Predicting Fluoroquinolone Resistance in *Mycobacterium tuberculosis*

Alice E. Brankin

St. Hugh's College,

University of Oxford

A thesis submitted for the degree of

Doctor of Philosophy

Trinity Term 2022

Word Count: 42,002

Table of Contents

Abstract

Tuberculosis killed an estimated 1.5 million people in 2020 and the spread of drug-resistant strains is an increasing threat. Fluoroquinolones are among the safest and most effective drugs used to treat drug-resistant tuberculosis, but fluoroquinolone resistance has emerged. Rapid resistance diagnosis is key to provide patients with effective treatment and monitor the spread of resistant strains. Sequence-based tools, including whole genome sequencing (WGS), provide a promising solution whereby known resistance associated mutations can be quickly detected and used to infer resistance. Several catalogues of mutations with known associations have now been produced, however, these catalogues will never be exhaustive owing to the inevitable emergence of novel and rare resistant mutations.

In this thesis I explore a geographically diverse dataset of thousands of *Mycobacterium tuberculosis* isolates with WGS and phenotypic resistance measurements, to improve our understanding of patterns associated with fluoroquinolone resistance. I investigate the reliability of the assumptions of current sequence-based diagnostics and show how resistance patterns can affect the performance of predictive tools. I apply and evaluate two methods, that utilise the chemical and structural interactions resulting from genetic mutations, to predict the effects of mutations on fluoroquinolone resistance; machine learning algorithms and free energy calculation.

I find that fluoroquinolone resistance is widespread and the associations with resistance are complex; genetic and geographic background, resistance to other antitubercular drugs and minor populations with resistance associated mutations are important. I show that

structure-based predictive methods, and the machine learning approach particularly, can successfully predict fluoroquinolone resistance and susceptibility. Such tools could increase the success of resistance prediction from WGS data by complimenting the catalogue-based predictive approach and predicting the effects of novel mutations. Overall, this work shows that genetics-based predictive diagnostics have the potential to provide personalised, effective treatment regimens for tuberculosis.

Preface

Acknowledgements

First and foremost, I would like to thank Dr. Philip Fowler and Prof. Ann Sarah Walker who have been outstanding supervisors, generous with both their time and wisdom. I express particular thanks to Phil, for his unwavering generosity and understanding during the pandemic and for inspiring and facilitating invaluable experiences outside of my study. I am thankful to all those involved in the CRyPTIC project, for their enormous contribution to the field of antimicrobial resistance in tuberculosis and without whom this work would not have been possible. I am especially grateful to Dr. Kerri Malone, for such a supportive and enjoyable collaboration. I would also like to thank the Modernising Medical Microbiology group for a wonderful time in Oxford and for their helpful advice during my study. There are so many people that have been imperative for me to have reached this point in my scientific career, but I am especially thankful to my parents, Richard and Lesley, for their unconditional love and support. My final big thank you is to Alexander Winch, for his patience, encouragement and fun over the last four years.

Funding

This work was funded by an NDM Prize Studentship from the Oxford Medical Research Council Doctoral Training Partnership and the Nuffield Department of Clinical Medicine. The computational aspects of this research were supported by the Wellcome Trust Core Award Grant Number 203141/Z/16/Z and the National Institute for Health Research (NIHR) Oxford Biomedical Research Centre (BRC). Access to the ARCHER supercomputer was provided through CompBioMed, an EU H2020 Centre of Excellence project, and the UK High-End

Computing Consortium for Biomolecular Simulation (HECBioSim, EP/R029407/1) which is supported by the Engineering and Physical Sciences Research Council (EPSRC).

Declaration

I, Alice Brankin, confirm that this thesis is my own work. I designed and conducted all of the analyses presented and wrote the thesis with appropriate support from my supervisors and collaborators. I here outline any specific support and contributions I have been grateful to receive from others.

In relation to chapters 3, 4 and 5:

The CRyPTIC Consortium partner laboratories collected all the *Mycobacterium tuberculosis* isolates, and collected the associated phenotypic and whole genome sequencing data. Citizen scientists read minimum inhibitory concentrations from plate photographs. Members of the CRyPTIC Analysis Group centrally processed all the phenotypic and sequencing data from the partner laboratories and calculated epidemiological cut-off values to segregate resistant and susceptible isolates.

In relation to chapters 3 and 6:

Chapters 3 and 6 are based on published work. I co-wrote the manuscript[1] on which chapter 3 is based with Dr. Kerri Malone but confirm that all the analysis, writing, and figures for the thesis chapter were conducted and produced by myself. I wrote the manuscript[2] on which chapter 6 is based and I confirm that all the analysis, writing, and figures for the thesis chapter were conducted and produced by myself. The .mdp files used for energy minimisation, equilibration and molecular dynamics simulations were provided by Dr. Philip Fowler, and edited by myself.

Publications

Relating to this thesis:

Chapter 3: **The CRyPTIC Consortium**, A data compendium associating the genomes of 12,289 *Mycobacterium tuberculosis* isolates with quantitative resistance phenotypes to 13 antibiotics. *PLOS Biology* **2022** 20(8), e3001721.

Chapter 6: **Brankin, A. E.**; Fowler, P. W., Predicting antibiotic resistance in complex protein targets using alchemical free energy methods. *Journal of Computational Chemistry* **2022**, *43* (26), 1771.

Manuscripts based on the results of chapters 4 and 5 are also being prepared.

Other work:

Brankin, A. E.; Fowler, P. W., Predicting Resistance Is (Not) Futile. *ACS Central Science* **2019,** *5* (8), 1312-1314.

Brankin, A.; Seifert, M.; Georghiou, S. B.; Walker, T. M.; Uplekar, S.; Suresh, A.; Colman, R. E., In silico evaluation of WHO-endorsed molecular methods to detect drug resistant tuberculosis. *Nature Scientific Reports* **2022**, *12*, 17741.

Young, B. C.; Bush, S. J.; Lipworth, S.; George, S.; Dingle, K. E.; Sanderson, N.; **Brankin, A**.; Walker, T.; Sharma, S.; Leong, J.; Plaha, P.; Hofer, M.; Chiodini, P.; Gottstein, B.; Furrer, L.; Crook, D.; Brent, A., Modern Solutions for Ancient Pathogens: Direct Pathogen Sequencing for Diagnosis of Lepromatous Leprosy and Cerebral Coenurosis. *Open Forum Infectious Diseases* **2022**, ofac428.

Walker, T. M.; Miotto, P.; Köser, C. U.; Fowler, P. W.; Knaggs, J.; Iqbal, Z.; Hunt, M.; Chindelevitch, L.; Farhat, M. R.; Cirillo, D. M.; Comas, I.; Posey, J.; Omar, S. V.; Peto, T. E. A.; Suresh, A.; Uplekar, S.; Laurent, S.; Colman, R. E.; Nathanson, C.-M.; Zignol, M.; Walker, A. S.; Crook, D. W.; Ismail, N.; Rodwell, T. C.; Walker, A. S.; Steyn, A. J. C.; Lalvani, A.; Baulard, A.; Christoffels, A.; Mendoza-Ticona, A.; Trovato, A.; Skrahina, A.; Lachapelle, A. S.; **Brankin, A**.; *et al.*, The 2021 WHO catalogue of Mycobacterium tuberculosis complex mutations associated with drug resistance: a genotypic analysis. *The Lancet Microbe* **2022**, *3* (4), e265.

The CRyPTIC Consortium, Epidemiological cutoff values for a 96-well broth microdilution plate for high-throughput research antibiotic susceptibility testing of M. tuberculosis. *European Respiratory Journal* **2022**, 2200239.

The CRyPTIC Consortium, Genome-wide association studies of global Mycobacterium tuberculosis resistance to 13 antimicrobials in 10,228 genomes identify new resistance mechanisms. *PLOS Biology* **2022,** *20* (8), e3001755.

Fowler, P. W.; Wright, C.; Spiers, H.; Zhu, T.; Baeten, E. M. L.; Hoosdally, S. W.; Gibertoni Cruz, A. L.; Roohi, A.; Kouchaki, S.; Walker, T. M.; Peto, T. E. A.; Miller, G.; Lintott, C.; Clifton, D.; Crook, D. W.; Walker, A. S.; The Zooniverse Volunteers; **The CRyPTIC Consortium**, A crowd of BashTheBug volunteers reproducibly and accurately measure the minimum inhibitory concentrations of 13 antitubercular drugs from photographs of 96-well broth microdilution plates. *eLife* **2022,** *11*, e75046.

