge penalises the model for the sum of squared value of the weights. This results in smaller absolute values for weights with extreme weights being penalised to ensure an even distribution of weights. An extended version with cross validation called Ridge Classifier CV is also available in scikit-learn.

Support Vector Classification (SVC) scikit-learn-SVC	Aims to find the best hyperplane in an n-dimensional space (n is the number of input features) that best separates the classes. Best is defined as the hyperplane that has the maximum distance between the data points from both classes. It is a memory efficient algorithm as it uses a subset of training data points in its decision function.
Linear Discriminant Analysis (LDA) scikit-learn-LDA	Used as a dimensionality reduction as well as classification tool. It works by maximising the separability among the known categories rather than maximising the variation like in Principal Component Analysis (PCA).
Tree based methods These use a series of if-then rules to generate predictions from one or more decision trees.	
Decision tree scikit-learn-DT	The foundation of all tree-based models, where data is divided into smaller and smaller subsets or nodes as the tree develops. It has three main parts; a root node (starting point of the tree), leaf node (decision criteria) and branches. Both the root and leaf node contain criteria to be answered with branches denoting the flow from question to answer. They are simple and fast, but are very sensitive to changes in input data.
Extra Trees scikit-learn-ETS	An ensemble method composed of a large number of decision trees where the final result accounts for predictions from every tree. It uses the entire input data at the start with the number of split nodes being randomly chosen unlike in the Random Forest model. At each node split, it fits randomised decision trees (i.e. extra trees) on sub samples of the data with averaging to improve accuracy and control over-fitting. In this way, it adds randomisation as well as optimisation to improve model performance.
Extra Tree scikit-learn-ET	An extremely randomised tree-based classifier. It differs from Decision Trees in their construction where the best split to separate the samples of a node into two groups is chosen based on random splits for the pre specified number of features to consider to obtain the best split.
Bagging based methods Bagging is the technique of constructing multiple decision tree models at a time by randomly sampling with replacement, or bootstrapping from the input data. It is one of the oldest and simplest ensemble-based methods applied to tree-based algorithms to enhance performance by reducing variance and over-fitting.	
Bagging classifier scikit-learn-BC	Base classifiers are fitted on random subsets of the input data. The final prediction is an aggregate of all the individual predictions calculated either by voting or averaging. Introducing randomisation in this manner is intended to reduce variance due to ensemble predictions being generated from random subsets of the data.

Random forest scikit-learn-RF	Constructing several decision trees which can be trained in parallel, where a group of trees is referred to as a forest. The final output is either the mean of all decision trees in a regression task or majority voting in a classification problem. Builds on the bagging method, where, on top of building several trees from the sampled dataset, each node is split on a random selection of the models input features.
Boosting based methods Boosting is a strategy where multiple simple models are combined into a single model, as such a combination is thought to result in a stronger predictor. In boosting terminology, the simple models are called weak models or weak learners. These algorithms are tree-based ensemble methods where trees are built sequentially. The most common algorithm used is a decision tree model.	
Adaptive Boosting Classifier (AdaBoost) scikit-learn-ABC	Starts by assigning equal weights to all data points, followed by assigning higher weights to points that are wrongly classified. Thus, all data points with higher weights are given more importance iteratively in subsequent models. As such the method focuses on training misclassified observations. The weak learners in this method are a basic form of decision tree, also known as stumps. Each learners prediction is weighted by its individual accuracy, with the final prediction based on a majority vote.
Gradient Boosting Classifier scikit-learn-GBC	Sequentially built models, with every subsequent model aimed at reducing the errors of the previous one. The new model is built on the errors or residuals of the previous model, with the aim of minimising the loss function of a learner. Weak learners are decision trees constructed in a greedy manner using the loss function and have equal weights.
Extreme Gradient Boosting (XGBoost) scikit-learn-nb-XGBoost	One of the most popular variants of gradient boosting which is optimised and distributed for efficiency and portability. It uses a pre-sorted and histogram-based algorithm for computing the best split, and implements parallel boosting. All data points for a given feature are split into discrete bins in order to find the split value of the histogram. The trees can have a varying number of terminal nodes, with proportional shrinkage applied. An extra randomisation parameter is added to reduce correlation between trees.
Naive Bayes (NB) Methods As the name implies, the algorithms in this family are predicated on Bayes theorem. However, they are based on simple (i.e. naive) assumptions, where each feature assumes independence, a feat seldom true in reality. NB Classifiers are easy and fast to implement.	
Gaussian NB scikit-learn-GNB	Computationally lightweight and built by calculating the mean and standard deviation of the training data.
Multinomial NB scikit-learn-MNB	Implements the NB algorithm for multinomially distributed data. The distribution is parametrised according to probability of a given feature belonging to a certain class.
Complement NB scikit-learn-CNB	Adaptation of the Multinomial NB algorithm designed to correct the severe assumptions made by the standard Multinomial NB model. It is particularly suited for imbalanced data.

Others	
Quadratic Discriminant Analysis (QDA) scikit-learn-QDA	A variant of LDA where an individual covariance matrix is estimated for every observation class. It has a quadratic decision boundary, but unlike LDA cannot be used as a dimensionality reduction technique.
K Nearest neighbours (KNN) scikit-learn-KNN	Uses the technique of nearest neighbours or feature similarity to predict which cluster the new data will fit into. An integer value of K is required to decide how many nearest data points the algorithm needs to consider, as well as a distance metric (Euclidean, Manhattan, Minkowski) to calculate the nearest neighbour. The standard value of K used is 5 with Euclidean as the distance metric. While computationally expensive, and sensitive to irrelevant features and imbalanced datasets, it is easy to implement.
Multi-Layer Perceptron Classifier (MLP) scikit-learn-MLP	A perceptron is a linear classifier. MLP is composed of more than one perceptron (multi-layer) and is a deep, artificial neural network (ANN). The layers are typically an input layer (receives input), output layer (makes the prediction) and an arbitrary amount of hidden layers. These hidden layers are the engine of this algorithm, with 100 being the default in scikit-learn.

Table 11.A.1: Summary of the classification models used in machine learning from scikit-learn, version 1.1.1

11.B Features used in machine learning

Feature name	Feature type	Feature category	Features (n)	Tools	Comment
mCSM-DUET stability change	numerical	Structure	35	mCSM-lig	A table with web URLs is available in Chapter 2, Methods, Appendix Table 2.A
FoldX stability change	numerical			FoldX	
DeepDDG stability change	numerical			DeepDDG	
Dynamut2 stability change	numerical			Dynamut2	
Drug binding affinity change	numerical			mCSM-lig	
Drug binding affinity change	numerical				
PP interface binding affinity change	numerical			mCSM-PPI2	
RNA binding affinity change	numerical			mCSM-NA	
Distance to drug	numerical			mCSM-lig	
Distance to PP interface	numerical			mCSM-PPI2	
Distance to RNA	numerical			mCSM-NA	
Molecular contacts	numerical			FoldX interaction components	
electro: rr, mm, sm, ss	numerical			FoldX interaction components	
disulfide: rr, mm, sm, ss	numerical			FoldX interaction components	
hbonds: rr, mm, sm, ss	numerical			FoldX interaction components	
partcov: rr, mm, sm	numerical			FoldX interaction components	
vdwclashes: sm, ss, rr, mm	numerical			FoldX interaction components	
volumetric: ss: rr, mm, ss	numerical			FoldX interaction components	
Relative Surface Area	numerical	Residue level properties	4	DSSP	
Residue Depth	numerical			RD depth server	
Residue hydrophobicity	numerical			Exapsy: Kyte & Doolittle	
Active site and binding partner residue indication	numerical			LigPlus, Arpeggio, PLIP	
ConSurf	numerical	Evolutionary conservation	3	ConSurf	
SNAP2	numerical			SNAP2	
PROVEAN	numerical			PROVEAN	
Average mutational frequency	numerical	Genomics	6	Calculated	
Mutational position frequency	numerical			Calculated	
Lineage contribution: count of distinct no. of lineages	numerical			Calculated	
Lineage contribution: total number of lineage occurrences	numerical			Calculated	
Lineage contribution: count of	numerical			Calculated	

Table 11.B.1: Summary of features used in machine learning

distinct lineages/all possible lineages					
Lineage contribution: total number of lineage occurrence/total number of samples for the gene	numerical			Calculated	
AA index property	numerical	AA index properties	>120	AA index	
Selected during feature selection from AAindex2 https://www.genome.jp/aaindex/AAindex/list_of_matrices					
BENS940104: Genetic code matrix	numerical				
DOSZ010103:An amino acid similarity matrix based on the THREADER force field					
GIAG010101:Residue substitutions					
MIYT790101:Amino acid pair distance					
OVEJ920102: Environment-specific amino acid substitution matrix for alpha residues					
RISJ880101: Scoring matrix					
Secondary structure information	categorical	Secondary structure	1	DSSP	
Residue property: water	categorical	Amino acid properties	4	Amino acid properties	
Residue property: polarity	categorical				
Residue property: charge	categorical				
Residue property: any change	categorical				

11.C ML Model Tables

Model Name	MCC Train	MCC Test	F1 Train	F1 Test	Accuracy Train	Accuracy Test	Recall Train	Recall Test	Precision Train	Precision Test	JCC Train	JCC Test
AdaBoost Classifier	0.22	0.57	0.29	0.63	0.84	0.89	0.24	0.60	0.39	0.67	0.17	0.46
Bagging Classifier	0.40	0.42	0.39	0.43	0.88	0.88	0.28	0.30	0.75	0.75	0.25	0.27
Complement NB	0.27	0.23	0.39	0.37	0.73	0.74	0.57	0.50	0.30	0.29	0.25	0.23
Decision Tree	0.24	0.40	0.35	0.47	0.80	0.86	0.38	0.40	0.35	0.57	0.23	0.31
Dummy Classifier	0.00	0.00	0.00	0.00	0.85	0.85	0.00	0.00	0.00	0.00	0.00	0.00
Extra Tree	0.09	0.23	0.22	0.37	0.77	0.74	0.24	0.50	0.22	0.29	0.13	0.23
Extra Trees	0.12	0.11	0.14	0.15	0.84	0.83	0.09	0.10	0.36	0.33	0.08	0.08
Gaussian NB	0.24	0.35	0.37	0.45	0.73	0.82	0.53	0.50	0.29	0.42	0.23	0.29
Gaussian Process	0.08	0.29	0.07	0.18	0.85	0.86	0.04	0.10	0.38	1.00	0.04	0.10
Gradient Boosting	0.31	0.52	0.33	0.53	0.86	0.89	0.24	0.40	0.59	0.80	0.20	0.36
K-Nearest Neighbors	0.09	-0.05	0.09	0.00	0.85	0.83	0.06	0.00	0.32	0.00	0.06	0.00
LDA	0.12	0.40	0.23	0.47	0.80	0.86	0.21	0.40	0.27	0.57	0.13	0.31
Logistic Regression	0.16	0.42	0.18	0.33	0.85	0.88	0.12	0.20	0.41	1.00	0.10	0.20
Logistic RegressionCV	0.03	0.29	0.04	0.18	0.85	0.86	0.02	0.10	0.12	1.00	0.02	0.10
MLP	0.18	0.20	0.25	0.27	0.83	0.83	0.20	0.20	0.37	0.40	0.15	0.15
Multinomial NB	0.12	-0.05	0.10	0.00	0.85	0.83	0.06	0.00	0.43	0.00	0.06	0.00
Passive Aggressive	0.14	0.29	0.19	0.18	0.77	0.86	0.24	0.10	0.36	1.00	0.11	0.10
QDA	-0.03	-0.08	0.03	0.00	0.82	0.82	0.02	0.00	0.12	0.00	0.01	0.00
Random Forest	0.22	0.42	0.16	0.33	0.86	0.88	0.09	0.20	0.67	1.00	0.09	0.20
Ridge Classifier	0.12	0.42	0.15	0.33	0.84	0.88	0.10	0.20	0.33	1.00	0.09	0.20
Ridge ClassifierCV	0.12	0.29	0.12	0.18	0.85	0.86	0.07	0.10	0.42	1.00	0.06	0.10
Stochastic GDescent	0.17	0.23	0.25	0.33	0.78	0.82	0.30	0.30	0.34	0.38	0.15	0.20
SVC	0.00	0.00	0.00	0.00	0.85	0.85	0.00	0.00	0.00	0.00	0.00	0.00
XGBoost	0.33	0.52	0.34	0.53	0.86	0.89	0.25	0.40	0.62	0.80	0.22	0.36

Table 11.C.1: Individual model evalutation metrics for ethambutol resistance prediction: all features
Data with 178 *embB* features for 858 SAVs was split into train and test sets using the scaling law principle, and model performance assessed using a stratified 10-fold cross validation method. Abbreviations used: CV: cross validation, LDA: linear discriminant analysis, MLP: multilayer perceptron, NB: naive Bayes, QDA: quadratic discriminant analysis, SVC: support vector classification, MCC: Matthews correlation coefficient, JCC: jaccard similarity coefficient, SAV: single amino acid variation.

Model Name	MCC Train	MCC Test	F1 Train	F1 Test	Accuracy Train	Accuracy Test	Recall Train	Recall Test	Precision Train	Precision Test	JCC Train	JCC Test
AdaBoost Classifier	0.03	-0.07	0.09	0.00	0.88	0.88	0.08	0.00	0.10	0.00	0.05	0.00
Bagging Classifier	0.04	0.00	0.05	0.00	0.92	0.92	0.03	0.00	0.10	0.00	0.03	0.00
Complement NB	0.08	-0.11	0.17	0.08	0.56	0.45	0.59	0.33	0.10	0.05	0.10	0.04
Decision Tree	0.06	0.00	0.13	0.00	0.86	0.92	0.14	0.00	0.13	0.00	0.08	0.00
Dummy Classifier	0.00	0.00	0.00	0.00	0.93	0.92	0.00	0.00	0.00	0.00	0.00	0.00
Extra Tree	0.02	0.18	0.08	0.25	0.88	0.85	0.08	0.33	0.08	0.20	0.05	0.14
Extra Trees	0.07	0.00	0.08	0.00	0.92	0.92	0.06	0.00	0.15	0.00	0.05	0.00
Gaussian NB	0.10	0.04	0.17	0.15	0.45	0.42	0.78	0.67	0.10	0.08	0.09	0.08
Gaussian Process	0.00	0.00	0.00	0.00	0.93	0.92	0.00	0.00	0.00	0.00	0.00	0.00
Gradient Boosting	0.06	0.00	0.08	0.00	0.91	0.92	0.06	0.00	0.15	0.00	0.05	0.00
K-Nearest Neighbors	-0.01	0.00	0.00	0.00	0.92	0.92	0.00	0.00	0.00	0.00	0.00	0.00
LDA	0.03	0.22	0.08	0.29	0.89	0.88	0.06	0.33	0.13	0.25	0.05	0.17
Logistic Regression	0.00	0.00	0.00	0.00	0.93	0.92	0.00	0.00	0.00	0.00	0.00	0.00
Logistic RegressionCV	0.00	0.00	0.00	0.00	0.93	0.92	0.00	0.00	0.00	0.00	0.00	0.00
MLP	-0.03	0.00	0.00	0.00	0.90	0.92	0.00	0.00	0.00	0.00	0.00	0.00
Multinomial NB	-0.04	0.00	0.00	0.00	0.89	0.92	0.00	0.00	0.00	0.00	0.00	0.00
Passive Aggressive	-0.00	0.00	0.04	0.00	0.83	0.92	0.11	0.00	0.03	0.00	0.02	0.00
QDA	0.16	-0.08	0.19	0.00	0.92	0.85	0.18	0.00	0.22	0.00	0.13	0.00
Random Forest	0.00	0.00	0.00	0.00	0.93	0.92	0.00	0.00	0.00	0.00	0.00	0.00
Ridge Classifier	0.00	0.00	0.00	0.00	0.93	0.92	0.00	0.00	0.00	0.00	0.00	0.00
Ridge ClassifierCV	-0.00	0.00	0.00	0.00	0.93	0.92	0.00	0.00	0.00	0.00	0.00	0.00
Stochastic GDescent	-0.01	0.00	0.01	0.00	0.90	0.92	0.03	0.00	0.01	0.00	0.01	0.00
SVC	0.00	0.00	0.00	0.00	0.93	0.92	0.00	0.00	0.00	0.00	0.00	0.00
XGBoost	0.09	0.00	0.09	0.00	0.92	0.92	0.06	0.00	0.20	0.00	0.06	0.00

Table 11.C.2: Individual model evalutation metrics for streptomycin resistance prediction: all features
Data with 178 *gidB* features for 531 SAVs was split into train and test sets using the scaling law principle, and model performance assessed using a stratified 10-fold cross validation method. Abbreviations used: CV: cross validation, LDA: linear discriminant analysis, MLP: multilayer perceptron, NB: naive Bayes, QDA: quadratic discriminant analysis, SVC: support vector classification, MCC: Matthews correlation coefficient, JCC: jaccard similarity coefficient, SAV: single amino acid variation.