List of figures

List of tables

Abbreviations

ATP	Adenosine triphosphate
AMI	Amikacin
AMR	Antimicrobial resistance
AUC	Area under curve
BAR	Bennet's acceptance ratio
BCG	Bacillus Calmette–Guérin
BDQ	Bedaquiline
CFZ	Clofazimine
COVID-19	Coronavirus disease 2019
CRyPTIC	Comprehensive resistance prediction for tuberculosis: an international consortium
DLM	Delamanid
DNA	Deoxyribonucleic acid
DST	Drug susceptibility test
ECOFF	Epidemiological cut off
EMB	Ethambutol
ETH	Ethionamide
FEP	Free energy perturbation
FQ	Fluoroquinolone
FRS	Fraction read support
HIV	Human immunodeficiency virus
Hr	Isoniazid resistant
INH	Isoniazid
KAN	Kanamycin
LEV	Levofloxacin
LJ	Löwenstien-Jensen
LPA	Line probe assay
LR	Logistic regression
LZD	Linezolid
MD	Molecular dynamics
MDR	Multi drug resistant
MGIT	Mycobacterial growth indicator tubes
MIC	Minimum inhibitory concentration
MTC	Mycobacterium tuberculosis complex
MXF	Moxifloxacin
NGS	Next generation sequencing

NRD	New or repurposed drug
OAT	Ordinal all threshold
OIT	Ordinal intermediate threshold
OR	Ordinal ridge
OSE	Ordinal squared error
PAS	Para-aminosalicylic acid
PCR	Polymerase chain reaction
PDB	Protein data bank
PZA	Pyrazinamide
QRDR	Quinolone resistance determining region
RBFE	Relative binding free energy
RFB	Rifabutin
RFC	Random forest classifier
RIF	Rifampicin
RMR	Rifampicin mono-resistant
RMSD	Root mean squared deviation
RNA	Ribonucleic acid
ROC	Receiver operating characteristic
RR	Rifampicin resistant
SEM	Standard error of the mean
TB	Tuberculosis
TDR	Totally drug resistant
TPT	Tuberculosis preventative treatment
TI	Thermodynamic integration
tNGS	Targeted next generation sequencing
WGS	Whole genome sequencing
WHO	World health organisation
XDR	Extensively drug resistant
XGBC	Extreme gradient boost classifier

1 Chapter 1: Introduction

Worldwide, infectious disease is at the forefront of the lives of many families, where a disproportionately high burden remains a constant threat for those with low incomes[3, 4]. The COVID-19 pandemic has re-alerted the western world to the catastrophe posed by an infectious disease outbreak; between January 2020 and December 2021, COVID-19 was estimated to be responsible for 18.2 million deaths[5]. Rapid research has enabled effective vaccines to be developed and produced and treatments identified, reducing the impact of the COVID-19 pandemic, and allowing our lives here in the United Kingdom to slowly return to normal. But as a result of the pandemic, there have been disastrous consequences for the control of tuberculosis[6] (TB); a deadly infectious disease most prevalent in low- and middle-income countries that, prior to COVID-19, was the biggest killer from infectious disease.

As a result of COVID-19, we are now also looking to how and where the next global pandemic could emerge, but a silent pandemic is already upon us in the form of antimicrobial resistance (AMR). Since the discovery of penicillin almost a century ago, antimicrobial drugs have become essential for treating common infections, preventing infections after injury or during surgery, and protecting patients with compromised immune systems (for example during cancer chemotherapy). The entire foundation of modern medicine is threatened by the spread of infectious pathogens that no longer respond to these drugs. It is vital that we monitor this silent pandemic and take action to ensure that our defences against infectious diseases remain viable.

My objective in this thesis, 'Predicting fluoroquinolone resistance in *Mycobacterium tuberculosis*', is to increase our understanding of, and provide tools for diagnosing resistance to fluoroquinolone antimicrobials in TB. The following introductory chapter provides a background to TB, drug resistant TB, diagnosis of drug resistance, the fluoroquinolone class of antibiotics and an outline of the subsequent thesis.

1.1 Tuberculosis

Tuberculosis (TB) is an infectious disease that has affected humans for thousands of years. The first known written descriptions of the disease date from 3,300 years ago[7] but there is evidence that the common causative agent of the disease, the bacterium *Mycobacterium tuberculosis*, has infected humans as long as 9,000 years ago[8]. Around a quarter of the world's population are now estimated to have a latent TB infection[9], where no disease symptoms are present, and the infection cannot be spread. Although not life threatening in this form, there is a 5-10% risk of progression to active disease[10] and immunocompromised patients are even more at risk[11]. Around 10 million people were estimated to have developed active TB in 2020[12]. The most common form is pulmonary TB which is characterised by symptoms including a chronic cough with bloody mucus, weight loss, excessive sweating and fever, but TB can infect any part of the body. Of those with active TB, 1.5 million died from the disease in 2020[12], making TB the biggest killer from infectious disease worldwide, second only to the recent COVID-19 pandemic. Although TB occurs in every part of the world, 95% of deaths were reported in low- and middle-income countries and 86% of new cases were from just 30 countries[12].

Active TB typically spreads from person to person through *M. tuberculosis* contaminated airborne water droplets, for example when a person with active disease coughs or sneezes. The infectious cycle of *M. tuberculosis* is shown in Figure 1. Once the bacteria are inhaled by another host and reaches the lungs, it is phagocytosed by macrophages in the alveoli where normally, the human immune response kills the bacteria[13]. If the bacteria survive, they can begin clonally replicating within the macrophage, diffuse to nearby cells and grow exponentially within the lungs or spread through the blood or lymphatic systems. The host immune response mobilises neutrophils and lymphocytes to the sites of infection that form granulomas to prevent further spread of the infection, probably by starving the bacteria of oxygen and promoting persistence in a dormant state[14], i.e. latency. However, the immune response may be insufficient, or the structure of the granuloma can decompose resulting in rapid bacterial growth and the destruction of lung tissue. This constitutes active TB disease: symptomatic, contagious, and requiring antibiotic treatment.

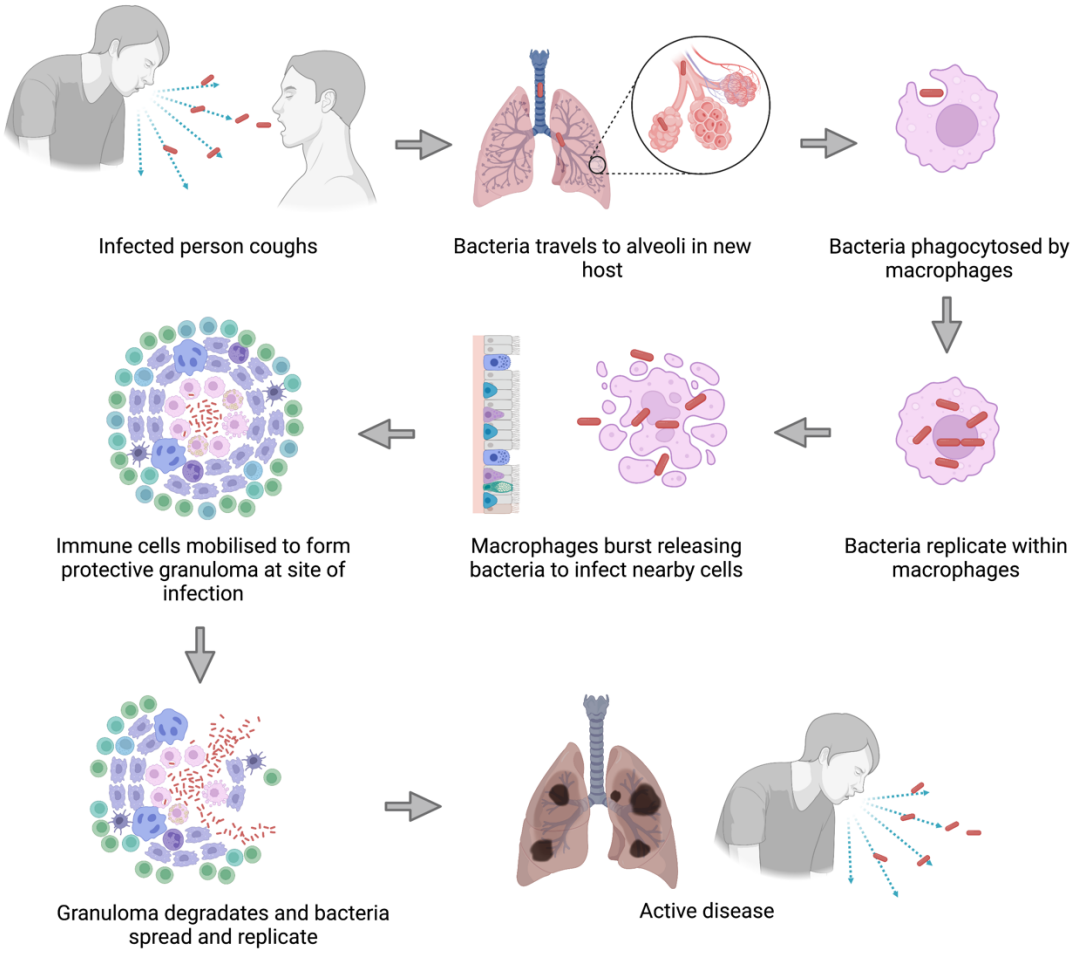

Figure 1 An example of progression of *M. tuberculosis* infection to active tuberculosis disease.

Despite giving no clinical symptoms in cases of latent infection, some *M. tuberculosis* bacteria are metabolically active[15] and can be found in tissues outside of granulomas[16-18]. A proposed explanation is that during latent infection most of the bacteria persist in a dormant state with a few bacteria actively replicating[19]. The replicating bacteria are killed by the immune system but are replenished by the reservoir of dormant TB. If the immune system is unable to kill the actively replicating bacteria (for example if the patient has a compromised immune system), uncontrolled replication is promoted resulting in a switch to active disease[20].

1.1.1 Mycobacterium tuberculosis

Human TB can be caused by *M. tuberculosis*, *M. africanum* and *M. canetti*[21], as well as other zoonotic members of the *M. tuberculosis* complex (MTC); *M. bovis*, *M. caprae* and *M. orygis* found in bovine animals, *M. microti* found in rodents and *M. pinnipedii* found in seals[22-26]. The main causative agent of human TB, the *M. tuberculosis* bacteria, was discovered by Robert Koch in 1882[27], and is a small, slow growing, aerobic bacillus that divides only once every 18-24 hours forming a cord-like colony structure. The complete genome sequence of the well characterized H37Rv *M. tuberculosis* reference strain is 4.4 million base pairs in length and is estimated to encode around 4,000 genes[28].

There are several key differences between *M. tuberculosis* and non-Mycobacterial pathogens. Pathogenic bacteria are often characterised as Gram negative or Gram positive based on the structure of their cell wall, however Mycobacteria do not respond to Gram staining because their unusual cell wall is composed of mycolic acid lipids. These lipids are essential for the survival and pathogenesis of TB-causing Mycobacteria; they help prevent dehydration and support growth in macrophages[29]. Compared to other bacteria, the *M. tuberculosis* genome has a high content of guanine and cytosine nucleotides (> 65% of the genome) and the bacterium significantly favours the amino acids Alanine, Glycine, Proline, Arginine and Tryptophan which are encoded by GC rich codons[28]. The high GC content could be a useful trait to survive in the warm conditions of the human host; GC pairs are more thermally stable than AT in structured nucleic acids[30] and there is a positive corelation between the GC content of bacteria and their optimal growth temperature[31].

Whole genome sequencing data used to reconstruct the evolutionary history of MTC suggests that it emerged around 70,000 years ago and accompanied the migration of humans out of Africa[32]. Study of the global genetic population structure of MTC shows that *M. tuberculosis* and *M. africanum* form seven distinct phylogenetic lineages that have locally adapted to, and likely evolved with, sympatric human populations[33-35], making lineage a proxy for geography in many cases[36]. Lineage 2 (East Asian) and Lineage 4 (Euro-American) are the most widespread, and together with Lineage 3 (East-African-Indian) form the "modern" clade that is characterised by a thousands of years old genetic deletion that increased virulence[37, 38]. The "ancient" lineages are more endemic, with recently discovered Lineage 7 generally restricted to Ethiopia[39] and *M. africanum* (Lineages 5 and 6) generally restricted to West Africa (with rare cases outside the region being in West African migrants)[40]. Lineage 1 is slightly more widespread, with the highest burden in South and South East Asia and East Africa[41]. As human migration has become more widespread, we now begin to see more infections outside of their historical geographical regions[33, 42].

1.1.2 Diagnosis of tuberculosis

Sputum culture is still widely considered the reference standard for diagnosis of active TB infection[43]. Here a patient sample may be cultured using Löwenstein–Jensen (LJ) or Middlebrook 7H10/11 solid agar medium, which can take up to 8 weeks to confirm diagnosis[44]. A quicker turn-around time is possible when using a liquid-based culture system such as for Mycobacteria Growth Indicator Tubes (MGIT). MGIT use a Middlebrook 7H9 Broth to culture the bacteria and a fluorescent indicator to detect growth. Despite the faster turn-around time compared to use of solid LJ media for culture, use of MGIT is more costly and can be more prone to contamination[45].

In countries with a high burden of TB, culture based diagnosis may not be affordable and diagnosis is instead most commonly performed using sputum smear microscopy[46]. The technique is relatively low cost and has fewer equipment requirements than culture[46], however it has lower diagnostic sensitivity[47]. In low-income countries many TB cases are still not bacteriologically confirmed, and diagnosis may rely on presented symptoms and chest X-rays to detect lung lesions; this kind of diagnosis has been shown to have low (< 70%) sensitivity and specificity for the diagnosis of TB[48].

Recently, the WHO updated guidelines to include recommendations for the use of PCR-based molecular tests that detect genetic regions specific to *M. tuberculosis* as a primary diagnostic tool for active TB[49]. These tests are recommended as they are accurate, and faster and easier to perform than culture[49]. Several molecular diagnostic tests have the added benefit of being able to predict drug resistance, which is discussed later in Section 1.3.2. The cost of using PCR-based tests can be cheaper in comparison to culture-based methods[50], although some tests are too costly for widespread implementation in low and middle income countries[51].

1.1.3 Tuberculosis treatment and prevention

Despite the grim statistics of global burden and deaths, TB is both curable and preventable. Antibiotics that target a variety of essential *M. tuberculosis* biochemical pathways can be used to treat TB and many target processes involved in the manufacture of the unusual mycolic acid rich cell wall (Table 1). The first antibiotic treatment for TB, streptomycin, was

first used in 1944[52] and further antibiotic development during the decade facilitated the successful use of a "triple therapy" comprised of streptomycin, para-aminosalicylic acid (PAS) and isoniazid in the 1950s[53]. However, PAS was difficult to tolerate and the treatment regimen took two years to complete[54]. PAS was later replaced by ethambutol as it was better tolerated, reducing treatment duration to 18 months[55]. Treatment time was then halved in the 1970s by the introduction of isoniazid and rifampicin in combination, and finally in the 1980s the addition of pyrazinamide reduced treatment duration to six months[53]. The 1980s saw a range of other drugs introduced for the treatment of TB including the fluoroquinolones[56], but the four drugs, isoniazid, pyrazinamide, ethambutol and rifampicin, still comprise the most common treatment regimen for drug-susceptible TB today. Since the 1980s no new drug classes were developed and approved for treating TB, until 2012 and 2014 when bedaquiline and delamanid were introduced[57, 58] and by the end of 2020, bedaquiline had been used to treat TB in 109 countries[12]. Other advances in TB treatment options include repurposing the drugs linezolid and clofazimine originally used for the treatment other *Mycobacterial* infections such as leprosy[59, 60].

Drug	Target(s)	Mechanism
Rifamycins: rifampicin (rifampin), rifabutin, rifapentine	RNA polymerase (RpoB)	Inhibits RNA synthesis
Isoniazid	enoyl acyl carrier protein reductase (InhA)	Inhibits cell wall synthesis
Thioamides: ethionamide, prothionamide	enoyl acyl carrier protein reductase (InhA)	Inhibits cell wall synthesis
Ethambutol	Glutamate racemase (MurI)[61]	Inhibits cell wall synthesis
Pyrazinamide	Aspartate decarboxylase (PanD)[62]	Inhibits coenzyme A synthesis
Fluoroquinolones: levofloxacin, moxifloxacin	DNA gyrase (GyrA/GyrB)	Inhibits DNA ligation and synthesis
Aminoglycosides: amikacin, streptomycin, capreomycin*, kanamycin*	16S ribosomal RNA	Inhibits protein synthesis
Bedaquiline	F-ATP synthase (AtpE)[63]	Inhibits ATP synthesis
Clofazimine	Outer membrane, respiratory chain and ion channels[64]	Respiratory poisoning, membrane disruption
Nitroimidazole pro-drugs: delamanid, pretomanid	Mycolic acid biosynthesis[65] and respiratory chain[66]	Inhibits cell wall synthesis, respiratory poisoning
***para*-aminosalicyclic acid**	Dihydrofolate reductase (DfrA)[67]	Inhibits folate synthesis
Carbapenems: imipenem-cilastatin, meropenem,	Penicillin binding protein 3 (FtsL)	Inhibits cell wall synthesis
Linezolid	23S Ribosomal RNA	Inhibits protein synthesis
Cycloserine, terizidone	alanine racemase (Alr) and D-alanine:D-alanine ligase (DdlA)	Inhibits cell wall synthesis

Table 1 Drugs used to treat tuberculosis. The molecular targets are not yet elucidated for clofazimine and nitroimidazoles. *Capreomycin and kanamycin are no longer recommended in treatment programs[68]

The standard treatment for drug susceptible TB today therefore consists of a six month regimen using daily doses of four drugs; isoniazid, rifampicin, pyrazinamide and ethambutol for two months and isoniazid plus rifampicin for four months[69]. Shorter regimens can be used, but these have special considerations. Using the same four drugs as the standard regimen, patients under 16 can be treated over four months; isoniazid, rifampicin, pyrazinamide and ethambutol for two months and isoniazid plus rifampicin for two months[69]. Another four-month regimen consisting of rifapentine, isoniazid, pyrazinamide and moxifloxacin, may also be used to treat drug susceptible TB but fluoroquinolone drug susceptibility testing is recommended and the cost of this treatment programme is

significantly more expensive than the standard regimen due the inclusion of rifapentine[69]. In the case a patient cannot be treated using one of these standard regimens, an individualised treatment regime can be devised based on priority groupings of other antitubercular drugs[68, 70] (Table 2).