Model Name	MCC Train	MCC Test	F1 Train	F1 Test	Accuracy Train	Accuracy Test	Recall Train	Recall Test	Precision Train	Precision Test	JCC Train	JCC Test
AdaBoost Classifier	0.27	0.35	0.59	0.64	0.64	0.68	0.57	0.64	0.61	0.64	0.42	0.47
Bagging Classifier	0.33	0.26	0.62	0.61	0.67	0.63	0.60	0.64	0.65	0.58	0.45	0.44
Complement NB	0.21	0.25	0.59	0.63	0.60	0.61	0.63	0.71	0.55	0.56	0.42	0.45
Decision Tree	0.20	0.11	0.55	0.56	0.60	0.55	0.55	0.64	0.56	0.50	0.39	0.39
Dummy Classifier	0.00	0.00	0.00	0.00	0.55	0.55	0.00	0.00	0.00	0.00	0.00	0.00
Extra Tree	0.12	-0.03	0.51	0.47	0.56	0.48	0.50	0.50	0.52	0.44	0.34	0.30
Extra Trees	0.29	0.17	0.59	0.58	0.65	0.58	0.57	0.64	0.62	0.53	0.42	0.41
Gaussian NB	0.28	0.40	0.62	0.69	0.63	0.69	0.68	0.75	0.58	0.64	0.46	0.52
Gaussian Process	0.23	0.20	0.56	0.59	0.62	0.60	0.53	0.64	0.59	0.55	0.39	0.42
Gradient Boosting	0.32	0.38	0.61	0.65	0.66	0.69	0.59	0.64	0.64	0.67	0.45	0.49
K-Nearest Neighbors	0.15	0.20	0.51	0.59	0.58	0.60	0.49	0.64	0.55	0.55	0.35	0.42
LDA	0.30	0.25	0.61	0.60	0.65	0.63	0.62	0.61	0.61	0.59	0.44	0.42
Logistic Regression	0.35	0.31	0.63	0.62	0.68	0.66	0.62	0.61	0.66	0.63	0.47	0.45
Logistic RegressionCV	0.33	0.22	0.62	0.57	0.67	0.61	0.60	0.57	0.65	0.57	0.45	0.40
MLP	0.31	0.19	0.62	0.56	0.65	0.60	0.62	0.57	0.62	0.55	0.45	0.39
Multinomial NB	0.21	0.30	0.57	0.65	0.61	0.65	0.58	0.71	0.56	0.59	0.40	0.48
Passive Aggressive	0.16	0.39	0.50	0.40	0.55	0.66	0.68	0.25	0.56	1.00	0.36	0.25
QDA	0.13	0.03	0.63	0.61	0.50	0.47	0.94	0.93	0.47	0.46	0.46	0.44
Random Forest	0.34	0.23	0.62	0.60	0.67	0.61	0.59	0.64	0.65	0.56	0.45	0.43
Ridge Classifier	0.33	0.28	0.62	0.61	0.67	0.65	0.61	0.61	0.64	0.61	0.46	0.44
Ridge ClassifierCV	0.36	0.28	0.64	0.61	0.68	0.65	0.63	0.61	0.66	0.61	0.48	0.44
Stochastic GDescent	0.27	0.22	0.55	0.62	0.61	0.60	0.64	0.71	0.62	0.54	0.40	0.44
SVC	0.32	0.23	0.62	0.60	0.66	0.61	0.60	0.64	0.63	0.56	0.45	0.43
XGBoost	0.32	0.35	0.61	0.64	0.66	0.68	0.60	0.64	0.63	0.64	0.44	0.47

Table 11.C.3: Individual model evalutation metrics for isoniazid resistance prediction: all features
Data with 178 *katG* features for 817 SAVs was split into train and test sets using the scaling law principle, and model performance assessed using a stratified 10-fold cross validation method. Abbreviations used: CV: cross validation, LDA: linear discriminant analysis, MLP: multilayer perceptron, NB: naive Bayes, QDA: quadratic discriminant analysis, SVC: support vector classification, MCC: Matthews correlation coefficient, JCC: jaccard similarity coefficient, SAV: single amino acid variation.

Model Name	MCC Train	MCC Test	F1 Train	F1 Test	Accuracy Train	Accuracy Test	Recall Train	Recall Test	Precision Train	Precision Test	JCC Train	JCC Test
AdaBoost Classifier	0.37	0.17	0.74	0.65	0.70	0.59	0.74	0.63	0.74	0.67	0.59	0.48
Bagging Classifier	0.45	0.40	0.77	0.78	0.73	0.72	0.78	0.84	0.77	0.73	0.63	0.64
Complement NB	0.29	0.37	0.70	0.72	0.65	0.69	0.71	0.68	0.70	0.76	0.55	0.57
Decision Tree	0.28	0.35	0.70	0.74	0.65	0.69	0.71	0.74	0.70	0.74	0.55	0.58
Dummy Classifier	0.00	0.00	0.74	0.75	0.59	0.59	1.00	1.00	0.59	0.59	0.59	0.59
Extra Tree	0.19	-0.09	0.67	0.62	0.61	0.50	0.67	0.68	0.67	0.57	0.50	0.45
Extra Trees	0.38	0.26	0.76	0.74	0.71	0.66	0.80	0.84	0.73	0.67	0.62	0.59
Gaussian NB	0.31	0.63	0.71	0.83	0.66	0.81	0.70	0.79	0.72	0.88	0.55	0.71
Gaussian Process	0.18	-0.04	0.70	0.65	0.62	0.53	0.76	0.74	0.66	0.58	0.54	0.48
Gradient Boosting	0.45	0.54	0.78	0.82	0.73	0.78	0.79	0.84	0.77	0.80	0.64	0.70
K-Nearest Neighbors	0.06	0.22	0.66	0.68	0.56	0.62	0.72	0.68	0.61	0.68	0.49	0.52
LDA	0.35	0.30	0.73	0.70	0.68	0.66	0.74	0.68	0.73	0.72	0.58	0.54
Logistic Regression	0.39	0.40	0.76	0.78	0.71	0.72	0.81	0.84	0.73	0.73	0.62	0.64
Logistic RegressionCV	0.40	0.40	0.77	0.78	0.71	0.72	0.82	0.84	0.73	0.73	0.63	0.64
MLP	0.30	0.22	0.72	0.68	0.66	0.62	0.73	0.68	0.71	0.68	0.56	0.52
Multinomial NB	0.28	0.42	0.72	0.76	0.66	0.72	0.76	0.74	0.69	0.78	0.57	0.61
Passive Aggressive	0.26	0.21	0.65	0.19	0.62	0.47	0.68	0.11	0.72	1.00	0.50	0.11
QDA	0.07	-0.15	0.72	0.72	0.59	0.56	0.91	0.95	0.60	0.58	0.57	0.56
Random Forest	0.49	0.58	0.80	0.84	0.75	0.78	0.82	1.00	0.78	0.73	0.67	0.73
Ridge Classifier	0.36	0.40	0.75	0.78	0.69	0.72	0.79	0.84	0.72	0.73	0.60	0.64
Ridge ClassifierCV	0.41	0.40	0.77	0.78	0.72	0.72	0.83	0.84	0.73	0.73	0.63	0.64
Stochastic GDescent	0.29	0.22	0.72	0.76	0.66	0.62	0.76	1.00	0.69	0.61	0.56	0.61
SVC	0.39	-0.01	0.78	0.63	0.71	0.53	0.86	0.68	0.71	0.59	0.64	0.46
XGBoost	0.40	0.33	0.76	0.76	0.71	0.69	0.76	0.84	0.75	0.70	0.61	0.62

Table 11.C.4: Individual model evalutation metrics for pyrazinamide resistance prediction: all features
Data with 176 *pncA* features for 424 SAVs was split into train and test sets using the scaling law principle, and model performance assessed using a stratified 10-fold cross validation method. Abbreviations used: CV: cross validation, LDA: linear discriminant analysis, MLP: multilayer perceptron, NB: naive Bayes, QDA: quadratic discriminant analysis, SVC: support vector classification, MCC: Matthews correlation coefficient, JCC: jaccard similarity coefficient, SAV: single amino acid variation.

Model Name	MCC Train	MCC Test	F1 Train	F1 Test	Accuracy Train	Accuracy Test	Recall Train	Recall Test	Precision Train	Precision Test	JCC Train	JCC Test
AdaBoost Classifier	0.41	0.44	0.56	0.60	0.77	0.78	0.52	0.56	0.62	0.64	0.40	0.42
Bagging Classifier	0.37	0.53	0.51	0.65	0.76	0.81	0.42	0.60	0.64	0.71	0.34	0.48
Complement NB	0.25	0.25	0.51	0.51	0.64	0.64	0.64	0.64	0.42	0.42	0.34	0.34
Decision Tree	0.21	0.30	0.45	0.53	0.67	0.67	0.46	0.64	0.44	0.46	0.29	0.36
Dummy Classifier	0.00	0.00	0.00	0.00	0.71	0.71	0.00	0.00	0.00	0.00	0.00	0.00
Extra Tree	0.30	0.23	0.49	0.45	0.71	0.68	0.49	0.44	0.51	0.46	0.33	0.29
Extra Trees	0.37	0.41	0.50	0.57	0.76	0.76	0.42	0.52	0.64	0.62	0.34	0.39
Gaussian NB	0.31	0.30	0.52	0.51	0.71	0.71	0.53	0.52	0.51	0.50	0.35	0.34
Gaussian Process	0.26	0.26	0.33	0.38	0.73	0.73	0.23	0.28	0.64	0.58	0.20	0.23
Gradient Boosting	0.43	0.53	0.55	0.65	0.78	0.81	0.47	0.60	0.68	0.71	0.38	0.48
K-Nearest Neighbors	0.23	0.23	0.38	0.42	0.72	0.71	0.30	0.36	0.53	0.50	0.24	0.26
LDA	0.35	0.43	0.51	0.60	0.75	0.76	0.45	0.60	0.59	0.60	0.34	0.43
Logistic Regression	0.40	0.47	0.51	0.61	0.77	0.79	0.42	0.56	0.68	0.67	0.35	0.44
Logistic RegressionCV	0.37	0.48	0.49	0.63	0.76	0.79	0.40	0.60	0.66	0.65	0.33	0.45
MLP	0.39	0.44	0.54	0.62	0.75	0.75	0.51	0.68	0.61	0.57	0.37	0.45
Multinomial NB	0.31	0.23	0.50	0.43	0.72	0.69	0.49	0.40	0.52	0.48	0.34	0.28
Passive Aggressive	0.31	0.32	0.42	0.36	0.71	0.75	0.38	0.24	0.67	0.75	0.27	0.22
QDA	0.11	-0.06	0.46	0.43	0.37	0.32	0.94	0.88	0.31	0.29	0.30	0.28
Random Forest	0.38	0.55	0.47	0.65	0.77	0.82	0.36	0.56	0.71	0.78	0.31	0.48
Ridge Classifier	0.37	0.50	0.50	0.64	0.76	0.80	0.41	0.60	0.65	0.68	0.33	0.47
Ridge ClassifierCV	0.38	0.52	0.48	0.62	0.77	0.81	0.37	0.52	0.70	0.76	0.32	0.45
Stochastic GDescent	0.31	0.36	0.43	0.58	0.72	0.64	0.45	0.84	0.55	0.44	0.29	0.40
SVC	0.31	0.37	0.34	0.44	0.75	0.76	0.22	0.32	0.77	0.73	0.21	0.29
XGBoost	0.45	0.41	0.59	0.57	0.79	0.76	0.52	0.52	0.68	0.62	0.42	0.39

Table 11.C.5: Individual model evalutation metrics for rifampicin resistance prediction: all features
Data with 180 *rpoB* features for 1132 SAVs was split into train and test sets using the scaling law principle, and model performance assessed using a stratified 10-fold cross validation method. Abbreviations used: CV: cross validation, LDA: linear discriminant analysis, MLP: multilayer perceptron, NB: naive Bayes, QDA: quadratic discriminant analysis, SVC: support vector classification, MCC: Matthews correlation coefficient, JCC: jaccard similarity coefficient, SAV: single amino acid variation.