Group A	Levofloxacin/moxifloxacin, bedaquiline, linezolid
Group B	Clofazimine, cycloserine/terizidone
Group C	Ethambutol, delamanid, pyrazinamide, imipenem-cilastatin/meropenem, amikacin/streptomycin, *p*-aminosalicylic acid, ethionamide/prothionamide

Table 2 Priority groupings of antitubercular drugs. Agents are ranked according to their effectiveness and safety profile[68, 70], group A drugs should be prioritised, followed by group B, then group C.

In addition to treatment there are several prevention measures that can reduce the incidence of TB, including TB preventative treatment (TBT), vaccination, and addressing of poverty-associated risk factors. TBT usually consists of 6-9 months of daily isoniazid and can be offered to anyone who has been in close contact with someone who has active TB or people with a compromised immune system, such as HIV infection[71]. TBT has been shown to reduce the progression of latent TB to active TB by up to 62% in HIV infected individuals[72-74], however the protection from TBT may wane soon after treatment completion[74]. Another preventative option is vaccination: the Bacille Calmette-Guérin (BCG) vaccine, based on live attenuated *M. bovis*, was first introduced over 100 years ago and remains effective at reducing the risk of active TB infection in children by around 74%[75]. However, the vaccine does not offer significant protection in older individuals[75] and the level of protection reduces to zero between ten and fifteen years after childhood vaccination[76]. Encouragingly, recent tests of a new vaccine against TB, based on two *M. tuberculosis* antigen proteins, significantly reduced the progression of latent TB to active disease and had 50% efficacy in

clinical trials[77]. Finally malnutrition, air pollution and overcrowded living conditions are significantly associated with TB prevalence[78] and social initiatives to address these problems can help significantly reduce TB incidence.

1.2 Drug resistant tuberculosis

Soon after the first introduction of antibiotic treatment for TB, patients were found to have *M. tuberculosis* bacilli that had acquired resistance to the treatment[79], and the phenomenon was eventually reduced by the introduction of combination therapy. Despite use of combination therapies for all TB treatment today, drug resistance remains a concern[6]. Any combination of mono- or poly-drug resistance can occur in TB, but the World Health Organisation (WHO) have defined several clinically important resistance phenotypes (Table 3). Of these, multi-drug resistance (MDR), which comprises resistance to the two most potent first-line drugs, isoniazid and rifampicin, is the most common; for patients with no prior history of TB, an estimated 3-4% of cases are MDR. For previously treated infections the global rate of MDR is 18-21%[12] and for some eastern European countries, such as Belarus, over 50% of previously treated infections have MDR[12, 80]. It is important to be aware that the MDR label encompasses a range of common resistance phenotypes and is not just limited to isoniazid and rifampicin resistance: a study showed that 40% of the MDR strains in Germany had resistance to all four first line drugs and streptomycin[81]. The spread of MDR strains is of particular concern because of its association with an increased risk of death[82] and the more resistant a strain, the higher the association[83].

Phenotype	Resistant to at least:
Rifampicin susceptible, isoniazid resistant (Hr TB)	Isoniazid but not rifampicin
Rifampicin resistant (RR)	Rifampicin
Multidrug resistant (MDR)	Rifampicin and isoniazid
Pre extensively drug resistant (pre-XDR)	Rifampicin and a fluoroquinolone
Extensively drug resistant (XDR)	Rifampicin, a fluoroquinolone and another group A agent (currently bedaquiline or linezolid)

Table 3 WHO defined *M. tuberculosis* resistance phenotypes. Reference[84].

1.2.1 Mechanisms of drug resistance in *M. tuberculosis*

Generally, acquired drug resistance is attributed to single nucleotide polymorphism (SNPs) or genetic insertions or deletions in either the drug targets or pro-drug activators[85]. For example, rifampicin resistance is most commonly conferred by a mutation of Serine to Leucine at amino acid position 450 in the target gene *rpoB*[86] and isoniazid resistance is most commonly conferred by a Serine to Threonine mutation at amino acid position 315 in *katG*, which encodes an enzyme that converts isoniazid into the active form[87]. Genetic mutations in promotor regions can also confer resistance by decreasing target expression, for example low-level resistance to isoniazid can be conferred by several mutations in the promoter of the antibiotic target *inhA*[88].

More general mechanisms of drug resistance relating to intracellular accumulation of the drug are also at play in *M. tuberculosis*. The *M. tuberculosis* genome contains a range of putative efflux pumps and transporters[89] for which increased expression or mutation are associated with increased resistance to a range of drugs including bedaquiline, clofazimine, isoniazid, streptomycin, amikacin, ethambutol, pyrazinamide and fluoroquinolones[89]. In addition to exporting drugs from the cytoplasm, *M. tuberculosis* can reduce the amount of drug molecules that enter the cell to increase the level of resistance. The mycolic acid rich

cell wall is hydrophobic which provides an intrinsic defence against the entry of hydrophilic drug molecules[90, 91] and *Mycobacteria* can also adapt to decrease the expression of porins to further prevent entry of hydrophilic drugs[91].

1.2.2 Evolution and transmission of drug resistance in *M. tuberculosis*

Isoniazid resistance tends to evolve prior to other drug resistance and almost always before rifampicin resistance[92]. Interestingly, this is not because of its earlier introduction into clinics, but could instead be due to the most common isoniazid conferring resistance mutation, *katG* S315T, not being associated with a fitness cost[92, 93]. This may also explain why Hr TB is much more common than RR TB[94]. Studies of historical TB samples from South Africa have suggested that after acquisition of isoniazid resistance, there is generally a stepwise acquisition of resistance to ethambutol, rifampicin (leading to MDR) and pyrazinamide[95]. Then, during MDR treatment, resistance to fluoroquinolones and other second-line drugs arises[96] and, despite their only recent introduction to clinics, resistance to the new drugs bedaquiline and delamanid have now been observed, both separately and together[97, 98].

Resistance and MDR has arisen independently in all lineages and countries investigated[1, 92] but it has been shown that Lineage 2 isolates are particularly associated with resistance[96, 99] and the transmission of resistance[99, 100]. In addition to genetic factors, the causes of an increase in local resistant TB levels may be different in different areas. For example the average ancestral age of resistance was found to be older in wealthier nations, suggesting that, in this setting, more resistance arises from transmission of strains from person to person rather than the emergence of resistance in patients during treatment

(amplification)[101]. Amplification frequently occurs if treatment courses are not adhered to or if the treatment courses are short and repeated[102, 103], which may be more likely in resource limited settings.

Interestingly, resistance doesn't always occur in the entire *M. tuberculosis* population in the host; mixed populations that have both wild-type and resistance-associated alleles at a particular locus can co-exist. Mixed *M. tuberculosis* infections are relatively common and can either be from transmission of multiple strains or evolution of an initially homogeneous infecting strain within the host[104]. In the latter instance, over time during treatment, even a very minor population with a resistance associated allele may completely outcompete the wild-type population[105, 106], but this is not always the case as the strains are also selected based on their fitness[105]. It is also suggested that within host evolution of part of the population may occur due to drugs having different penetration profiles in different areas of the lung[107]. This phenomenon, often termed 'hetero-resistance' in the literature, has been implicated in resistance to many antitubercular drugs including isoniazid, rifampicin, ethambutol, fluoroquinolones and bedaquiline[101,108].

1.2.3 Treatment of drug resistant tuberculosis

The WHO treatment recommendations were recently updated to advocate the use of a shorter (6-9 month) all-oral treatment programme that includes bedaquiline for drug resistant TB[109]. The clinically important TB resistance phenotypes (Table 3) are used to help direct the treatment a patient with resistant TB should be put on and the current recommendations for each phenotype are presented in Table 4. Alarmingly, it is estimated that only 38% of RR or MDR patients were enroled in an appropriate treatment regimen in

2019[110] and, despite yearly increases in treatment success for rifampicin resistant (RR) or MDR cases, the current level of treatment success sits at just 59%[6]. This is much lower than for drug susceptible TB[6], but it is hoped that these new recommendations may help increase the level of successful treatment for resistant infections.

Phenotype	Current treatment recommendations:
Rifampicin susceptible, isoniazid resistant (Hr TB)	6 month regimen comprising rifampicin, ethambutol, pyrazinamide and levofloxacin[69]
Rifampicin resistant (RR)	6-month regimen, comprising bedaquiline, pretomanid, linezolid and moxifloxacin (BPaLM)[109] OR 9-month regimen comprising bedaquiline (used for 6 months), in combination with levofloxacin/moxifloxacin, ethionamide, ethambutol, isoniazid (high-dose), pyrazinamide and clofazimine for 4 months (with the possibility of extending to 6 months if the patient remains sputum smear positive at the end of 4 months); followed by 5 months of treatment with levofloxacin/moxifloxacin, clofazimine and ethambutol. Ethionamide can be replaced by 2 months of linezolid[109] OR Individualised longer regimen for patients who are not eligible for or have failed shorter treatment regimens designed using the priority grouping of medicines recommended in current WHO guidelines[109]
Multidrug resistant (MDR)	6-month (BPaLM) regimen, comprising bedaquiline, pretomanid, linezolid and moxifloxacin[109] OR 9-month regimen comprising 6 months bedaquiline, in combination with levofloxacin/moxifloxacin, ethionamide, ethambutol, isoniazid (high-dose), pyrazinamide and clofazimine for 4 months, followed by 5 months of treatment with levofloxacin/moxifloxacin, clofazimine and ethambutol. 4 months Ethionamide can be replaced by 2 months of linezolid[109] OR Individualised longer regimen for patients who are not eligible for or have failed shorter treatment regimens designed using the priority grouping of medicines recommended in current WHO guidelines[109]
Pre extensively drug resistant (pre-XDR)	6-month BPaLM regimen without moxifloxacin[109] OR Individualised longer regimen for patients who are not eligible for or have failed shorter treatment regimens designed using the priority grouping of medicines recommended in current WHO guidelines[109]
Extensively drug resistant (XDR)	Individualized longer regimen designed using the priority grouping of medicines recommended in current WHO guidelines[109]

Table 4 Current WHO treatment recommendations for drug resistant TB.

1.3 Diagnosis of drug susceptibility and resistance in *M. tuberculosis*

In order to reduce TB incidence by 80% and deaths by 90%, meeting END TB goals[111], the World Health Organisation (WHO) identifies universal drug susceptibility testing (DST) as a key component[111, 112]. The proportion of patients having DST for rifampicin in particular is increasing, and 71% of patients with confirmed TB in 2020 were tested for rifampicin resistance[12]. Increased use of DST is also important for resistance surveillance, so that outbreaks can be monitored and prevented from spreading further. Rapid DST will be particularly important to enable patients to be placed on the most effective and least toxic treatment regimens as soon as possible to increase the chance of successful treatment and prevent the spread of resistant strains. This section will introduce the current methods used for DST; phenotypic DST, molecular diagnostic tests and next generation sequencing (NGS).

1.3.1 Phenotypic drug susceptibility testing

Traditional culture based DST is considered the gold standard DST method by the WHO[113]. The most common phenotypic DST method is the indirect proportion method using solid media, which was first proposed by Canetti *et al*.[114]. In this qualitative method, a culture is used to inoculate media containing a critical concentration of an antitubercular and two 10-fold serial dilutions of the culture are used to inoculate control media. The growth (i.e. the number of colonies corrected for the dilution factor) on the control media is compared with the growth on the drug containing-media and resistance is detected when at least 1% of the control growth is seen in the media containing the critical concentration of the drug. The critical concentrations used are based on clinical evidence and expert opinion[115], although

have been subject to updates[115, 116], and are defined as the lowest concentration of a drug that inhibits growth of 99% of wild-type *M. tuberculosis* strains in the selected media.

The proportion method can also be implemented using liquid culture. In particular, the WHO recommended BACTEC-MGIT-960 system[113] can deliver DST results within 2 weeks and the turn-around time is significantly quicker than for using solid media[117-120]. Briefly, two MGIT tubes are inoculated with the *M. tuberculosis* culture, one diluted 100-fold compared to the other. A predetermined critical concentration[115] of a drug is added to the MGIT tube with the undiluted culture, and both tubes are incubated. Bacterial growth increases the fluorescence of the medium which is monitored by the MGIT 960 instrument and is directly proportional to bacterial growth. Growth in the tube containing the drug is compared with the control growth once growth in the control tube reaches a predetermined threshold. If the growth in the drug-containing tube is equal to or exceeds the growth in the control tube, the isolate is labelled drug resistant, else the isolate is called drug susceptible[121].

Quantitative culture-based dilution methods also can be used, where the amount of resistance an isolate has to a particular drug is interpreted in the form of a minimum inhibitory concentration (MIC), which can be used to inform the critical concentrations used for qualitative methods. The MIC is defined as the lowest concentration (in mg/L) of a drug that completely inhibits visual growth of the *M. tuberculosis* isolate *in vitro*. The MIC will be dependent on the media used, but a reference protocol to measure the MIC using a broth dilution susceptibility test in enriched Middlebrook 7H9 media has recently been validated by the European Committee on Antimicrobial Susceptibility Testing (EUCAST)[122, 123]. After two weeks of pre-culture on solid media, an inoculum of the *M. tuberculosis* sample is

diluted in broth to give two preparations with one being a 100-fold dilution of the other. These are used to each inoculate two control wells on a 96 well plate, and the more concentrated suspension is used to inoculate at least eight wells containing sequential doubling dilutions of antitubercular drug in broth. After 7 to 21 days of plate incubation, and assuming there is adequate growth in the control wells, positive or negative growth in each well can be detected using an inverted mirror and the MIC determined. Several 96 well microtitre plate designs containing different concentrations of different antitubercular drugs have been developed and validated for MIC measurement using broth dilution[124-126].

Due to culturing requirements, all phenotypic DST methods are time-consuming, with the quickest method, MGIT, having an average of 17.9 days turn-around time from delivery of clinical specimens to reporting DST results in a routine lab setting[127], and it could take longer to report DST results for pyrazinamide[113]. A further limitation of phenotypic DST as a universal DST solution is that laboratory testing is not always possible in resource limited settings due to the infrastructure and trained personnel required[113].

1.3.2 Molecular diagnostic tests

By omitting the culturing steps, polymerase chain reaction (PCR)-based molecular diagnostic tests offer a simpler and more rapid solution to detect both the presence of *M. tuberculosis* and drug resistance in comparison to phenotypic methods. TB lends itself to this technology because the majority of drug resistance can be attributed to well characterized genetic mutations, including single nucleotide polymorphisms (SNPs) and insertions or deletions[85]. Several molecular diagnostic tests for drugs including rifampicin, isoniazid, pyrazinamide, aminoglycosides and fluoroquinolones have now been approved by the WHO[49]. The Xpert

MTB/RIF assay is the most widely used globally; it can detect both the presence of TB and RIF resistance in as little as two hours[128], and has facilitated increased testing and resistance surveillance in areas of high TB burden[110, 129].

There are two main flavours of molecular diagnostic tests: line probe assays (LPAs) and real-time PCR technology. LPAs are implemented by amplifying genetic material from clinical samples using PCR, with resistance detected by nitrocellulose hybridised probes in one of two ways; either the amplified material does not bind to wild type probes implying that there is a mutation present, or the amplified material binds to a resistant probe which specifically detects a known resistance conferring mutation. Alternatively for real-time PCR assays such as Xpert MTB/RIF, during PCR molecular beacons fluoresce when bound to a matching wild type sequence; if a particular beacon does not fluoresce this implies the presence of a resistance conferring mutation[130]. More recently, the GeneXpert technology has been combined with melting analysis, whereby specific mutation patterns can be identified using the melting temperatures of sloppy molecular beacon probes, this allows a greater number of targets and specific mutations to be interrogated in the Xpert MTB/XDR assay[131].

However, reliance on singular tests can be problematic as they may not be comprehensive. This is exemplified by the *rpoB* I491F mutation that was behind an MDR outbreak in Eswatini, the mutation is not detected by Xpert MTB/RIF[132] which led to further spread of the RIF resistant strain in South Africa[133]. A further limitation is that tests cannot always identify the exact resistance conferring mutation, which could be important as different mutations can confer higher or lower level resistance which may be an important

consideration when devising a suitable treatment plan[134]. Molecular diagnostic tests can also give false positive resistance diagnosis; this can prevent patients being treated with the most effective drugs and result in either more toxic or last line antitubercular drugs being used unnecessarily. For instance, a mutation that does not confer resistance, *gyrA* A90G, prevents binding of a wild-type probe in the Hain Genotype MTBDRsl v1 and v2 assays leading to a test interpretation of resistant to fluoroquinolones[135, 136]. Several tests also use the proxy of rifampicin resistance for diagnosing MDR, and therefore could result in a high false positive rate of MDR prediction in areas where RMR is prevalent[1]. Finally, there are currently no tests designed to detect resistance to the new and repurposed drugs; bedaquiline, clofazimine, delamanid and linezolid.