Model Name	MCC Train	MCC Test	F1 Train	F1 Test	Accuracy Train	Accuracy Test	Recall Train	Recall Test	Precision Train	Precision Test	JCC Train	JCC Test	Features Selected (n=9)
AdaBoost Classifier	0.31	0.27	0.34	0.40	0.86	0.72	0.26	0.60	0.56	0.30	0.22	0.25	
Bagging Classifier	0.36	0.32	0.38	0.42	0.87	0.83	0.27	0.40	0.66	0.44	0.24	0.27	
Complement NB	0.27	0.30	0.39	0.41	0.72	0.65	0.62	0.80	0.29	0.28	0.25	0.26	
Decision Tree	0.27	0.35	0.38	0.45	0.80	0.82	0.41	0.50	0.36	0.42	0.24	0.29	
Dummy Classifier	0.00	0.00	0.00	0.00	0.85	0.85	0.00	0.00	0.00	0.00	0.00	0.00	
Extra Tree	0.23	0.29	0.34	0.41	0.80	0.74	0.37	0.60	0.33	0.32	0.21	0.26	
Extra Trees	0.38	0.32	0.37	0.42	0.87	0.83	0.26	0.40	0.76	0.44	0.23	0.27	
Gaussian NB	0.24	0.49	0.31	0.57	0.84	0.86	0.27	0.60	0.40	0.55	0.19	0.40	
Gaussian Process	0.13	0.42	0.10	0.33	0.86	0.88	0.06	0.20	0.45	1.00	0.06	0.20	ConSurf,
Gradient Boosting	0.35	0.45	0.37	0.53	0.87	0.86	0.28	0.50	0.63	0.56	0.24	0.36	SNAP2,
K-Nearest Neighbors	0.21	0.52	0.25	0.46	0.85	0.89	0.18	0.30	0.45	1.00	0.15	0.30	PROVEAN,
LDA	0.14	0.25	0.13	0.29	0.85	0.85	0.08	0.20	0.42	0.50	0.07	0.17	DeepDDG,
Logistic Regression	0.11	0.29	0.08	0.18	0.85	0.86	0.04	0.10	0.40	1.00	0.04	0.10	MAF,
Logistic RegressionCV	0.12	0.42	0.10	0.33	0.85	0.88	0.06	0.20	0.40	1.00	0.06	0.20	SAV frequency,
MLP	0.23	0.36	0.23	0.40	0.86	0.86	0.15	0.30	0.50	0.60	0.14	0.25	Residue depth,
Multinomial NB	0.10	0.42	0.06	0.33	0.86	0.88	0.03	0.20	0.40	1.00	0.03	0.20	EMB distance,
Passive Aggressive	0.17	0.26	0.17	0.35	0.82	0.83	0.20	0.30	0.40	0.43	0.10	0.21	Mutation site
QDA	0.17	0.28	0.20	0.36	0.84	0.46	0.15	1.00	0.42	0.22	0.11	0.22	frequency in dataset
Random Forest	0.38	0.26	0.35	0.35	0.88	0.83	0.24	0.30	0.82	0.43	0.22	0.21	
Ridge Classifier	0.10	0.00	0.06	0.00	0.85	0.85	0.03	0.00	0.40	0.00	0.03	0.00	
Ridge ClassifierCV	0.09	-0.05	0.06	0.00	0.85	0.83	0.03	0.00	0.35	0.00	0.03	0.00	
SVC	0.11	0.00	0.06	0.00	0.86	0.85	0.03	0.00	0.40	0.00	0.03	0.00	
Stochastic GDescent	0.09	-0.05	0.06	0.00	0.85	0.83	0.03	0.00	0.35	0.00	0.03	0.00	
XGBoost	0.41	0.40	0.45	0.47	0.87	0.86	0.35	0.40	0.66	0.57	0.29	0.31	

Table 11.C.6: Individual model evalutation metrics for ethambutol resistance prediction: post feature selection
Data with 178 *embB* features for 858 SAVs was split into train and test sets using the scaling law principle followed by the Boruta feature selection process[11] which identified 9 features optimised for the MCC score. Abbreviations used: CV: cross validation, LDA: linear discriminant analysis, MLP: multilayer perceptron, NB: naive Bayes, QDA: quadratic discriminant analysis, SVC: support vector classification, MAF: minor allele frequency, MCC: Matthews correlation coefficient, JCC: jaccard similarity coefficient, SAV: single amino acid variation, EMB: ethambutol.

Model Name	MCC Train	MCC Test	F1 Train	F1 Test	Accuracy Train	Accuracy Test	Recall Train	Recall Test	Precision Train	Precision Test	JCC Train	JCC Test	Features Selected (N/A)
AdaBoost Classifier	-0.03	0.37	0.00	0.40	0.91	0.92	0.00	0.33	0.00	0.50	0.00	0.25	
Bagging Classifier	0.03	0.56	0.04	0.50	0.93	0.95	0.03	0.33	0.05	1.00	0.02	0.33	
Complement NB	0.04	0.08	0.15	0.17	0.53	0.50	0.54	0.67	0.08	0.10	0.08	0.09	
Decision Tree	0.08	0.56	0.12	0.50	0.89	0.95	0.14	0.33	0.13	1.00	0.07	0.33	
Dummy Classifier	0.00	0.00	0.00	0.00	0.93	0.92	0.00	0.00	0.00	0.00	0.00	0.00	
Extra Tree	0.04	-0.07	0.09	0.00	0.88	0.88	0.12	0.00	0.08	0.00	0.06	0.00	
Extra Trees	0.04	0.00	0.05	0.00	0.92	0.92	0.03	0.00	0.10	0.00	0.03	0.00	
Gaussian NB	0.13	0.37	0.14	0.40	0.91	0.92	0.09	0.33	0.30	0.50	0.09	0.25	
Gaussian Process	0.00	0.00	0.00	0.00	0.93	0.92	0.00	0.00	0.00	0.00	0.00	0.00	
Gradient Boosting	0.06	0.37	0.08	0.40	0.92	0.92	0.09	0.33	0.07	0.50	0.05	0.25	
K-Nearest Neighbors	-0.00	0.00	0.00	0.00	0.93	0.92	0.00	0.00	0.00	0.00	0.00	0.00	
LDA	0.00	0.00	0.00	0.00	0.93	0.92	0.00	0.00	0.00	0.00	0.00	0.00	
Logistic Regression	0.00	0.00	0.00	0.00	0.93	0.92	0.00	0.00	0.00	0.00	0.00	0.00	
Logistic RegressionCV	0.00	0.00	0.00	0.00	0.93	0.92	0.00	0.00	0.00	0.00	0.00	0.00	
MLP	0.00	0.00	0.00	0.00	0.93	0.92	0.00	0.00	0.00	0.00	0.00	0.00	
Multinomial NB	0.00	0.00	0.00	0.00	0.93	0.92	0.00	0.00	0.00	0.00	0.00	0.00	
Passive Aggressive	0.01	0.00	0.02	0.00	0.85	0.92	0.10	0.00	0.01	0.00	0.01	0.00	
QDA	-0.01	0.00	0.00	0.00	0.92	0.92	0.00	0.00	0.00	0.00	0.00	0.00	
Random Forest	0.00	0.00	0.00	0.00	0.93	0.92	0.00	0.00	0.00	0.00	0.00	0.00	
Ridge Classifier	0.00	0.00	0.00	0.00	0.93	0.92	0.00	0.00	0.00	0.00	0.00	0.00	
Ridge ClassifierCV	0.00	0.00	0.00	0.00	0.93	0.92	0.00	0.00	0.00	0.00	0.00	0.00	
SVC	0.00	0.00	0.00	0.00	0.93	0.92	0.00	0.00	0.00	0.00	0.00	0.00	
Stochastic GDescent	0.00	0.00	0.00	0.00	0.93	0.92	0.00	0.00	0.00	0.00	0.00	0.00	
XGBoost	0.06	0.00	0.08	0.00	0.91	0.92	0.07	0.00	0.13	0.00	0.05	0.00	

Table 11.C.7: Individual model evalutation metrics for streptomycin resistance prediction: post feature selection
Data with 178 *gidB* features for 531 SAVs was split into train and test sets using the scaling law principle followed by the Boruta feature selection process[11] which identified 2 features optimised for the MCC score. However, since most MCC scores were ∼0.0, results from this analysis were inconclusive and not considered further. Abbreviations used: CV: cross validation, LDA: linear discriminant analysis, MLP: multilayer perceptron, NB: naive Bayes, QDA: quadratic discriminant analysis, SVC: support vector classification, MCC: Matthews correlation coefficient, JCC: jaccard similarity coefficient, SAV: single amino acid variation.

Model Name	MCC Train	MCC Test	F1 Train	F1 Test	Accuracy Train	Accuracy Test	Recall Train	Recall Test	Precision Train	Precision Test	JCC Train	JCC Test	Features Selected (n=11)
AdaBoost Classifier	0.29	0.21	0.60	0.60	0.65	0.60	0.58	0.68	0.62	0.54	0.43	0.43	
Bagging Classifier	0.34	0.28	0.63	0.61	0.68	0.65	0.61	0.61	0.65	0.61	0.46	0.44	
Complement NB	0.34	0.22	0.66	0.62	0.66	0.60	0.74	0.71	0.60	0.54	0.50	0.44	
Decision Tree	0.28	-0.03	0.60	0.39	0.64	0.50	0.60	0.36	0.60	0.43	0.43	0.24	
Dummy Classifier	0.00	0.00	0.00	0.00	0.55	0.55	0.00	0.00	0.00	0.00	0.00	0.00	
Extra Tree	0.16	0.15	0.53	0.54	0.58	0.58	0.53	0.54	0.54	0.54	0.37	0.37	
Extra Trees	0.32	0.32	0.61	0.64	0.67	0.66	0.59	0.68	0.64	0.61	0.45	0.48	
Gaussian NB	0.33	0.25	0.65	0.63	0.66	0.61	0.69	0.71	0.61	0.56	0.48	0.45	ConSurf, SNAP2, PROVEAN, DeepDDG, FoldX, PPI, Residue depth, Average stability, INH distance, AAindex2: BENS940104[13] GIAG010101[14]
Gaussian Process	0.36	0.29	0.64	0.63	0.68	0.65	0.64	0.68	0.65	0.59	0.48	0.46	
Gradient Boosting	0.33	0.12	0.62	0.51	0.67	0.56	0.61	0.50	0.64	0.52	0.45	0.34	
K-Nearest Neighbors	0.24	0.12	0.58	0.53	0.63	0.56	0.57	0.54	0.59	0.52	0.41	0.36	
LDA	0.37	0.42	0.65	0.71	0.69	0.69	0.66	0.82	0.66	0.62	0.49	0.55	
Logistic Regression	0.37	0.32	0.65	0.64	0.69	0.66	0.65	0.68	0.65	0.61	0.49	0.48	
Logistic RegressionCV	0.35	0.42	0.65	0.71	0.68	0.69	0.65	0.82	0.64	0.62	0.48	0.55	
MLP	0.33	0.28	0.63	0.65	0.67	0.63	0.62	0.75	0.64	0.57	0.46	0.48	
Multinomial NB	0.28	0.25	0.54	0.57	0.65	0.63	0.46	0.54	0.66	0.60	0.37	0.39	
Passive Aggressive	0.15	0.19	0.43	0.63	0.56	0.56	0.54	0.82	0.43	0.51	0.31	0.46	
QDA	0.37	0.14	0.64	0.60	0.69	0.55	0.61	0.75	0.67	0.50	0.47	0.43	
Random Forest	0.33	0.26	0.63	0.61	0.67	0.63	0.63	0.64	0.64	0.58	0.46	0.44	
Ridge Classifier	0.37	0.42	0.66	0.71	0.69	0.69	0.66	0.82	0.65	0.62	0.49	0.55	
Ridge ClassifierCV	0.37	0.29	0.66	0.62	0.69	0.65	0.66	0.64	0.65	0.60	0.49	0.45	
SVC	0.35	0.32	0.63	0.63	0.68	0.66	0.61	0.64	0.65	0.62	0.46	0.46	
Stochastic GDescent	0.32	0.38	0.58	0.65	0.65	0.69	0.61	0.64	0.66	0.67	0.42	0.49	
XGBoost	0.34	0.19	0.63	0.58	0.67	0.60	0.62	0.61	0.64	0.55	0.46	0.40	

Table 11.C.8: Individual model evalutation metrics for isoniazid resistance prediction: post feature selection
Data with 178 *katG* features for 817 SAVs was split into train and test sets using the scaling law principle followed by the Boruta feature selection process[11] which identified 11 features optimised for the MCC score. Abbreviations used: CV: cross validation, LDA: linear discriminant analysis, MLP: multilayer perceptron, NB: naive Bayes, QDA: quadratic discriminant analysis, SVC: support vector classification, MCC: Matthews correlation coefficient, JCC: jaccard similarity coefficient, SAV: single amino acid variation, INH: isoniazid.

Model Name	MCC Train	MCC Test	F1 Train	F1 Test	Accuracy Train	Accuracy Test	Recall Train	Recall Test	Precision Train	Precision Test	JCC Train	JCC Test	Features Selected (n=14)
AdaBoost Classifier	0.44	0.40	0.77	0.79	0.73	0.72	0.77	0.89	0.77	0.71	0.63	0.65	
Bagging Classifier	0.52	0.25	0.80	0.76	0.77	0.66	0.79	0.89	0.81	0.65	0.67	0.61	
Complement NB	0.43	0.40	0.77	0.79	0.72	0.72	0.77	0.89	0.77	0.71	0.62	0.65	
Decision Tree	0.35	0.13	0.72	0.68	0.68	0.59	0.71	0.74	0.75	0.64	0.57	0.52	
Dummy Classifier	0.00	0.00	0.74	0.75	0.59	0.59	1.00	1.00	0.59	0.59	0.59	0.59	
Extra Tree	0.29	0.21	0.68	0.70	0.65	0.62	0.67	0.74	0.72	0.67	0.53	0.54	ConSurf, SNAP2, PROVEAN, DeepDDG, FoldX, MAF, SAV frequency, Residue depth, PZA distance, RSA, Mutation site frequency in the dataset, Mutation site frequency in gene, AAindex2: DOSZ010103,[15]
Extra Trees	0.47	0.46	0.79	0.81	0.74	0.72	0.80	1.00	0.78	0.68	0.65	0.68	
Gaussian NB	0.48	0.64	0.76	0.86	0.74	0.81	0.72	1.00	0.81	0.76	0.62	0.76	
Gaussian Process	0.45	0.52	0.79	0.83	0.73	0.75	0.85	1.00	0.74	0.70	0.66	0.70	
Gradient Boosting	0.53	0.33	0.80	0.77	0.77	0.69	0.80	0.89	0.81	0.68	0.67	0.63	
K-Nearest Neighbors	0.36	0.58	0.75	0.84	0.69	0.78	0.78	1.00	0.72	0.73	0.60	0.73	
LDA	0.45	0.52	0.79	0.83	0.73	0.75	0.85	1.00	0.74	0.70	0.66	0.70	
Logistic Regression	0.46	0.52	0.79	0.83	0.74	0.75	0.84	1.00	0.75	0.70	0.66	0.70	
Logistic RegressionCV	0.45	0.46	0.79	0.81	0.74	0.72	0.85	1.00	0.74	0.68	0.66	0.68	
MLP	0.50	0.52	0.80	0.83	0.76	0.75	0.83	1.00	0.78	0.70	0.67	0.70	
Multinomial NB	0.44	0.35	0.80	0.78	0.73	0.69	0.91	0.95	0.71	0.67	0.66	0.64	
Passive Aggressive	0.29	0.37	0.77	0.72	0.67	0.69	0.95	0.68	0.65	0.76	0.63	0.57	
QDA	0.45	0.42	0.75	0.80	0.72	0.72	0.71	0.95	0.80	0.69	0.61	0.67	
Random Forest	0.51	0.46	0.80	0.81	0.76	0.72	0.80	1.00	0.80	0.68	0.66	0.68	
Ridge Classifier	0.45	0.52	0.79	0.83	0.73	0.75	0.85	1.00	0.74	0.70	0.66	0.70	
Ridge ClassifierCV	0.42	0.52	0.78	0.83	0.72	0.75	0.86	1.00	0.73	0.70	0.65	0.70	
SVC	0.42	0.52	0.79	0.83	0.72	0.75	0.87	1.00	0.72	0.70	0.65	0.70	
Stochastic GDescent	0.38	0.39	0.65	0.79	0.67	0.69	0.68	1.00	0.79	0.66	0.52	0.66	
XGBoost	0.51	0.33	0.79	0.77	0.76	0.69	0.80	0.89	0.80	0.68	0.66	0.63	

Table 11.C.9: Individual model evalutation metrics for pyrazinamide resistance prediction: post feature selection
Data with all 176 *pncA* features for 424 SAVs was split into train and test sets using the scaling law principle followed by 14 features optimised for the MCC score. Abbreviations used: CV: cross validation, LDA: linear discriminant analysis, MLP: multilayer perceptron, NB: naive Bayes, QDA: quadratic discriminant analysis, SVC: support vector classification, MCC: Matthews correlation coefficient, JCC: jaccard similarity coefficient, RSA: relative surface area, SAV: single amino acid variation, PZA: pyrazinamide.