1.3.3 Next generation sequencing (NGS) and catalogue-based prediction

Next-generation sequencing (NGS) offers a more comprehensive alternative to molecular diagnostic testing; there is no limitation in the number of regions that can be probed and complete information is available for all the genetic mutations present within said regions. Therefore, one could identify; resistance to many drugs simultaneously, the presence of high- and low-level resistance markers, and resistance patterns that may not be detected by molecular diagnostic tests.

Whole genome sequencing (WGS) has the capability to provide a complete picture of the genetic variation present in a *M. tuberculosis* isolate. WGS is well suited to *M. tuberculosis* in comparison to other bacteria, as *M. tuberculosis* has a relatively small genome and generally does not contain extrachromosomal genetic information on plasmids – drug resistance

exclusively evolves through chromosomal mutations[137]. The plasmid copy number, which is associated with different levels of resistance in other bacteria, is hard to estimate and the presence of multiple plasmids can make the full genome difficult to assemble using short read sequencing. It should be noted however that short read sequencing can struggle to map highly repetitive reads in some regions of the *M. tuberculosis* genome[138, 139].

A description of a typical WGS pipeline[138, 140] is presented in Figure 2. The process begins with decontamination of the clinical sputum sample which is then cultured. The DNA is then extracted, and a library of DNA fragments with oligonucleotide adaptors comprising the whole *M. tuberculosis* genome is prepared. Short read sequencing can then be performed, for example using the Illumina MiSeq™ platform, which amplifies single stranded DNA fragments and uses DNA polymerase to synthesise complimentary strands of DNA to the single stranded targets, emitting a unique fluorescent signal depending on the nucleotide added which is detected by the MiSeq™ system. The sequencing process continues until a desired length of DNA is reached, for Illumina sequencing this can be up to 300 bases, and this fragment DNA is called a read. The process happens in a massively parallel fashion for all the library fragments. Occasionally, there may be a sequencing error because the wrong nucleotide is incorporated or the signal is misinterpreted; typically sequencing error occurs in 0.1-1% of bases sequenced[141].

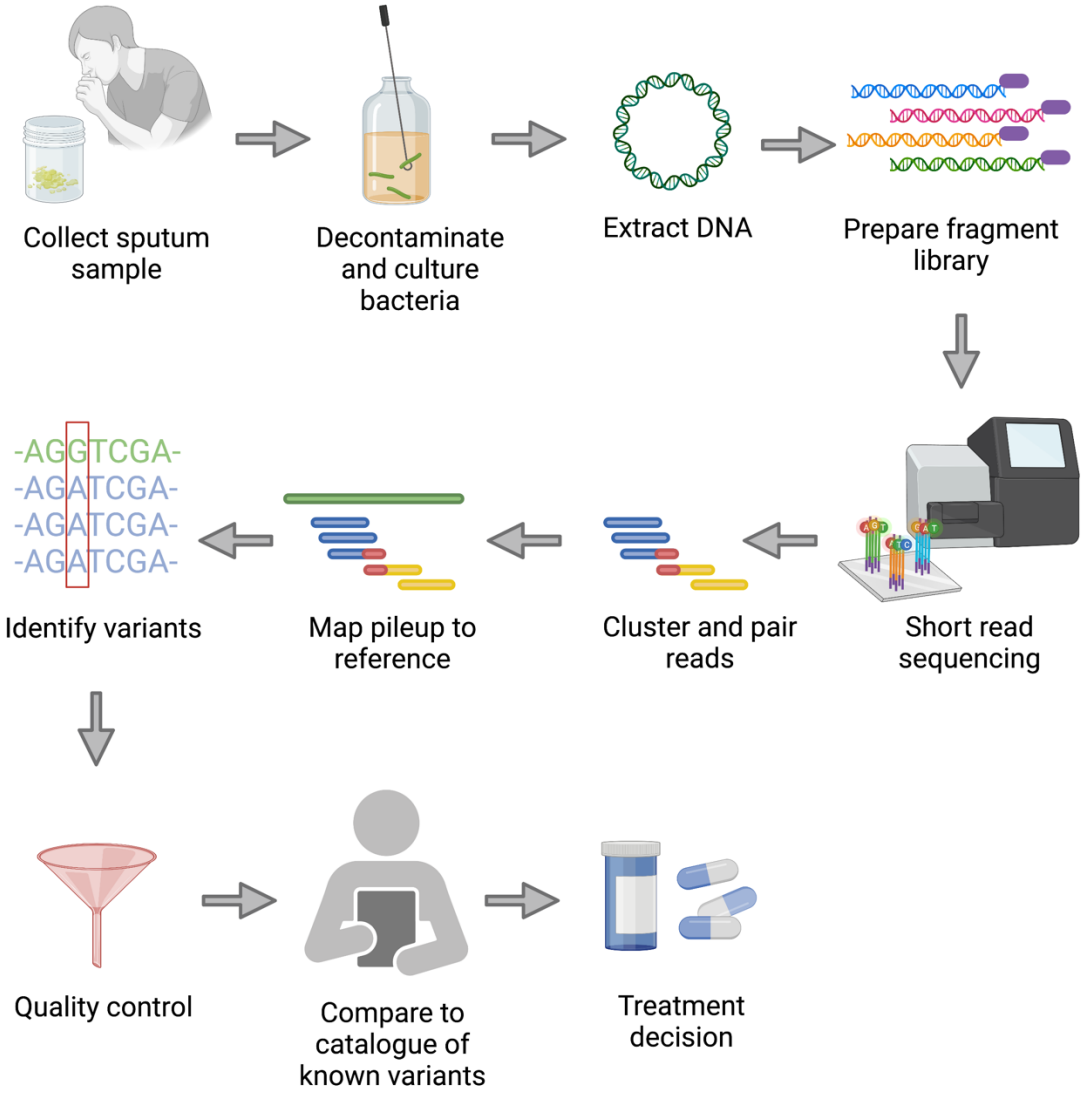

Figure 2 A typical whole genome sequencing pipeline for detecting drug-resistant *M. tuberculosis* from a clinical sample.

The reads for all the fragments are then separated based on their unique adapters, similar sequences are clustered, and matching forward and reverse strands are paired to build a contiguous sequence. The 'pileup' of clustered, ordered reads is then 'mapped' by aligning to a reference sequence, usually H37Rv[142], to detect genomic variants. At this point, quality control filters can be put in place to remove potential false positive variants arising from sequencing error or contamination. These can include only removing variants where there are an insufficient number of reads covering the position in the reference (minimum depth),

or where there is not a sufficiently high proportion of reads supporting one nucleotide at a certain position (fraction of read support). However, these filters can lead to difficulty in identifying variants for *M. tuberculosis* samples that contain more than one strain. The final list of variants in the sample is then compared to a catalogue of known variants associated with resistance to predict whether the sample is susceptible or resistant to each drug. WGS with a catalogue-based approach has been shown to successfully predict resistance and susceptibility to several first and second line antitubercular drugs[143-145], although the approach tends to perform less well for second-line drugs[146].

The major benefit of WGS is that it has a faster turnaround time for DST results compared to phenotypic methods[147] and the costs have greatly reduced in recent years. As such WGS pipelines are used in some high-income countries for *M. tuberculosis* DST, including England, Scotland and Wales where all TB samples now undergo WGS[148]. WGS also has other benefits; it can be used to improve TB surveillance via analysis of global movement and transmission[149, 150] and importantly WGS information is needed to build and enrich catalogues of resistance conferring mutations[42]. However, there are some disadvantages to the WGS pipeline. Firstly there is currently no standardized workflow meaning that genomes collected by different countries may be inconsistent and incomparable[151]. Secondly, similarly to phenotypic DST, the current WGS pipeline has time and infrastructure limitations because *M. tuberculosis* bacilli generally need to be isolated and cultured from the clinical sample prior to sequencing to ensure there is sufficient DNA[152]. With this being said, performing WGS directly from clinical samples is feasible, potentially reducing the turn-around time for DST results to c. 5 days[153]. Another option, targeted NGS (tNGS), can also be implemented directly from clinical samples and has been shown to perform comparably to WGS platforms for identifying resistance conferring mutations[154-156] and the technology

could be more appropriate for resistance diagnosis in resource limited settings[157]. Although tNGS can be used to amplify many genetic regions in order to provide a broad picture of drug resistance present in a TB isolate, less information can be gleaned from the resulting output in comparison to WGS data.

Ultimately, the success of any NGS-based method for resistance prediction relies on the production of comprehensive catalogues of genetic mutations associated with resistance. This requirement has resulted in the collection of large, global, matched genotypic and phenotypic datasets, from which statistical tests, machine learning or genome wide association studies can be used to identify the mutations that are associated with phenotypic resistance and those that are not[1]. Most recently, the WHO has produced a catalogue of over 17,000 such mutations from a collection of 50,306 *M. tuberculosis* isolates using a standardized statistical approach that it recommends for interpretation of TB sequencing data[42, 145]. This catalogue is not exhaustive[42, 145, 158]; indeed no catalogue will ever be exhaustive as novel and rare resistance conferring mutations will arise and when detected are unlikely to meet the statistical thresholds required to be highlighted as associated with resistance. Therefore, novel predictive approaches will always be necessary to compliment the catalogues.

1.4 Fluoroquinolones

The first fluoroquinolone was developed in 1976[159], and the general scaffold is based on a quinoline ring system with a fluorine atom at the C6 position (Figure 3). Additions or modifications can be introduced at several points on the scaffold and the keto acid part of

the molecule can form complexes with metal ions including Magnesium, Copper, Rhenium and Technetium[160].

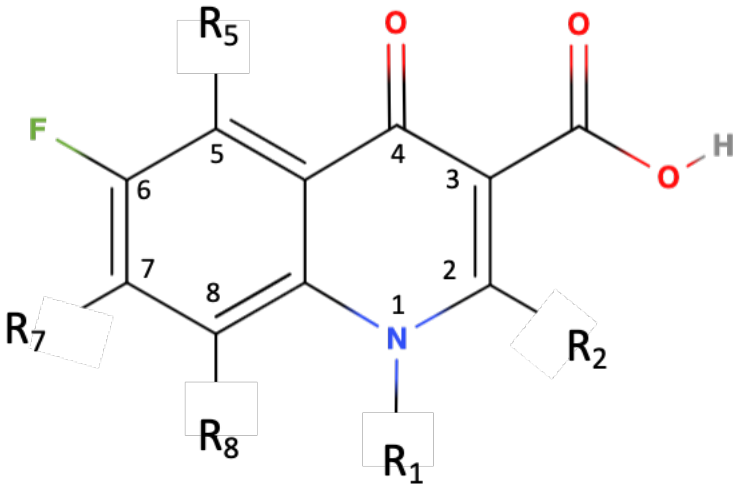

Figure 3 The fluoroquinolone scaffold. R shows where additional moieties may be introduced to the scaffold. Note: structures drawn using molview software[161].

Many different fluoroquinolone antibiotics have been developed to specifically treat a wide range of Gram positive and Gram negative bacterial infections and their use is particularly implicated in respiratory, urinary tract and sexually transmitted infections[162]. Despite their many uses, fluoroquinolone drugs are often over- or inappropriately prescribed[163]. Fluoroquinolones are now are one of the most commonly prescribed classes of antibiotics worldwide[164] and their global usage in humans is increasing[165]. Concerningly, but not unexpectedly, an increased usage, in both humans and animals, has been associated with increased levels of resistance in several bacteria[166-169].

1.4.1 Fluoroquinolones for the treatment of *M. tuberculosis*

Fluoroquinolones have been used in tuberculosis treatment programs since the 1980s when ofloxacin was trialled for treatment of pulmonary tuberculosis[56], and are now essential in many MDR and drug-susceptible TB treatment regimens (Table 4). Since ofloxacin, several newer generation fluoroquinolones have been developed which have greater *in vitro* and *in vivo* activity against *M. tuberculosis*[170-172]. Specifically, substituents at the C8 and C7 carbon and N1 nitrogen and the presence of fluorine at C6 (Figure 3) are desirable structural features for targeting *M. tuberculosis* as they can form interactions with the fluoroquinolone binding pocket in the target, DNA gyrase[173]. The newer fluoroquinolones include levofloxacin, the S form optical isomer of ofloxacin, and moxifloxacin and gatifloxacin which contain an 8-methoxy moiety that increases the bactericidal activity by forming a stronger complex with the DNA gyrase target[172] (Figure 4).

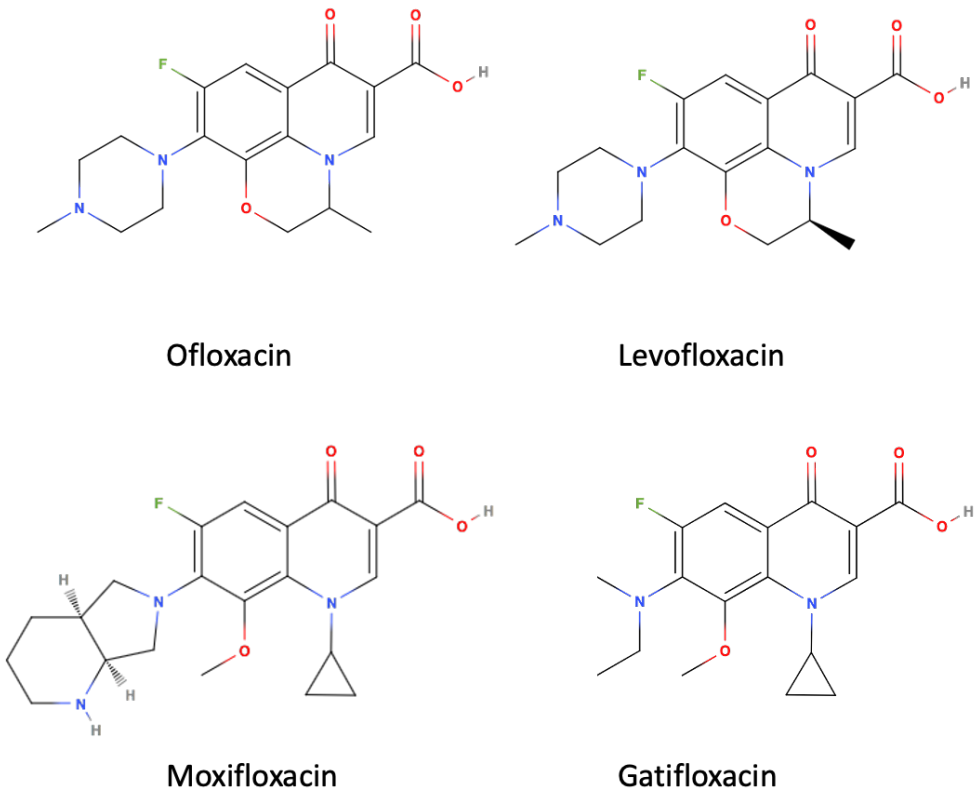

Figure 4 Chemical structures of fluoroquinolones used to treat TB. Note: structures drawn using molview software[161]

The fluoroquinolones are generally safe and well tolerated for TB treatment[174], but there is a small risk of fatal arrythmia due to a prolonged QT interval time[175] and the QT interval is less prolonged when patients are treated with levofloxacin compared to moxifloxacin[176]. Gatifloxacin is not currently recommended for the treatment of TB because it has been removed from the market over concern about its effects on blood glucose levels[70, 177]. The WHO considers both levofloxacin and moxifloxacin 'group A' agents, with high effectiveness and a good safety profile (Table 2) and a Korean study showed no difference in clinical outcome when using recommended doses of levofloxacin versus moxifloxacin to treat MDR TB[178]. However, *in vitro*, moxifloxacin has been shown to have better performance than levofloxacin[179] and kills bacteria more quickly[180], suggesting moxifloxacin may be a preferable choice in settings where there is a higher risk of non-compliance or treatment interruption.

1.4.2 Mechanism of action

The cellular target of fluoroquinolones in *M. tuberculosis* is DNA gyrase, a type II topoisomerase enzyme formed by a tetrameric complex of two *gyrA* and two *gyrB* subunits (Figure 5). The DNA gyrase creates and re-ligates double stranded breaks in DNA in an energy (adenosine triphosphate) dependent manner[165]. Creating and re-ligating double stranded breaks is necessary to unwind and untangle DNA prior to the essential processes of transcription and translation[165]. In *M. tuberculosis*, the DNA gyrase is the sole enzyme responsible for this task[28], making it a good antibiotic target. Humans also have type II topoisomerase enzymes although they are, generally, sufficiently different from bacterial ones as the *gyrA* and *gyrB* domains are fused to not lead to off target effects[181].