Model Name	MCC Train	MCC Test	F1 Train	F1 Test	Accuracy Train	Accuracy Test	Recall Train	Recall Test	Precision Train	Precision Test	JCC Train	JCC Test	Features Selected (n=22)
AdaBoost Classifier	0.38	0.58	0.53	0.67	0.76	0.84	0.47	0.56	0.61	0.82	0.36	0.50	
Bagging Classifier	0.43	0.55	0.55	0.65	0.78	0.82	0.47	0.56	0.68	0.78	0.39	0.48	
Complement NB	0.39	0.45	0.55	0.62	0.76	0.76	0.51	0.64	0.61	0.59	0.38	0.44	ConSurf,
Decision Tree	0.28	0.30	0.49	0.53	0.70	0.67	0.49	0.64	0.49	0.46	0.32	0.36	SNAP2,
Dummy Classifier	0.00	0.00	0.00	0.00	0.71	0.71	0.00	0.00	0.00	0.00	0.00	0.00	PROVEAN,
Extra Tree	0.25	0.07	0.46	0.39	0.69	0.56	0.45	0.48	0.47	0.33	0.30	0.24	mCSM-DUET,
Extra Trees	0.44	0.48	0.55	0.63	0.79	0.79	0.45	0.60	0.72	0.65	0.38	0.45	DeepDDG,
Gaussian NB	0.33	0.23	0.45	0.50	0.75	0.62	0.35	0.64	0.64	0.41	0.30	0.33	Four FoldX interactions,
Gaussian Process	0.42	0.45	0.52	0.57	0.78	0.79	0.42	0.48	0.71	0.71	0.36	0.40	RFP distance,
Gradient Boosting	0.41	0.49	0.54	0.62	0.77	0.80	0.46	0.56	0.67	0.70	0.37	0.45	RFP affinity,
K-Nearest Neighbors	0.38	0.41	0.53	0.57	0.76	0.76	0.46	0.52	0.62	0.62	0.36	0.39	NA distance,
LDA	0.42	0.34	0.53	0.46	0.78	0.75	0.43	0.36	0.71	0.64	0.37	0.30	NA affinity,
Logistic Regression	0.40	0.39	0.51	0.52	0.77	0.76	0.41	0.44	0.69	0.65	0.35	0.35	PPI distance,
Logistic RegressionCV	0.36	0.48	0.49	0.59	0.76	0.80	0.40	0.48	0.65	0.75	0.33	0.41	RSA,
MLP	0.41	0.41	0.54	0.57	0.77	0.76	0.46	0.52	0.67	0.62	0.37	0.39	Mutation site frequency in
Multinomial NB	0.32	0.41	0.30	0.42	0.75	0.78	0.19	0.28	0.84	0.88	0.18	0.27	gene,
Passive Aggressive	0.31	0.40	0.48	0.60	0.66	0.67	0.59	0.84	0.55	0.47	0.33	0.43	Mutation site
QDA	0.37	0.19	0.49	0.49	0.76	0.49	0.40	0.84	0.66	0.35	0.33	0.33	frequency in
Random Forest	0.44	0.55	0.55	0.67	0.79	0.82	0.46	0.60	0.71	0.75	0.39	0.50	dataset,
Ridge Classifier	0.41	0.30	0.50	0.42	0.78	0.74	0.39	0.32	0.72	0.62	0.34	0.27	Lineage,
Ridge ClassifierCV	0.40	0.30	0.50	0.42	0.78	0.74	0.39	0.32	0.71	0.62	0.34	0.27	Active site,
SVC	0.34	0.37	0.41	0.44	0.76	0.76	0.30	0.32	0.70	0.73	0.27	0.29	Non-active site,
Stochastic GDescent	0.35	0.30	0.42	0.42	0.76	0.74	0.36	0.32	0.71	0.62	0.29	0.27	AAindex2: MIYT790101[17]
XGBoost	0.45	0.49	0.59	0.60	0.78	0.80	0.54	0.52	0.65	0.72	0.42	0.43	

Table 11.C.10: Individual model evalutation metrics for rifampicin resistance prediction: post feature selection
Data with 180 *rpoB* features for 1132 SAVs was split into train and test sets using the scaling law principle followed by the Boruta feature selection process[11] which identified 22 features optimised for the MCC score. Abbreviations used: CV: cross validation, LDA: linear discriminant analysis, MLP: multilayer perceptron, NB: naive Bayes, QDA: quadratic discriminant analysis, SVC: support vector classification, NA: nucleic acid, PPI: protein-protein interface, MCC: Matthews correlation coefficient, JCC: jaccard similarity coefficient, RSA: relative surface area, SAV: single amino acid variation, RFP: rifampicin.

Model Name	MCC Train	MCC Test	F1 Train	F1 Test	Accuracy Train	Accuracy Test	Recall Train	Recall Test	Precision Train	Precision Test	JCC Train	JCC Test
AdaBoost Classifier	0.45	0.23	0.62	0.33	0.76	0.82	0.58	0.31	0.68	0.36	0.45	0.20
Bagging Classifier	0.47	0.29	0.63	0.39	0.77	0.82	0.57	0.39	0.70	0.40	0.46	0.24
Complement NB	0.18	0.09	0.51	0.27	0.57	0.60	0.67	0.50	0.42	0.19	0.35	0.16
Decision Tree	0.32	0.14	0.55	0.29	0.69	0.70	0.57	0.43	0.54	0.22	0.38	0.17
Dummy Classifier	0.00	0.00	0.00	0.00	0.66	0.85	0.00	0.00	0.00	0.00	0.00	0.00
Extra Tree	0.25	0.06	0.50	0.24	0.66	0.68	0.51	0.33	0.50	0.18	0.34	0.13
Extra Trees	0.41	0.26	0.57	0.35	0.75	0.83	0.49	0.31	0.69	0.40	0.40	0.21
Gaussian NB	0.27	0.15	0.55	0.31	0.63	0.67	0.67	0.51	0.47	0.22	0.38	0.19
Gaussian Process	0.31	0.14	0.46	0.24	0.72	0.81	0.36	0.20	0.65	0.30	0.30	0.14
Gradient Boosting	0.49	0.28	0.64	0.38	0.78	0.83	0.59	0.36	0.71	0.41	0.47	0.24
K-Nearest Neighbors	0.21	0.09	0.44	0.24	0.67	0.73	0.38	0.29	0.51	0.21	0.28	0.14
LDA	0.44	0.24	0.62	0.35	0.76	0.81	0.57	0.34	0.67	0.36	0.45	0.21
Logistic Regression	0.44	0.24	0.61	0.33	0.76	0.82	0.56	0.30	0.67	0.38	0.44	0.20
Logistic RegressionCV	0.00	0.00	0.00	0.00	0.66	0.85	0.00	0.00	0.00	0.00	0.00	0.00
MLP	0.45	0.17	0.62	0.29	0.75	0.79	0.60	0.30	0.67	0.29	0.45	0.17
Multinomial NB	0.21	0.07	0.49	0.24	0.64	0.72	0.52	0.29	0.47	0.20	0.33	0.13
Passive Aggressive	0.32	0.03	0.46	0.26	0.70	0.15	0.46	1.00	0.63	0.15	0.32	0.15
QDA	0.09	0.03	0.51	0.26	0.37	0.18	0.98	0.98	0.35	0.15	0.35	0.15
Random Forest	0.46	0.27	0.60	0.34	0.77	0.85	0.50	0.26	0.73	0.47	0.42	0.20
Ridge Classifier	0.45	0.25	0.61	0.35	0.76	0.83	0.55	0.31	0.69	0.39	0.44	0.21
Ridge ClassifierCV	0.44	0.24	0.61	0.34	0.76	0.83	0.55	0.30	0.68	0.39	0.44	0.20
SVC	0.43	0.26	0.58	0.36	0.76	0.82	0.50	0.33	0.70	0.39	0.41	0.22
Stochastic GDescent	0.40	0.24	0.57	0.35	0.72	0.82	0.59	0.33	0.64	0.37	0.40	0.21
XGBoost	0.44	0.27	0.61	0.38	0.76	0.81	0.56	0.39	0.67	0.37	0.44	0.23

Table 11.C.11: Combined model evaluation metrics for ethambutol resistance prediction: all features
Data comprised of 178 features, with the training set consisting of 2904 SAVs combined from four genes (*gidB, katG, pncA*, and *rpoB*) based on the 'leave-one-gene-out' approach to test predictions on the 'left out' (i.e. test) gene. The test set comprised of 858 SAVs from *embB*. Train-test data split was undertaken using the scaling law principle, and model performance assessed using a stratified 10-fold cross validation method. Abbreviations used: CV: cross validation, LDA: linear discriminant analysis, MLP: multilayer perceptron, NB: naive Bayes, QDA: quadratic discriminant analysis, SVC: support vector classification, MCC: Matthews correlation coefficient, JCC: jaccard similarity coefficient, SAV: single amino acid variation.

Model Name	MCC Train	MCC Test	F1 Train	F1 Test	Accuracy Train	Accuracy Test	Recall Train	Recall Test	Precision Train	Precision Test	JCC Train	JCC Test
AdaBoost Classifier	0.45	0.09	0.61	0.17	0.77	0.67	0.56	0.47	0.68	0.10	0.44	0.09
Bagging Classifier	0.45	0.05	0.61	0.15	0.76	0.61	0.55	0.47	0.68	0.09	0.43	0.08
Complement NB	0.24	0.07	0.54	0.15	0.61	0.36	0.68	0.79	0.44	0.08	0.37	0.08
Decision Tree	0.27	0.03	0.52	0.14	0.67	0.49	0.52	0.58	0.51	0.08	0.35	0.08
Dummy Classifier	0.00	0.00	0.00	0.00	0.67	0.93	0.00	0.00	0.00	0.00	0.00	0.00
Extra Tree	0.23	-0.03	0.48	0.10	0.66	0.57	0.48	0.34	0.49	0.06	0.32	0.05
Extra Trees	0.40	0.01	0.56	0.12	0.75	0.63	0.48	0.37	0.67	0.07	0.39	0.07
Gaussian NB	0.30	0.04	0.56	0.14	0.66	0.24	0.64	0.87	0.50	0.08	0.39	0.08
Gaussian Process	0.35	-0.00	0.49	0.12	0.73	0.66	0.39	0.32	0.67	0.07	0.33	0.06
Gradient Boosting	0.48	-0.03	0.63	0.10	0.78	0.60	0.56	0.32	0.72	0.06	0.46	0.05
K-Nearest Neighbors	0.25	-0.02	0.45	0.11	0.69	0.65	0.39	0.29	0.55	0.07	0.29	0.06
LDA	0.43	-0.07	0.60	0.02	0.76	0.82	0.54	0.03	0.67	0.02	0.43	0.01
Logistic Regression	0.45	-0.02	0.60	0.07	0.76	0.81	0.55	0.11	0.68	0.06	0.43	0.04
Logistic RegressionCV	0.41	-0.05	0.57	0.07	0.75	0.74	0.49	0.13	0.68	0.05	0.40	0.04
MLP	0.42	0.01	0.58	0.11	0.75	0.76	0.53	0.21	0.66	0.07	0.41	0.06
Multinomial NB	0.27	0.04	0.53	0.14	0.67	0.45	0.55	0.63	0.50	0.08	0.36	0.08
Passive Aggressive	0.29	-0.04	0.38	0.00	0.67	0.91	0.38	0.00	0.70	0.00	0.25	0.00
QDA	0.09	0.03	0.51	0.13	0.37	0.08	0.98	1.00	0.34	0.07	0.34	0.07
Random Forest	0.45	-0.01	0.59	0.11	0.77	0.62	0.51	0.34	0.72	0.07	0.42	0.06
Ridge Classifier	0.44	-0.03	0.59	0.06	0.76	0.82	0.53	0.08	0.68	0.05	0.42	0.03
Ridge ClassifierCV	0.43	-0.04	0.59	0.07	0.76	0.76	0.52	0.13	0.68	0.05	0.42	0.04
SVC	0.42	0.00	0.56	0.12	0.76	0.69	0.47	0.29	0.70	0.07	0.39	0.06
Stochastic GDescent	0.31	-0.01	0.40	0.09	0.69	0.77	0.39	0.16	0.72	0.06	0.26	0.05
XGBoost	0.44	0.03	0.60	0.13	0.76	0.61	0.55	0.42	0.67	0.08	0.43	0.07

Table 11.C.12: Combined model evaluation metrics for streptomycin resistance prediction: all features
Data comprised of 176 features, with the training set consisting of 3231 SAVs combined from four genes (*embB, katG, pncA*, and *rpoB*) based on the 'leave-one-gene-out' approach to test predictions on the 'left out' (i.e. test) gene. The test set comprised of 531 SAVs from *gidB*. Train-test data split was undertaken using the scaling law principle, and model performance assessed using a stratified 10-fold cross validation method. Abbreviations used: CV: cross validation, LDA: linear discriminant analysis, MLP: multilayer perceptron, NB: naive Bayes, QDA: quadratic discriminant analysis, SVC: support vector classification, MCC: Matthews correlation coefficient, JCC: jaccard similarity coefficient, SAV: single amino acid variation.

Model Name	MCC Train	MCC Test	F1 Train	F1 Test	Accuracy Train	Accuracy Test	Recall Train	Recall Test	Precision Train	Precision Test	JCC Train	JCC Test
AdaBoost Classifier	0.46	0.22	0.57	0.38	0.81	0.62	0.49	0.27	0.68	0.70	0.40	0.24
Bagging Classifier	0.48	0.19	0.57	0.30	0.82	0.60	0.48	0.19	0.71	0.71	0.40	0.18
Complement NB	0.20	0.17	0.45	0.59	0.59	0.58	0.66	0.66	0.34	0.53	0.29	0.41
Decision Tree	0.32	0.22	0.50	0.47	0.74	0.62	0.51	0.37	0.48	0.64	0.33	0.31
Dummy Classifier	0.00	0.00	0.00	0.00	0.75	0.55	0.00	0.00	0.00	0.00	0.00	0.00
Extra Tree	0.21	0.07	0.41	0.38	0.70	0.55	0.42	0.30	0.41	0.51	0.26	0.23
Extra Trees	0.42	0.17	0.50	0.27	0.81	0.59	0.38	0.17	0.72	0.68	0.33	0.16
Gaussian NB	0.25	0.24	0.48	0.57	0.65	0.63	0.63	0.54	0.39	0.60	0.31	0.39
Gaussian Process	0.29	0.12	0.31	0.17	0.78	0.57	0.20	0.10	0.71	0.69	0.19	0.10
Gradient Boosting	0.50	0.20	0.59	0.27	0.83	0.60	0.50	0.17	0.73	0.74	0.42	0.16
K-Nearest Neighbors	0.21	0.10	0.34	0.28	0.74	0.57	0.27	0.18	0.49	0.57	0.21	0.16
LDA	0.46	0.22	0.56	0.38	0.81	0.61	0.49	0.26	0.67	0.69	0.39	0.23
Logistic Regression	0.46	0.22	0.56	0.37	0.81	0.61	0.47	0.25	0.69	0.71	0.39	0.23
Logistic RegressionCV	0.40	0.19	0.48	0.33	0.81	0.60	0.39	0.22	0.64	0.69	0.33	0.20
MLP	0.43	0.16	0.54	0.26	0.80	0.59	0.47	0.16	0.66	0.67	0.37	0.15
Multinomial NB	0.22	0.17	0.42	0.45	0.71	0.60	0.41	0.36	0.42	0.59	0.26	0.29
Passive Aggressive	0.32	0.30	0.38	0.60	0.73	0.65	0.42	0.59	0.67	0.62	0.25	0.43
QDA	0.07	0.14	0.41	0.63	0.29	0.49	0.98	0.98	0.26	0.47	0.26	0.46
Random Forest	0.47	0.17	0.53	0.13	0.82	0.58	0.41	0.07	0.78	0.90	0.37	0.07
Ridge Classifier	0.46	0.20	0.55	0.31	0.82	0.60	0.45	0.20	0.71	0.70	0.38	0.18
Ridge ClassifierCV	0.45	0.20	0.54	0.31	0.81	0.60	0.43	0.20	0.72	0.70	0.37	0.18
SVC	0.39	0.15	0.42	0.19	0.80	0.58	0.29	0.11	0.77	0.74	0.27	0.10
Stochastic GDescent	0.36	0.35	0.45	0.65	0.73	0.67	0.48	0.68	0.63	0.63	0.29	0.49
XGBoost	0.50	0.18	0.60	0.28	0.82	0.59	0.53	0.18	0.71	0.71	0.43	0.16

Table 11.C.13: Combined model evaluation metrics for isoniazid resistance prediction: all features
Data comprised of 176 features, with the training set consisting of 2945 SAVs combined from four genes (*embB, gidB, pncA*, and *rpoB*) based on the 'leave-one-gene-out' approach to test predictions on the 'left out' (i.e. test) gene. The test set comprised of 817 SAVs from *katG*. Train-test data split was undertaken using the scaling law principle, and model performance assessed using a stratified 10-fold cross validation method. Abbreviations used: CV: cross validation, LDA: linear discriminant analysis, MLP: multilayer perceptron, NB: naive Bayes, QDA: quadratic discriminant analysis, SVC: support vector classification, MCC: Matthews correlation coefficient, JCC: jaccard similarity coefficient, SAV: single amino acid variation.

Model Name	MCC Train	MCC Test	F1 Train	F1 Test	Accuracy Train	Accuracy Test	Recall Train	Recall Test	Precision Train	Precision Test	JCC Train	JCC Test
AdaBoost Classifier	0.38	0.23	0.50	0.68	0.78	0.62	0.42	0.66	0.62	0.69	0.33	0.51
Bagging Classifier	0.42	0.38	0.53	0.78	0.80	0.71	0.44	0.88	0.66	0.70	0.36	0.64
Complement NB	0.16	0.24	0.43	0.73	0.56	0.65	0.65	0.82	0.33	0.66	0.28	0.58
Decision Tree	0.29	0.17	0.47	0.69	0.72	0.61	0.48	0.73	0.47	0.65	0.31	0.52
Dummy Classifier	0.00	0.00	0.00	0.00	0.74	0.41	0.00	0.00	0.00	0.00	0.00	0.00
Extra Tree	0.15	0.13	0.38	0.57	0.67	0.55	0.38	0.50	0.37	0.66	0.23	0.40
Extra Trees	0.34	0.29	0.42	0.60	0.78	0.62	0.32	0.48	0.64	0.78	0.27	0.43
Gaussian NB	0.24	0.17	0.47	0.60	0.64	0.58	0.62	0.53	0.38	0.68	0.31	0.42
Gaussian Process	0.21	0.24	0.24	0.37	0.76	0.53	0.15	0.23	0.60	0.87	0.14	0.22
Gradient Boosting	0.43	0.39	0.53	0.77	0.80	0.71	0.43	0.83	0.69	0.72	0.36	0.63
K-Nearest Neighbors	0.17	0.20	0.32	0.50	0.73	0.55	0.25	0.38	0.45	0.74	0.19	0.33
LDA	0.39	0.31	0.50	0.73	0.79	0.67	0.42	0.74	0.64	0.71	0.34	0.57
Logistic Regression	0.39	0.34	0.49	0.74	0.79	0.68	0.39	0.78	0.66	0.71	0.32	0.59
Logistic RegressionCV	0.00	0.00	0.00	0.00	0.74	0.41	0.00	0.00	0.00	0.00	0.00	0.00
MLP	0.38	0.20	0.48	0.68	0.79	0.62	0.39	0.70	0.64	0.67	0.32	0.52
Multinomial NB	0.17	0.23	0.35	0.59	0.70	0.59	0.31	0.49	0.41	0.73	0.22	0.41
Passive Aggressive	0.26	0.20	0.35	0.48	0.66	0.55	0.46	0.35	0.61	0.75	0.22	0.32
QDA	0.06	0.06	0.42	0.74	0.29	0.59	0.97	0.99	0.26	0.59	0.26	0.59
Random Forest	0.36	0.37	0.42	0.70	0.79	0.68	0.30	0.63	0.72	0.78	0.27	0.53
Ridge Classifier	0.37	0.32	0.46	0.71	0.79	0.67	0.35	0.70	0.67	0.73	0.30	0.56
Ridge ClassifierCV	0.36	0.35	0.43	0.73	0.79	0.69	0.32	0.73	0.69	0.74	0.28	0.58
SVC	0.25	0.24	0.23	0.43	0.77	0.54	0.14	0.30	0.76	0.81	0.13	0.28
Stochastic GDescent	0.34	0.39	0.41	0.79	0.77	0.71	0.36	0.91	0.66	0.69	0.27	0.65
XGBoost	0.43	0.41	0.54	0.78	0.80	0.72	0.46	0.82	0.65	0.73	0.37	0.63

Table 11.C.14: Combined model evaluation metrics for pyrazinamide resistance prediction: all features
Data comprised of 176 features, with the training set consisting of 3338 SAVs combined from four genes (*embB, gidB, katG*, and *rpoB*) based on the 'leave-one-gene-out' approach to test predictions on the 'left out' (i.e. test) gene. The test set comprised of 424 SAVs from *pncA*. Train-test data split was undertaken using the scaling law principle, and model performance assessed using a stratified 10-fold cross validation method. Abbreviations used: CV: cross validation, LDA: linear discriminant analysis, MLP: multilayer perceptron, NB: naive Bayes, QDA: quadratic discriminant analysis, SVC: support vector classification, MCC: Matthews correlation coefficient, JCC: jaccard similarity coefficient, SAV: single amino acid variation.

Model Name	MCC Train	MCC Test	F1 Train	F1 Test	Accuracy Train	Accuracy Test	Recall Train	Recall Test	Precision Train	Precision Test	JCC Train	JCC Test
AdaBoost Classifier	0.45	0.29	0.60	0.42	0.78	0.74	0.55	0.32	0.66	0.60	0.43	0.26
Bagging Classifier	0.48	0.31	0.61	0.39	0.79	0.75	0.54	0.27	0.71	0.68	0.44	0.24
Complement NB	0.20	0.25	0.49	0.50	0.58	0.65	0.68	0.58	0.38	0.43	0.33	0.33
Decision Tree	0.32	0.19	0.53	0.43	0.71	0.67	0.54	0.43	0.52	0.42	0.36	0.27
Dummy Classifier	0.00	0.00	0.00	0.00	0.70	0.71	0.00	0.00	0.00	0.00	0.00	0.00
Extra Tree	0.20	0.12	0.44	0.35	0.66	0.66	0.45	0.31	0.44	0.39	0.29	0.21
Extra Trees	0.41	0.31	0.53	0.37	0.77	0.75	0.44	0.25	0.68	0.70	0.37	0.23
Gaussian NB	0.28	0.26	0.53	0.48	0.65	0.70	0.67	0.48	0.44	0.48	0.36	0.32
Gaussian Process	0.28	0.22	0.37	0.23	0.74	0.73	0.26	0.14	0.65	0.69	0.23	0.13
Gradient Boosting	0.50	0.35	0.62	0.43	0.80	0.76	0.55	0.30	0.72	0.71	0.45	0.27
K-Nearest Neighbors	0.22	0.20	0.41	0.32	0.70	0.72	0.34	0.23	0.50	0.53	0.26	0.19
LDA	0.42	0.29	0.58	0.40	0.77	0.74	0.53	0.30	0.63	0.62	0.41	0.25
Logistic Regression	0.45	0.36	0.59	0.46	0.78	0.76	0.54	0.35	0.66	0.68	0.42	0.30
Logistic RegressionCV	0.39	0.33	0.52	0.45	0.77	0.75	0.47	0.35	0.59	0.64	0.37	0.29
MLP	0.41	0.33	0.56	0.49	0.76	0.74	0.52	0.42	0.62	0.58	0.39	0.32
Multinomial NB	0.21	0.25	0.44	0.40	0.67	0.72	0.44	0.32	0.45	0.54	0.29	0.25
Passive Aggressive	0.30	0.30	0.46	0.46	0.65	0.73	0.59	0.40	0.56	0.56	0.31	0.30
QDA	0.06	0.09	0.46	0.46	0.32	0.35	0.98	0.96	0.30	0.30	0.30	0.30
Random Forest	0.45	0.35	0.56	0.37	0.79	0.76	0.45	0.24	0.73	0.80	0.39	0.23
Ridge Classifier	0.43	0.33	0.57	0.41	0.78	0.76	0.50	0.29	0.67	0.69	0.40	0.26
Ridge ClassifierCV	0.43	0.33	0.57	0.41	0.78	0.76	0.50	0.29	0.67	0.69	0.40	0.26
SVC	0.40	0.34	0.52	0.41	0.77	0.76	0.41	0.28	0.70	0.72	0.35	0.25
Stochastic GDescent	0.37	0.28	0.50	0.26	0.74	0.74	0.51	0.16	0.62	0.81	0.35	0.15
XGBoost	0.48	0.33	0.61	0.43	0.79	0.76	0.55	0.31	0.69	0.67	0.44	0.27

Table 11.C.15: Combined model evaluation metrics for rifampicin resistance prediction: all features
Data comprised of 176 features, with the training set consisting of 2630 SAVs combined from four genes (*embB, gidB, katG*, and *pncA*) based on the 'leave-one-gene-out' approach to test predictions on the 'left out' (i.e. test) gene. The test set comprised of 1132 SAVs from *rpoB*. Train-test data split was undertaken using the scaling law principle, and model performance assessed using a stratified 10-fold cross validation method. Abbreviations used: CV: cross validation, LDA: linear discriminant analysis, MLP: multilayer perceptron, NB: naive Bayes, QDA: quadratic discriminant analysis, SVC: support vector classification, MCC: Matthews correlation coefficient, JCC: jaccard similarity coefficient, SAV: single amino acid variation.

Model Name	MCC Train	MCC Test	F1 Train	F1 Test	Accuracy Train	Accuracy Test	Recall Train	Recall Test	Precision Train	Precision Test	JCC Train	JCC Test
AdaBoost Classifier	0.45	0.25	0.62	0.34	0.76	0.83	0.57	0.31	0.67	0.39	0.45	0.21
Bagging Classifier	0.45	0.28	0.61	0.39	0.76	0.82	0.55	0.38	0.69	0.40	0.44	0.24
Complement NB	0.35	0.28	0.59	0.40	0.68	0.77	0.68	0.53	0.53	0.32	0.42	0.25
Decision Tree	0.32	0.18	0.55	0.32	0.69	0.72	0.55	0.45	0.55	0.25	0.38	0.19
Dummy Classifier	0.00	0.00	0.00	0.00	0.66	0.85	0.00	0.00	0.00	0.00	0.00	0.00
Extra Tree	0.30	0.10	0.54	0.27	0.68	0.67	0.54	0.42	0.54	0.20	0.37	0.16
Extra Trees	0.46	0.28	0.61	0.37	0.77	0.84	0.54	0.31	0.71	0.44	0.44	0.22
Gaussian NB	0.36	0.24	0.57	0.36	0.71	0.81	0.56	0.36	0.58	0.35	0.40	0.22
Gaussian Process	0.44	0.25	0.61	0.35	0.75	0.82	0.56	0.32	0.66	0.39	0.44	0.21
Gradient Boosting	0.49	0.28	0.64	0.37	0.78	0.83	0.58	0.33	0.71	0.42	0.47	0.23
K-Nearest Neighbors	0.41	0.21	0.60	0.34	0.74	0.78	0.57	0.39	0.63	0.30	0.43	0.21
LDA	0.42	0.30	0.59	0.39	0.75	0.84	0.54	0.35	0.67	0.45	0.42	0.24
Logistic Regression	0.44	0.27	0.61	0.36	0.76	0.83	0.55	0.31	0.67	0.41	0.44	0.22
Logistic RegressionCV	0.36	0.24	0.48	0.32	0.74	0.84	0.39	0.26	0.64	0.41	0.33	0.19
MLP	0.43	0.25	0.60	0.36	0.75	0.82	0.55	0.34	0.67	0.38	0.43	0.22
Multinomial NB	0.15	0.16	0.24	0.15	0.67	0.85	0.16	0.09	0.57	0.50	0.14	0.08
Passive Aggressive	0.32	0.00	0.55	0.26	0.61	0.15	0.74	1.00	0.54	0.15	0.39	0.15
QDA	0.35	0.17	0.52	0.29	0.73	0.79	0.44	0.29	0.65	0.30	0.35	0.17
Random Forest	0.47	0.34	0.62	0.40	0.77	0.86	0.55	0.31	0.72	0.54	0.45	0.25
Ridge Classifier	0.43	0.31	0.59	0.39	0.75	0.84	0.53	0.34	0.68	0.46	0.42	0.24
Ridge ClassifierCV	0.45	0.29	0.60	0.38	0.76	0.84	0.53	0.33	0.70	0.44	0.43	0.23
SVC	0.44	0.27	0.61	0.36	0.76	0.83	0.54	0.33	0.69	0.40	0.43	0.22
Stochastic GDescent	0.41	0.24	0.55	0.30	0.73	0.85	0.53	0.22	0.68	0.46	0.39	0.18
XGBoost	0.46	0.29	0.62	0.39	0.76	0.82	0.58	0.38	0.68	0.40	0.45	0.24

Table 11.C.16: Combined model evaluation metrics for ethambutol resistance prediction: post feature selection
Data comprised of 178 features, with the training set consisting of 2904 SAVs combined from four genes (*gidB, katG, pncA*, and *rpoB*) based on the 'leave-one-gene-out' approach to test predictions on the 'left out' (i.e. test) gene. The test set comprised of 858 SAVs from *embB*. Train-test data split was undertaken using the scaling law principle, followed by the Boruta feature selection process[11] which identified 32 features optimised for the MCC score. Abbreviations used: CV: cross validation, LDA: linear discriminant analysis, MLP: multilayer perceptron, NB: naive Bayes, QDA: quadratic discriminant analysis, SVC: support vector classification, MCC: Matthews correlation coefficient, JCC: jaccard similarity coefficient, SAV: single amino acid variation.

Model Name	MCC Train	MCC Test	F1 Train	F1 Test	Accuracy Train	Accuracy Test	Recall Train	Recall Test	Precision Train	Precision Test	JCC Train	JCC Test
AdaBoost Classifier	0.45	0.08	0.62	0.16	0.77	0.63	0.57	0.50	0.68	0.10	0.45	0.09
Bagging Classifier	0.44	-0.03	0.60	0.10	0.76	0.60	0.52	0.32	0.69	0.06	0.42	0.05
Complement NB	0.41	0.04	0.62	0.14	0.72	0.47	0.68	0.63	0.57	0.08	0.45	0.08
Decision Tree	0.29	-0.04	0.53	0.11	0.68	0.42	0.54	0.50	0.52	0.06	0.36	0.06
Dummy Classifier	0.00	0.00	0.00	0.00	0.67	0.93	0.00	0.00	0.00	0.00	0.00	0.00
Extra Tree	0.25	0.04	0.50	0.14	0.67	0.59	0.51	0.47	0.50	0.08	0.34	0.08
Extra Trees	0.44	-0.00	0.59	0.12	0.76	0.60	0.52	0.37	0.70	0.07	0.42	0.06
Gaussian NB	0.39	0.06	0.59	0.14	0.73	0.12	0.58	1.00	0.60	0.07	0.42	0.07
Gaussian Process	0.45	-0.03	0.60	0.09	0.76	0.71	0.54	0.21	0.68	0.06	0.43	0.05
Gradient Boosting	0.47	-0.04	0.62	0.10	0.77	0.60	0.55	0.32	0.71	0.06	0.45	0.05
K-Nearest Neighbors	0.38	0.03	0.57	0.14	0.73	0.59	0.53	0.45	0.62	0.08	0.40	0.07
LDA	0.43	0.00	0.59	0.10	0.76	0.77	0.53	0.18	0.68	0.07	0.42	0.05
Logistic Regression	0.43	-0.02	0.59	0.09	0.76	0.77	0.53	0.16	0.67	0.06	0.42	0.05
Logistic RegressionCV	0.42	-0.00	0.58	0.11	0.76	0.75	0.52	0.21	0.67	0.07	0.41	0.06
MLP	0.42	-0.02	0.59	0.09	0.76	0.76	0.52	0.16	0.67	0.06	0.42	0.05
Multinomial NB	0.26	-0.06	0.33	0.07	0.71	0.70	0.22	0.16	0.69	0.04	0.20	0.04
Passive Aggressive	0.27	-0.01	0.31	0.06	0.70	0.87	0.24	0.05	0.74	0.06	0.20	0.03
QDA	0.34	0.06	0.56	0.15	0.69	0.69	0.57	0.39	0.58	0.09	0.39	0.08
Random Forest	0.47	-0.03	0.61	0.11	0.77	0.59	0.53	0.34	0.72	0.06	0.44	0.06
Ridge Classifier	0.43	0.00	0.59	0.11	0.76	0.78	0.52	0.18	0.68	0.07	0.42	0.06
Ridge ClassifierCV	0.43	0.00	0.59	0.11	0.76	0.78	0.52	0.18	0.68	0.07	0.42	0.06
SVC	0.42	-0.04	0.58	0.08	0.76	0.70	0.49	0.18	0.69	0.05	0.41	0.04
Stochastic GDescent	0.38	0.01	0.51	0.11	0.73	0.81	0.46	0.16	0.69	0.08	0.35	0.06
XGBoost	0.44	-0.03	0.60	0.11	0.76	0.52	0.54	0.42	0.68	0.06	0.43	0.06

Table 11.C.17: Combined model evaluation metrics for streptomycin resistance prediction: post feature selection
Data comprised of 176 features, with the training set consisting of 3231 SAVs combined from four genes (*embB, katG, pncA*, and *rpoB*) based on the 'leave-one-gene-out' approach to test predictions on the 'left out' (i.e. test) gene. The test set comprised of 531 SAVs from *gidB*. Train-test data split was undertaken using the scaling law principle, followed by the Boruta feature selection process[11] which identified 39 features optimised for the MCC score. Abbreviations used: CV: cross validation, LDA: linear discriminant analysis, MLP: multilayer perceptron, NB: naive Bayes, QDA: quadratic discriminant analysis, SVC: support vector classification, MCC: Matthews correlation coefficient, JCC: jaccard similarity coefficient, SAV: single amino acid variation.

Model Name	MCC Train	MCC Test	F1 Train	F1 Test	Accuracy Train	Accuracy Test	Recall Train	Recall Test	Precision Train	Precision Test	JCC Train	JCC Test
AdaBoost Classifier	0.46	0.23	0.57	0.33	0.81	0.61	0.49	0.21	0.68	0.75	0.40	0.19
Bagging Classifier	0.50	0.16	0.60	0.28	0.83	0.59	0.51	0.18	0.72	0.67	0.43	0.16
Complement NB	0.34	0.28	0.53	0.63	0.71	0.64	0.65	0.67	0.45	0.59	0.36	0.46
Decision Tree	0.33	0.13	0.50	0.40	0.74	0.58	0.51	0.31	0.49	0.56	0.33	0.25
Dummy Classifier	0.00	0.00	0.00	0.00	0.75	0.55	0.00	0.00	0.00	0.00	0.00	0.00
Extra Tree	0.33	0.10	0.50	0.39	0.75	0.57	0.51	0.31	0.50	0.53	0.34	0.25
Extra Trees	0.50	0.15	0.58	0.19	0.83	0.58	0.48	0.11	0.75	0.74	0.41	0.10
Gaussian NB	0.36	0.28	0.52	0.49	0.76	0.64	0.52	0.38	0.52	0.69	0.35	0.33
Gaussian Process	0.47	0.23	0.56	0.42	0.82	0.62	0.46	0.30	0.72	0.68	0.39	0.27
Gradient Boosting	0.50	0.16	0.59	0.25	0.83	0.59	0.50	0.15	0.73	0.70	0.42	0.14
K-Nearest Neighbors	0.42	0.15	0.54	0.42	0.79	0.59	0.49	0.33	0.62	0.57	0.37	0.27
LDA	0.48	0.24	0.57	0.39	0.82	0.62	0.48	0.27	0.70	0.71	0.40	0.24
Logistic Regression	0.45	0.26	0.55	0.44	0.81	0.63	0.45	0.32	0.70	0.71	0.38	0.28
Logistic RegressionCV	0.45	0.26	0.54	0.42	0.81	0.63	0.44	0.30	0.71	0.72	0.37	0.27
MLP	0.45	0.20	0.55	0.37	0.81	0.61	0.46	0.25	0.69	0.67	0.38	0.23
Multinomial NB	0.20	0.11	0.25	0.14	0.76	0.57	0.16	0.08	0.58	0.69	0.14	0.08
Passive Aggressive	0.34	0.17	0.43	0.33	0.71	0.59	0.51	0.22	0.61	0.65	0.28	0.20
QDA	0.31	0.14	0.50	0.43	0.71	0.59	0.54	0.35	0.48	0.57	0.33	0.28
Random Forest	0.51	0.18	0.59	0.15	0.83	0.58	0.47	0.08	0.78	0.88	0.41	0.08
Ridge Classifier	0.44	0.21	0.52	0.31	0.81	0.61	0.42	0.20	0.72	0.74	0.36	0.18
Ridge ClassifierCV	0.45	0.23	0.52	0.32	0.81	0.61	0.41	0.20	0.74	0.77	0.35	0.19
SVC	0.46	0.20	0.53	0.32	0.82	0.60	0.41	0.21	0.75	0.70	0.36	0.19
Stochastic GDescent	0.41	0.31	0.45	0.58	0.80	0.66	0.35	0.53	0.77	0.66	0.30	0.41
XGBoost	0.50	0.11	0.60	0.24	0.82	0.57	0.53	0.15	0.70	0.61	0.43	0.14

Table 11.C.18: Combined model evaluation metrics for isoniazid resistance prediction: post feature selection
Data comprised of 176 features, with the training set consisting of 2945 SAVs combined from four genes (*embB, gidB, pncA*, and *rpoB*) based on the 'leave-one-gene-out' approach to test predictions on the 'left out' (i.e. test) gene. The test set comprised of 817 SAVs from *katG*. Train-test data split was undertaken using the scaling law principle, followed by the Boruta feature selection process[11] which identified 31 features optimised for the MCC score. Abbreviations used: CV: cross validation, LDA: linear discriminant analysis, MLP: multilayer perceptron, NB: naive Bayes, QDA: quadratic discriminant analysis, SVC: support vector classification, MCC: Matthews correlation coefficient, JCC: jaccard similarity coefficient, SAV: single amino acid variation.

Model Name	MCC Train	MCC Test	F1 Train	F1 Test	Accuracy Train	Accuracy Test	Recall Train	Recall Test	Precision Train	Precision Test	JCC Train	JCC Test
AdaBoost Classifier	0.40	0.27	0.51	0.71	0.79	0.65	0.43	0.74	0.64	0.69	0.34	0.55
Bagging Classifier	0.42	0.40	0.52	0.79	0.80	0.71	0.44	0.88	0.66	0.71	0.36	0.65
Complement NB	0.31	0.40	0.52	0.79	0.67	0.71	0.68	0.95	0.42	0.68	0.35	0.66
Decision Tree	0.27	0.17	0.47	0.68	0.72	0.61	0.48	0.70	0.46	0.66	0.30	0.51
Dummy Classifier	0.00	0.00	0.00	0.00	0.74	0.41	0.00	0.00	0.00	0.00	0.00	0.00
Extra Tree	0.25	0.17	0.45	0.64	0.71	0.59	0.46	0.61	0.44	0.67	0.29	0.47
Extra Trees	0.40	0.46	0.49	0.79	0.80	0.74	0.38	0.82	0.69	0.76	0.33	0.65
Gaussian NB	0.30	0.31	0.46	0.76	0.74	0.68	0.44	0.88	0.50	0.67	0.30	0.62
Gaussian Process	0.39	0.33	0.48	0.74	0.79	0.68	0.38	0.78	0.66	0.71	0.32	0.59
Gradient Boosting	0.43	0.40	0.52	0.78	0.80	0.72	0.43	0.86	0.68	0.72	0.36	0.64
K-Nearest Neighbors	0.36	0.26	0.49	0.68	0.77	0.63	0.42	0.66	0.59	0.70	0.32	0.52
LDA	0.38	0.42	0.48	0.78	0.79	0.72	0.38	0.82	0.66	0.74	0.32	0.63
Logistic Regression	0.39	0.39	0.48	0.76	0.79	0.71	0.37	0.79	0.67	0.73	0.31	0.61
Logistic RegressionCV	0.05	0.00	0.05	0.00	0.75	0.41	0.03	0.00	0.15	0.00	0.03	0.00
MLP	0.39	0.31	0.49	0.77	0.79	0.68	0.40	0.89	0.65	0.67	0.33	0.62
Multinomial NB	0.04	0.00	0.02	0.00	0.74	0.41	0.01	0.00	0.40	0.00	0.01	0.00
Passive Aggressive	0.25	0.04	0.32	0.01	0.73	0.41	0.35	0.00	0.50	1.00	0.21	0.00
QDA	0.29	0.27	0.43	0.73	0.76	0.66	0.36	0.78	0.54	0.68	0.28	0.57
Random Forest	0.42	0.44	0.50	0.79	0.80	0.73	0.39	0.83	0.72	0.75	0.34	0.65
Ridge Classifier	0.37	0.36	0.44	0.73	0.79	0.69	0.32	0.72	0.70	0.74	0.28	0.58
Ridge ClassifierCV	0.36	0.37	0.42	0.72	0.79	0.69	0.30	0.69	0.72	0.76	0.26	0.57
SVC	0.36	0.40	0.41	0.74	0.79	0.70	0.29	0.71	0.72	0.77	0.26	0.59
Stochastic GDescent	0.33	0.32	0.38	0.59	0.77	0.62	0.30	0.47	0.72	0.80	0.24	0.42
XGBoost	0.40	0.40	0.53	0.78	0.78	0.72	0.47	0.85	0.60	0.72	0.36	0.64

Table 11.C.19: Combined model evaluation metrics for pyrazinamide resistance prediction: post feature selection
Data comprised of 176 features, with the training set consisting of 3338 SAVs combined from four genes (*embB, gidB, katG*, and *rpoB*) based on the 'leave-one-gene-out' approach to test predictions on the 'left out' (i.e. test) gene. The test set comprised of 424 SAVs from *pncA*. Train-test data split was undertaken using the scaling law principle, followed by the Boruta feature selection process[11] which identified 24 features optimised for the MCC score. Abbreviations used: CV: cross validation, LDA: linear discriminant analysis, MLP: multilayer perceptron, NB: naive Bayes, QDA: quadratic discriminant analysis, SVC: support vector classification, MCC: Matthews correlation coefficient, JCC: jaccard similarity coefficient, SAV: single amino acid variation.

Model Name	MCC Train	MCC Test	F1 Train	F1 Test	Accuracy Train	Accuracy Test	Recall Train	Recall Test	Precision Train	Precision Test	JCC Train	JCC Test
AdaBoost Classifier	0.46	0.35	0.60	0.45	0.79	0.76	0.55	0.33	0.67	0.68	0.43	0.29
Bagging Classifier		0.31		0.40		0.75		0.29		0.67		0.25
Complement NB	0.38	0.35	0.59	0.54	0.68	0.73	0.76	0.55	0.48	0.53	0.42	0.37
Decision Tree	0.37	0.23	0.56	0.46	0.73	0.68	0.57	0.46	0.55	0.45	0.39	0.30
Dummy Classifier	0.00	0.00	0.00	0.00	0.70	0.71	0.00	0.00	0.00	0.00	0.00	0.00
Extra Tree	0.29	0.19	0.50	0.44	0.70	0.65	0.50	0.48	0.50	0.41	0.33	0.28
Extra Trees	0.45	0.34	0.58	0.42	0.78	0.76	0.51	0.30	0.69	0.70	0.41	0.27
Gaussian NB	0.37	0.30	0.56	0.45	0.73	0.74	0.58	0.37	0.55	0.57	0.39	0.29
Gaussian Process	0.45	0.37	0.59	0.48	0.78	0.77	0.53	0.37	0.67	0.68	0.42	0.31
Gradient Boosting	0.50	0.40	0.62	0.47	0.80	0.78	0.55	0.34	0.72	0.75	0.45	0.31
K-Nearest Neighbors	0.38	0.31	0.54	0.45	0.75	0.74	0.49	0.37	0.61	0.58	0.37	0.29
LDA	0.46	0.35	0.59	0.44	0.78	0.76	0.53	0.32	0.68	0.69	0.42	0.28
Logistic Regression	0.46	0.37	0.59	0.48	0.79	0.77	0.53	0.36	0.69	0.69	0.42	0.31
Logistic RegressionCV	0.45	0.37	0.58	0.47	0.78	0.77	0.51	0.35	0.68	0.70	0.41	0.31
MLP	0.46	0.39	0.60	0.48	0.79	0.77	0.55	0.36	0.68	0.72	0.43	0.32
Multinomial NB	-0.02	0.20	0.00	0.11	0.70	0.73	0.00	0.06	0.00	1.00	0.00	0.06
Passive Aggressive	0.32	0.38	0.48	0.58	0.66	0.70	0.62	0.73	0.53	0.49	0.34	0.41
QDA	0.37	0.22	0.54	0.38	0.75	0.71	0.49	0.30	0.60	0.52	0.37	0.23
Random Forest	0.49	0.36	0.61	0.42	0.80	0.77	0.52	0.29	0.73	0.75	0.43	0.27
Ridge Classifier	0.45	0.36	0.58	0.44	0.79	0.77	0.50	0.32	0.69	0.71	0.41	0.28
Ridge ClassifierCV	0.44	0.38	0.57	0.47	0.78	0.77	0.49	0.34	0.68	0.72	0.40	0.30
SVC	0.45	0.38	0.59	0.47	0.78	0.77	0.51	0.35	0.69	0.71	0.41	0.31
Stochastic GDescent	0.45	0.39	0.57	0.45	0.77	0.77	0.56	0.32	0.66	0.76	0.41	0.29
XGBoost	0.45	0.39	0.59	0.47	0.78	0.77	0.54	0.35	0.67	0.72	0.42	0.31

Table 11.C.20: Combined model evaluation metrics for rifampicin resistance prediction: post feature selection
Data comprised of 176 features, with the training set consisting of 2630 SAVs combined from four genes (*embB, gidB, katG*, and *pncA*) based on the 'leave-one-gene-out' approach to test predictions on the 'left out' (i.e. test) gene. The test set comprised of 1132 SAVs from *rpoB*. Train-test data split was undertaken using the scaling law principle, followed by the Boruta feature selection process[11] which identified 30 features optimised for the MCC score. Abbreviations used: CV: cross validation, LDA: linear discriminant analysis, MLP: multilayer perceptron, NB: naive Bayes, QDA: quadratic discriminant analysis, SVC: support vector classification, MCC: Matthews correlation coefficient, JCC: jaccard similarity coefficient, SAV: single amino acid variation.

11.D AI/ML Model Explorer dashboard

Figure 11.D.1: AI/ML Model Explorer interface
The web interface for the AI/ML Model Explorer makes it easy to visually compare the various ML approaches applied to the gene-specific and gene-agnostic data. The upper plot shows the 'baseline' model with all features (highlighted in red). The lower plot shows models after feature selection (highlighted in blue). On the sidebar (highlighted in green) are controls appropriate to the current graph model ('Combined' or 'Gene', highlighted in yellow). Several resampling techniques (also highlighted in green) may be explored to illustrate differences in model performance. Note that the Feature Selection plot is not affected by this option.

Chapter 12

Discussion, Conclusion, and Future Work

12.1 Discussion

The thesis focussed on three aspects of understanding resistance development in six structural *M. tuberculosis* genes: 1) investigating the relationship between genomic and biophysical measures of SAV mutations in six structural *M. tuberculosis* genes, with a web-based visualisation tool developed to investigate these multiple targets interactively, 2) resistance profiling of SAVs by lineage, and 3) resistance prediction in a machine learning framework in a gene-target as well as a gene-agnostic manner. To the best of our knowledge, this project is the first attempt to combine genomic and structural impacts of SAV mutations to further the understanding of resistance development in *M. tuberculosis*. The project utilised the largest dataset available to date incorporating over 35,000 clinical isolates from more than 100 countries containing 12,935 unique SAVs in the protein coding region of the recorded genes in the dataset. Nearly 80% (n=28,217) of the isolates displayed over 4000 SAVs across the six genes investigated in this thesis. The thesis also explored the GidB and EmbB protein structural landscape in a manner not previously attempted. Standard statistical and machine learning techniques were used systematically to explore these dynamics individually as well as in a combined manner. While some individual genes and their structural mutational landscapes have been explored by others,[1–5] this thesis brings together a large genomic dataset, protein structure and machine learning for a comprehensive insight into the genotype-phenotype relationship.

Chapters 3-8 elucidate the complex genotype-phenotype relationship related to resistance development. Computational tools based on different underlying methodologies provide different information with respect to mutational effects. Sequence based methods (ConSurf, PROVEAN, and SNAP2) rely on evolutionary conservation and substitution matrices to estimate impact on protein function, while structure based methods (DeepDDG, Dynamut2, mCSM-DUET, and FoldX) consider the protein structural environment to assess mutational impact on structure stability. The evolutionary-based methods are aimed at predicting pathogenic effects of variants while the structure-based tools are aimed at predicting structure stability consequences without regard for pathogenicity. Hence, a variant classified as 'deleterious' to protein conservation may display gain-of-function in the presence of a drug through optimised protein stability. Therefore, when assessing specific proteins, different methodological strategies should be used as these benefit from distinct and complementary insights to understand the genotype-phenotype interrelationship.[6–8]

With SAVs distributed across the protein and extending beyond the active site in all six structural genes, allosteric mutational effects becomes necessary to consider. Allostery is defined as the transmission of information spatially from one site to another in a protein.[9] In the context of drug resistance,

these are those phenotypically resistant mutations that occur outside the active site. Allosteric effects are brought about by conferring protomer stability and non-ligand interaction affinities without directly affecting the drug binding affinity. Further, it also appeared that SAVs imparting drastic enthalpic changes associated with high fitness penalty for one biophysical measure (e.g. protein stability) is compensated by more favourable effects on other measures (e.g. NA affinity). In this way compensatory effects related to different enthalpic changes for a given SAV, as well compensatory effects between SAVs play a role in acquisition of resistance. Another notion that emerged was that selection pressure from multiple drugs owing to treatment regimens may also contribute to the resistance development. For example, it has been reported that the *embB* 306 mutation predisposes strains to acquire resistance to INH and RFP.[10]

In the course of these analyses, it became apparent that DST data currently used to confirm resistance in TB diagnosis is suboptimal. This is primarily due to the time-consuming culture-based methods for DST, as well as the difficulty in quantifying resistance thresholds, for certain drugs like EMB (poor solubility[11] and DCS (poor reliability).[12] DST is recognised to be highly variable with respect to the cut-off thresholds for different drugs as well as in a drug for a given mutation from different strains[13,14] highlighting major issues in its widespread use for classifying mutations. It would appear that DST data results would benefit from including corresponding MIC values that help quantify the extent of drug resistance.

Another important consideration brought to light was the need to consider multiple SAVs (in the same gene or across multiple genes) occurring in the same clinical isolate. This becomes especially important when warranting the use of whole genome sequencing or GWAS based resistance inference, as a way of substituting DST in *M. tuberculosis*. In the context of estimating the biophysical impact of SAVs, an improved strategy might include modelling the protein structure with co-occurring mutations to get a better and direct estimates of protomer stability and drug/PPI/NA binding affinities. Alternatively, computational tools predicting local and global biophysical changes of multiple mutations (concomitant or haplotype combinations) can improve mutational effect predictions. In general, it is thought that mutations with a low bacterial or protein fitness cost get 'fixed' in the population. A classic example is that of *katG* S315T mutation, that has a selection advantage, where it is low fitness and high resistance –the 'sweet spot', and thus becomes enriched or abundant in the bacterial population.[15] If certain combination of mutations occur frequently in a population, it implies that concomitant haplotypes for such gene-drug combinations improve protein and bacterial fitness especially in the presence of a drug. Conversely, if a mutation always occurs with compensatory mutations, it is thought to have a high fitness cost. Protein fitness and resistance acquisition thus operates in a complex landscape

involving interactions of different enthalpic measures, concomitant and compensatory mutations, and mutational frequency in interacting proteins.

The assumption of independence of individual polymorphisms is important at first to help establish the 'baseline' understanding of mutational impact on protein structure and binding partner affinity. As such, in these analyses, Odds Ratio for each SAV was calculated based on a contingency table used to classify outcomes (resistance) in rows and columns from the entire data set. The number in each cell indicates the frequency, whether a given SAV is resistant or sensitive according to DST. Such an approach is not able to account for concomitant mutations. Appropriate statistical methods akin to logistic regression and the kinship matrix can help account for the dependency among SAVs correlated with phenotypes due to the underlying *M. tuberculosis* population structure. Further improvements can be made where DST results are accompanied by corresponding MIC values that help quantify the extent of resistance. Furthermore, the strengths of GWAS studies could be better utilised by incorporating added measures (dN/dS: rate of non-synonymous to synonymous mutations) able to quantify selection pressures, to improve the combined genomics and structure approach to understanding resistance development.

With the identified need to account for the underlying population structure of *M. tuberculosis*, an initial attempt was made to investigate mutational effects across lineages. **Chapter 10** extended the lineage information to explore the epistatic landscape of *M. tuberculosis* with respect to the six gene targets. While only a preliminary analysis, the insights gained were valuable and highlighted that a given mutation can potentially be resistant in one lineage but sensitive in another. Consequently, strain diversity (i.e. lineage effects) becomes important to consider, not only for a holistic understanding of resistance development,[16–19] but also for clinical management and personalised therapy in TB.[20,21] Taken together, these insights from **Chapter 10** help to illustrate the complex resistance landscape and the importance of considering the genetic background in understanding resistance evolution in TB. Ultimately, the work in this thesis resonates with other findings that suggest that the relevance of lineage and strain effects are more pronounced in *M. tuberculosis* than originally believed. Pervasive epistatic interactions involving compensatory mutations invariably affect the fitness of drug resistant strains,[16,17,22] as well as evidence for seemingly incompatible mutations becoming fixed in different lineages in general.[23] Altogether, these observations and inferences warrant the need to understand resistance development in response to phenotypic pressure exerted by treatment regimens, lineage information and co-occurring SAVs in clinical isolates. The analysis performed here presents a framework to analyse mutational data with respect to *M. tuberculosis* lineages, and to inform resistance profiling of DST data to utilise *M. tuberculosis* strain diversity.

Building on the detailed analyses of mutations and their effects explored in **Chapters 9**, and the importance of considering strain diversity in **Chapter 10**, the work in **Chapter 11** sought to leverage all observations (features) analysed across the multiple targets to build machine learning models to predict resistance in a gene-target, as well as in a combined, gene-agnostic manner. There are several resistance prediction tools for *M. tuberculosis* exploiting WGS data directly[13,24–27] with TB-Profiler[25] and Mykrobe[26] currently being the two leading ones. Additionally, ML driven approaches have built on these advancements of resistance prediction[27,28] while others have furthered these by including additional data (features) like protein stability, binding affinity (drug, nucleic acid, PPI), residue level properties (surface area, depth, etc.) towards resistance prediction.[2,4,29,30]

Most ML tools reporting on the performance of their respective resistance predictions primarily focus on accuracy and ROC curves,[28,29] and to a lesser extent on precision and recall scores,[27,31] with only a handful of tools reporting the more balanced MCC metric.[2,30] The SUSPECT-PZA[2] and SUSPECT-RIF[30] tools are freely available, fast and easy to use and of great benefit to the wider scientific community. The ML analysis performed in this project optimised for the MCC score. Considering that existing tools such as SUSPECT-PZA[2] and SUSPECT-RIF[30] are closely aligned due to using shared features like protein stability, binding affinity changes, and residue level properties, comparing performance with these tools serves as a good indicator of the overall reliability in the method for building individual gene-drug target models, as well as the potential of a combined model. While both these tools marginally outperform the PncA-PZA and RpoB-RFP individual models developed during this project, their differences brought useful insights to the fore. Firstly, ML features used in my analysis included estimates from other computational stability predictor tools like FoldX and DeepDDG, as well as from genomic features like allele frequency, lineage information, and mutational site frequency. In SUSPECT-PZA,[2] the final model performance achieved an MCC score of 0.6, however the authors do not separately report the train and test MCC scores to inform about the 'learning' capacity of the model. If we consider the baseline model with all features included and do not apply the <0.1 difference in train-test MCC scores, then the PncA-PZA model implemented here is equally able to achieve an MCC score of 0.64 using Random Forest, the same model used in SUSPECT-PZA. However, when respecting the criteria set out in the current analysis, the MCC score of 0.5 is thought to be comparable. Additionally, the features generated in these analyses span a wide range of different tools and include genomics features to allow an inter-disciplinary approach to predicting resistance.

Similarly, SUSPECT-RIF,[30] an ML driven RFP resistance predictor, also achieved a final MCC score of 0.6 using the KNN model. The study generated additional structural features and then performed

a 'greedy' feature selection via in-house scripts. Greedy feature selection is time consuming and computationally intensive as it needs to iteratively add or drop features at each stage to derive the final model. Due to its exhaustive nature, it is therefore more effective when used on smaller data sets. The final MCC score in my analysis for RpoB-RFP was 0.5 using XGBoost, which, when including the non-overlapping features and the use of a different and efficient feature selection strategy (BoruatPy, see **Chapter 11**, methods section) used, is comparable. When the RFECV greedy feature selection algorithm (available from Python scikit-learn[32]) was attempted in this analysis, major run-time issues were encountered, with some algorithms taking >15 hours to run on a 96 core Xeon workstation with 128 GiB of RAM. This was for a single gene-target and single train-test split type, making it unfeasible to attempt multiple iterations for streamlining the ML pipeline.

The mCSM based SUSPECT studies mentioned above generate a multitude of protein structure-related features using mCSM-graph based signatures (both the overall estimate as well as the individual graph-based signatures e.g. changes in the pi-pi interaction, hydrogen-bond donors, etc.). This process has the advantage of acutely capturing all the small-to-large scale molecular interactions contributing to resistance prediction. However, the models may only be reproduced by running the in-house scripts that generate the individual graph-based features, and the additional greedy feature selection. The analysis presented here, however, shows that by using a breadth of different features, it is possible to achieve comparable performance, with insight into complementary features like mutational frequency and lineage contribution to advance the understanding of resistance prediction. An example of when different contributing features could be better suited at resistance prediction is illustrated as follows: Discovery of a novel variant with no mutational or lineage frequency information, tools exploiting the structural features will initially be most useful. However, as the mutation becomes a circulating variant in a population, its evolution across the lineages, interaction with other genes, mutational frequency, and co-occurrence with other mutations becomes much more relevant in dictating its trajectory towards resistance, making ML tools informed by genomic measures better suited.

As currently a general AMR predictor in *M. tuberculosis*, does not exist, to the best of our knowledge, I have attempted to develop a combined model facilitating 'learning' across targets, to reveal the potential of pursuing a gene-agnostic approach towards resistance prediction. The success of this approach would lend support for 'cross-learning', and would enhance our understanding of the complex genotype-phenotype relationship. The combined model achieved the target MCC score of at least 0.4, offering the promise of cross-learning between genes. The modelling also revealed, however, that the test MCC score is biased towards the quality of the 'test' gene data **(Chapter 11)**. As is the case with first version of anything, the combined model requires further optimisations, focused around

establishing DST data integrity, preferably using resistance inference from WGS data, or alternatively pursuing regression-based (using MIC values from DST) and unsupervised ML learning. Further improvements could come from the application of knowledge-based weighting strategies to specific gene-targets **(Chapter 11)** and the inclusion of an independent blind test dataset, a limitation of the current analysis. Harnessing the efforts led by the CRyPTIC consortium, diverse GWAS datasets can certainly improve the data integrity, and for drugs like DCS, where predicting resistance has proven to be challenging, a general AMR predictor can help.

Overall, ML driven solutions are powerful, scalable, and easily integrable with big-data in the post-genomics era. Developing accurate, and explainable ML/AI solutions, and using open-access tools and technology, can speed up digital transformation in healthcare and clinical research. However, it is important to raise awareness regarding some ML fundamentals. A lack of consensus on the use of a unified score metric when evaluating ML based classification tasks, despite it being a crucial issue, highlights the need for caution when evaluating ML powered solutions in clinical decision making.[33] For example, in the context of supervised ML classification tasks, the MCC score makes use of all four categories of a confusion matrix (true and false positive and negatives), and as such is a more 'truthful' and informative score, but seldom ends up being reported.[33] This may well depend on the field and the research question at hand, and thus it is recognised that different metrics are useful in different cases. For example, the recall or sensitivity score is the metric of choice to optimise for when the use case warrants maximisation of true positives (true and false), for instance when developing clinical screening tests. Whereas when the use case involves administering a drug with potential side effects, the metric to optimise for would be specificity that reduces false positives. When a use case is not as straightforward and requires a combination of both, the F1 score is used (see **Chapter 11**, methods section for more details).

Similarly, the widely reported 'accuracy' score metric is only valid when measured relative to a Dummy Classifier. As the name suggests, a Dummy Classifier is a model that makes predictions without finding any patterns in the data. Typically it predicts the most frequent class in the data, although this can be changed to predict the class of interest. A Dummy Classifier helps establish a baseline from which to assess other model performances, in the absence of which, imbalanced data risks being misrepresented in an overly optimistic way. Despite this, and for the reasons mentioned above, adoption of the MCC score in assessing model performance has been limited across the wider scientific community. There are also extended metrics such as the Brier score,[34] which help quantify the 'confidence' of a model using predicted probabilities, for example if two models have the same accuracy or MCC scores, the Brier score will assist in choosing the appropriate model.

In general, the thesis supports the efforts of the CRyPTIC consortium (Comprehensive Resistance Prediction for Tuberculosis) aimed at replacing culture-based DST testing for TB with whole genome sequencing to allow faster and more accurate identification and management of DR TB. The thesis has also identified the need to extend resistance prediction from binary classification to a regression (extent of resistance) based approach, with emphasis placed on the need to include strain diversity, concomitant and compensatory mutations.

12.2 Conclusion

This thesis has contributed to the development of an integrated approach using protein structure and genomics data to investigate and predict drug resistance in *M. tuberculosis*. Specifically, the thesis has contributed in three key areas:

- A large-scale detailed analysis of the consequences of mutations associated with resistance across a wide range of TB drug targets. The results of these analyses with respect to protein structure including protein stability changes in lineages can be interrogated in an interactive visualisation tool.
- The importance of considering strain diversity in understanding the resistance landscape.
- Development of a gene-specific and gene-agnostic ML-driven resistance prediction tool.

Chapters 3-8 systematically investigated the impact of SAV mutations on the protein structure and binding affinities of six drugs used in TB treatment, including mutational impact on nucleic acid, protein-protein interface binding affinity, and functional effects. **Chapters 9** integrated the findings from these chapters and clarified the relationship between biophysical and genomic measures, establishing that frequently occurring mutations have fitness advantages, while the relationship of resistant mutations and protein stability is influenced by additional factors requiring further investigation. Building on this, **Chapter 10** provided preliminary yet important insights on the importance of considering strain diversity, epistasis and compensatory mutations in resistance development. The intent is to inform personalised TB treatment regimens to limit the spread of resistance. Such personalised therapy can be tailored to the genotypic and phenotypic data based on clinical isolates identified from individual patients. Where genotypic data from sequencing technologies can rapidly help identify mutations and lineage in specific genes, phenotypic data is then able to associate DST to *M. tuberculosis* lineage to determine specific drug resistance outcome. Thus, the ability to determine factors (lineage specific mutations) in isolates extracted from patients, and relating these to appropriate DST, will help

to accurately tailor treatment regimens to improve disease and thus resistance outcome. **Chapter 11** explored the development of a gene-focussed and gene-agnostic, ML driven resistance prediction tool. Further improvements in refining this have the potential to inform resistance prediction for gene-drug targets with mutations that are especially challenging to assay using DST.

With poor reliability and other challenges with DST, rapid turnaround from sequencing technologies is increasingly guiding clinical management. With this wealth of data to exploit, ML approaches are likely to become the mainstay in TB healthcare. However, the effective delivery of ML/AI powered solutions can only come about from systematic and more fundamental research into resistance evolution in *M. tuberculosis*. Therefore, it is incumbent upon us to engage effectively with existing global efforts like the CRyPTIC consortium for TB at mapping and predicting variants in *M. tuberculosis* by developing standard protocols in line with the FAIR principles[35] for aiding Findability, Accessibility, Interoperability and Reproducibility in research data management and delivery. In line with this, the variant data analysed in this thesis, along with its biophysical estimates, is intended to be integrated into the Protein Data Bank in Europe-Knowledge Base (PDBe-KB)[36] for the wider benefit of the scientific community. As such, this thesis contributes to this wider context, with the promise of a general AMR predictor in TB informed by mutational interactions between several genes.

12.3 Future Work

Future avenues of research that could follow from this work might be to develop lineage-specific structural models to assess mutational impact. Mutational impact could in turn be further improved by including epistasis and compensatory mutation modelling. There are also opportunities to both leverage WGS datasets in improving existing supervised ML tools, and to develop new unsupervised learning approaches for resistance prediction. Emphasis should also be placed on using regression based ML models as compared with classification models to the assessment of samples for resistance prediction. The general methodology used in this work can be further extended to other AMR pathogens.

References

[1] Stephanie Portelli et al. "Understanding Molecular Consequences of Putative Drug Resistant Mutations in Mycobacterium Tuberculosis". In: *Scientific Reports* (2018). ISSN: 20452322. DOI: 10.1038/s41598-018-33370-6.

[2] Malancha Karmakar et al. *SUSPECT-PZA | Home*. 2018.

[3] Malancha Karmakar et al. "Empirical Ways to Identify Novel Bedaquiline Resistance Mutations in AtpE". In: *PloS One* 14.5 (2019), e0217169. ISSN: 1932-6203. DOI: 10.1371/journal.pone.0217169.

[4] Erol S. Kavvas et al. “Machine Learning and Structural Analysis of Mycobacterium Tuberculosis Pan-Genome Identifies Genetic Signatures of Antibiotic Resistance”. In: *Nature Communications* 9.1 (Dec. 2018), p. 4306. ISSN: 2041-1723. DOI: 10.1038/s41467-018-06634-y.

[5] Arnold Amusengeri, Asifullah Khan, and Özlem Tastan Bishop. “The Structural Basis of Mycobacterium Tuberculosis RpoB Drug-Resistant Clinical Mutations on Rifampicin Drug Binding”. In: *Molecules* 27.3 (Jan. 28, 2022), p. 885. ISSN: 1420-3049. DOI: 10.3390/molecules27030885.

[6] Tanushree Tunstall et al. “Combining Structure and Genomics to Understand Antimicrobial Resistance”. In: *Computational and Structural Biotechnology Journal* 18 (2020), pp. 3377–3394. ISSN: 2001-0370. DOI: 10.1016/j.csbj.2020.10.017.

[7] Douglas Pires, David B. Ascher, and Tom L. Blundell. “DUET: A Server for Predicting Effects of Mutations on Protein Stability Using an Integrated Computational Approach”. In: *Nucleic Acids Research* 42.W1 (2014), pp. 314–319. ISSN: 13624962. DOI: 10.1093/nar/gku411.

[8] Huali Cao et al. “DeepDDG: Predicting the Stability Change of Protein Point Mutations Using Neural Networks”. In: *Journal of Chemical Information and Modeling* 59.4 (Apr. 22, 2019), pp. 1508–1514. ISSN: 1549-960X. DOI: 10.1021/acs.jcim.8b00697.

[9] Andre J. Faure et al. “Mapping the Energetic and Allosteric Landscapes of Protein Binding Domains”. In: *Nature* 604.7904 (7904 Apr. 2022), pp. 175–183. ISSN: 1476-4687. DOI: 10.1038/s41586-022-04586-4.

[10] Manzour Hernando Hazbón et al. “Role of embB Codon 306 Mutations in Mycobacterium Tuberculosis Revisited: A Novel Association with Broad Drug Resistance and IS6110 Clustering Rather than Ethambutol Resistance”. In: *Antimicrobial Agents and Chemotherapy* 49.9 (Sept. 2005), pp. 3794–3802. ISSN: 0066-4804. DOI: 10.1128/AAC.49.9.3794-3802.2005.

[11] Hassan Safi et al. “Evolution of High-Level Ethambutol-Resistant Tuberculosis through Interacting Mutations in Decaprenylphosphoryl--d-Arabinose Biosynthetic and Utilization Pathway Genes”. In: *Nature genetics* 45.10 (Oct. 2013), pp. 1190–1197. ISSN: 1061-4036. DOI: 10.1038/ng.2743.

[12] Ruibai Wang, Xiuqin Zhao, and Kanglin Wan. “Deterioration of Cycloserine in Drug Susceptibility Testing of Mycobacterium”. In: *Infection and Drug Resistance* 15 (Jan. 13, 2022), pp. 135–140. ISSN: 1178-6973. DOI: 10.2147/IDR.S348428.

[13] Jinli Li et al. “Whole-Genome Sequencing for Resistance Level Prediction in Multidrug-Resistant Tuberculosis”. In: *Microbiology Spectrum* 10.3 (June 6, 2022), e02714–21. DOI: 10.1128/spectrum.02714-21.

[14] Viola Schleusener et al. “Mycobacterium Tuberculosis Resistance Prediction and Lineage Classification from Genome Sequencing: Comparison of Automated Analysis Tools”. In: *Scientific Reports* 7 (Apr. 20, 2017), p. 46327. ISSN: 2045-2322. DOI: 10.1038/srep46327.

[15] S. Ramaswamy and J. M. Musser. “Molecular Genetic Basis of Antimicrobial Agent Resistance in Mycobacterium Tuberculosis: 1998 Update”. In: *Tubercle and Lung Disease: The Official Journal of the International Union Against Tuberculosis and Lung Disease* 79.1 (1998), pp. 3–29. ISSN: 0962-8479. DOI: 10.1054/tuld.1998.0002.

[16] Mireilla Coscolla and Sebastien Gagneux. “Does M. Tuberculosis Genomic Diversity Explain Disease Diversity?” In: *Drug discovery today. Disease mechanisms* 7.1 (2010), e43–e59. ISSN: 1740-6765. DOI: 10.1016/j.ddmec.2010.09.004.

[17] Sònia Borrell and Sebastien Gagneux. “Strain Diversity, Epistasis and the Evolution of Drug Resistance in Mycobacterium Tuberculosis”. In: *Clinical microbiology and infection : the official publication of the European Society of Clinical Microbiology and Infectious Diseases* 17.6 (June 2011), pp. 815–820. ISSN: 1198-743X. DOI: 10.1111/j.1469-0691.2011.03556.x.

[18] J. S. Verma et al. “Evaluation of gidB Alterations Responsible for Streptomycin Resistance in Mycobacterium Tuberculosis”. In: *Journal of Antimicrobial Chemotherapy* 69.11 (Nov. 1, 2014), pp. 2935–2941. ISSN: 0305-7453, 1460-2091. DOI: 10.1093/jac/dku273.

[19] Sebastian M. Gygli et al. “Antimicrobial Resistance in Mycobacterium Tuberculosis: Mechanistic and Evolutionary Perspectives”. In: *FEMS microbiology reviews* 41.3 (May 1, 2017), pp. 354–373. ISSN: 1574-6976. DOI: 10.1093/femsre/fux011.

[20] Nikhat Khan and Aparup Das. "Can the Personalized Medicine Approach Contribute in Controlling Tuberculosis in General and India in Particular?" In: *Precision Clinical Medicine* 3.3 (June 4, 2020), pp. 240–243. ISSN: 2096-5303. DOI: 10.1093/pcmedi/pbaa021.

[21] Kartik Kumar and Onn Min Kon. "Personalised Medicine for Tuberculosis and Non-Tuberculous Mycobacterial Pulmonary Disease". In: *Microorganisms* 9.11 (Oct. 26, 2021), p. 2220. ISSN: 2076-2607. DOI: 10.3390/microorganisms9112220.

[22] Amel Kevin Alame Emane et al. "Drug Resistance, Fitness and Compensatory Mutations in Mycobacterium Tuberculosis". In: *Tuberculosis* 129 (July 1, 2021), p. 102091. ISSN: 1472-9792. DOI: 10.1016/j.tube.2021.102091.

[23] Mark Lunzer, G. Brian Golding, and Antony M. Dean. "Pervasive Cryptic Epistasis in Molecular Evolution". In: *PLoS genetics* 6.10 (Oct. 21, 2010), e1001162. ISSN: 1553-7404. DOI: 10.1371/journal.pgen.1001162.

[24] Pierre Mahé and Maud Tournoud. "Predicting Bacterial Resistance from Whole-Genome Sequences Using k-Mers and Stability Selection". In: *BMC bioinformatics* 19.1 (Oct. 17, 2018), p. 383. ISSN: 1471-2105. DOI: 10.1186/s12859-018-2403-z.

[25] Jody E. Phelan et al. "Integrating Informatics Tools and Portable Sequencing Technology for Rapid Detection of Resistance to Anti-Tuberculous Drugs". In: *Genome Medicine* 11.1 (June 24, 2019), p. 41. ISSN: 1756-994X. DOI: 10.1186/s13073-019-0650-x.

[26] Martin Hunt et al. "Antibiotic Resistance Prediction for Mycobacterium Tuberculosis from Genome Sequence Data with Mykrobe". In: *Wellcome Open Research* 4 (2019), p. 191. ISSN: 2398-502X. DOI: 10.12688/wellcomeopenres.15603.1.

[27] Matthias I. Gröschel et al. "GenTB: A User-Friendly Genome-Based Predictor for Tuberculosis Resistance Powered by Machine Learning". In: *Genome Medicine* 13.1 (Aug. 30, 2021), p. 138. ISSN: 1756-994X. DOI: 10.1186/s13073-021-00953-4.

[28] Michael L. Chen et al. "Beyond Multidrug Resistance: Leveraging Rare Variants with Machine and Statistical Learning Models in Mycobacterium Tuberculosis Resistance Prediction". In: *EBioMedicine* 43 (May 1, 2019), pp. 356–369. ISSN: 2352-3964. DOI: 10.1016/j.ebiom.2019.04.016.

[29] Salma Jamal et al. "Artificial Intelligence and Machine Learning Based Prediction of Resistant and Susceptible Mutations in Mycobacterium Tuberculosis". In: *Scientific Reports* 10.1 (1 Mar. 26, 2020), p. 5487. ISSN: 2045-2322. DOI: 10.1038/s41598-020-62368-2.

[30] Stephanie Portelli et al. "Prediction of Rifampicin Resistance beyond the RRDR Using Structure-Based Machine Learning Approaches". In: *Scientific Reports* 10.1 (1 Oct. 22, 2020), p. 18120. ISSN: 2045-2322. DOI: 10.1038/s41598-020-74648-y.

[31] Wouter Deelder et al. "Machine Learning Predicts Accurately Mycobacterium Tuberculosis Drug Resistance From Whole Genome Sequencing Data". In: *Frontiers in Genetics* 10 (2019). ISSN: 1664-8021.

[32] F. Pedregosa et al. "Scikit-Learn: Machine Learning in Python". In: *Journal of Machine Learning Research* 12 (2011), pp. 2825–2830.

[33] Davide Chicco and Giuseppe Jurman. "The Advantages of the Matthews Correlation Coefficient (MCC) over F1 Score and Accuracy in Binary Classification Evaluation". In: *BMC genomics* 21.1 (Jan. 2, 2020), p. 6. ISSN: 1471-2164. DOI: 10.1186/s12864-019-6413-7.

[34] Glenn W Brier. "Verification of Forecasts Expressed in Terms of Probability". In: 78.1 (Jan. 1950), pp. 1–3. DOI: 10.1175/1520-0493(1950)078<0001:VOFEIT>2.0.CO;2.

[35] Mark D. Wilkinson et al. "The FAIR Guiding Principles for Scientific Data Management and Stewardship". In: *Scientific Data* 3.1 (1 Mar. 15, 2016), p. 160018. ISSN: 2052-4463. DOI: 10.1038/sdata.2016.18.

[36] PDBe-KB consortium. "PDBe-KB: A Community-Driven Resource for Structural and Functional Annotations". In: *Nucleic Acids Research* 48.D1 (Jan. 8, 2020), pp. D344–D353. ISSN: 0305-1048. DOI: 10.1093/nar/gkz853.