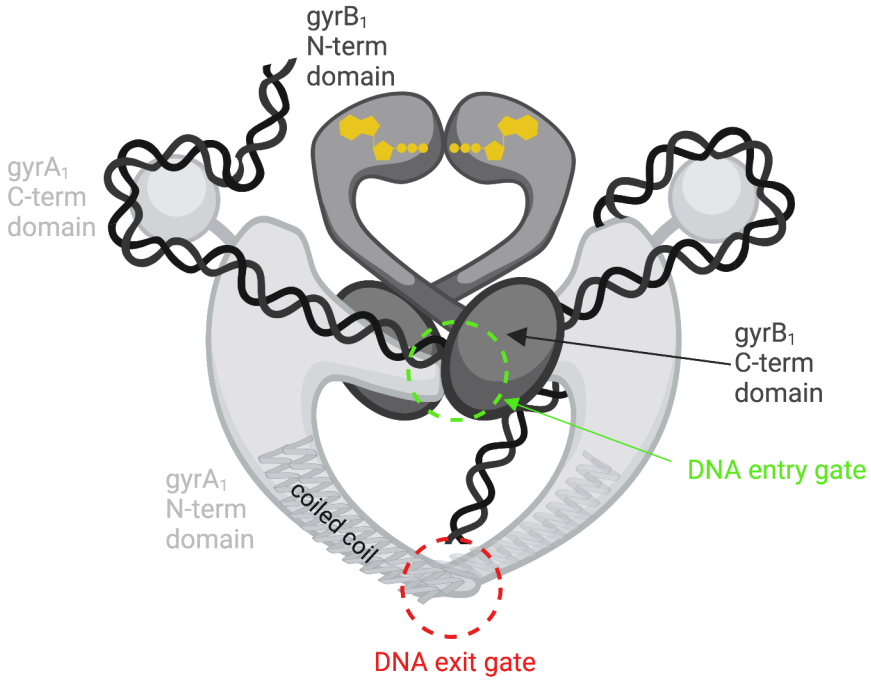

Figure 5 Schematic figure of a bacterial DNA gyrase. The DNA gyrase is formed of two *gyrA* and two *gyrB* subunits. The *gyrA* N terminal domains are responsible for breaking and re-ligating DNA and the *gyrA* C terminal domains bind DNA non-specifically for untangling. The *gyrB* N-terminal domain is responsible for ATP hydrolysis and the C-terminal domain for binding the subunits together and coordinating catalytic ions.

The DNA gyrase enzyme carries out its function using a "two metal mechanism"[182, 183], that in *M. tuberculosis* relies on Magnesium ions (Mg^{2+}) coordinated by the two *gyrB* subunits[184]. The reaction results in a staggered double stranded break in the DNA backbone and, to maintain genomic integrity prior to re-ligation, an evolutionarily conserved Tyrosine residue in each *gyrA* subunit forms a covalent bond with one of the newly cleaved 5' DNA ends[165,185]. In this state the complex is called a 'cleavage complex'[185]. Fluoroquinolone bound DNA gyrase cleavage complex structures show how the drugs exploit the enzymatic process; the fluoroquinolones intercalate into the DNA to produce a physical block that prevents DNA re-ligation[184] (Figure 6). Through the continued action of fluoroquinolones, the function of DNA gyrase (and thus transcription) is impaired and the concentration of poisoned cleavage complexes increases over the genome[186]. When the replication or transcription machinery

encounter the cleavage complexes, permanent chromosomal breaks are formed and if there are more permanent breaks than can be repaired, apoptosis is triggered and the bacterium dies[185, 186].

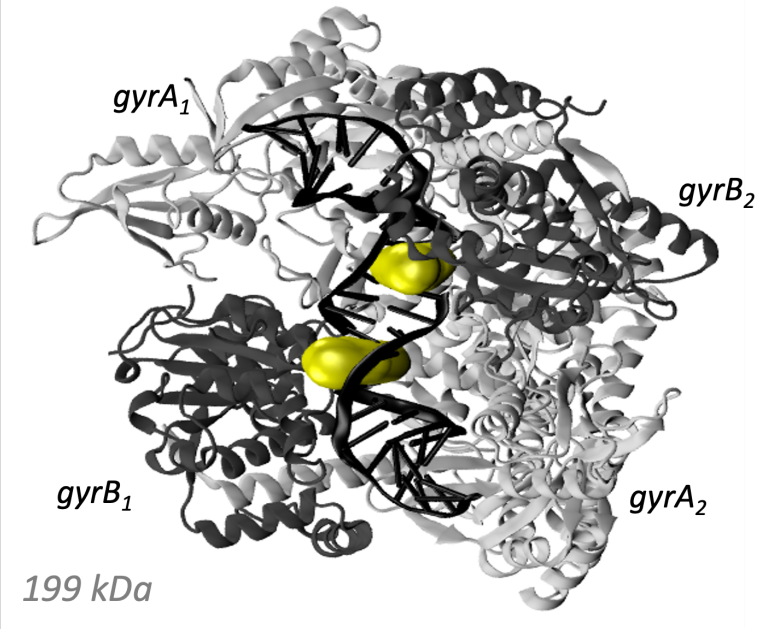

Figure 6 *Mycobacterium tuberculosis* DNA gyrase cleavage complex with a 19bp strand of double stranded DNA (black) and two bound moxifloxacin molecules (yellow). Moxifloxacin molecules are shown in a surface representation. Image created from PDB 5BS8[184] using VMD[187]. Note: the structure of the cleavage complex does not include the *gyrB* N-terminal ATPase (Figure 5).

Crystal structures show that levofloxacin and moxifloxacin interact with the *M. tuberculosis* DNA gyrase complex in a highly similar manner[184] (Figure 7). The drug molecules form only one direct interaction with DNA gyrase through coordination of a magnesium ion (Mg^{2+}) by the C3-C4 keto acid group. The Mg^{2+} forms a water bridge network with *gyrA* Aspartate residues at position 94. Both levofloxacin and moxifloxacin also bind in close enough proximity to potentially form contacts with a range of other *gyrA* and *gyrB* residues that are highlighted in Figure 7.

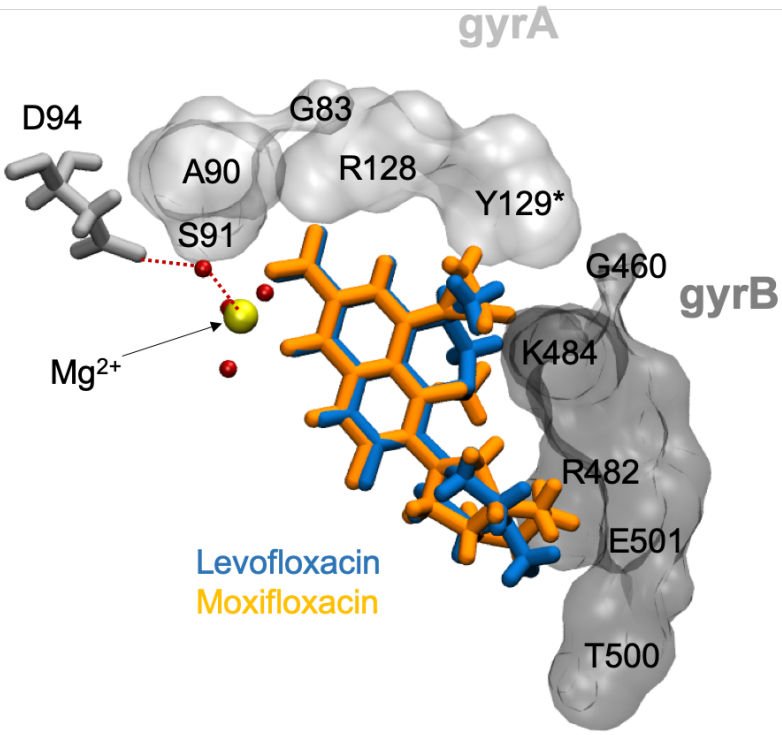

Figure 7 Levofloxacin and moxifloxacin binding poses in *M. tuberculosis* DNA gyrase cleavage complex. For clarity, DNA residues have been removed. Mg^{2+} coordinated crystal waters are shown as red dots and the water-ion bridge as red dashed lines. *gyrA* and *gyrB* residues within 5Å of either levofloxacin or moxifloxacin are highlighted. Image created in VMD[187] from PDB structures 5BS8 and 5BTG[184].

1.4.3 Fluoroquinolone resistance

Fluoroquinolone resistance arises during the treatment of MDR TB[188] and is associated with fluoroquinolone usage[188-190], although resistance may be able to arise independently of usage[188, 191]. Most fluoroquinolone resistance is found in MDR TB isolates, although a small proportion (1-3%) can be found in drug-susceptible backgrounds[192-194]. Using data from the past 15 years, the WHO estimates that around 20% of MDR TB infections are fluoroquinolone resistant[110], however there is limited data compared to rifampicin, because of more limited drug susceptibility testing[12]. In 2020, only 50% of notified MDR TB cases were tested for fluoroquinolone resistance and in some high burden TB regions, levels were below 30%[12].

Resistance to fluoroquinolones can be conferred by several mutations in the *gyrA* and *gyrB* genes[195-197] that can impact the structure and properties of the DNA gyrase fluoroquinolone binding site. These mutations tend to occur in the *gyrA* "quinolone resistance determining region" (QRDR), which spans *gyrA* residues 74 to 113[198], or less frequently, the proposed *gyrB* QRDR that encompasses *gyrB* residues 461 to 501[197]. The most observed resistance conferring mutations are *gyrA* D94G, which is thought to disrupt the ion water bridge network, and *gyrA* A90V, which likely causes a steric hinderance to fluoroquinolone binding[184, 199-201] (Figure 7). These common mutations have also been seen as part of mixed *M. tuberculosis* infections that have fluoroquinolone resistance[202].

The WHO catalogue of *M. tuberculosis* resistance-associated mutations identified 14 mutations associated with both levofloxacin and moxifloxacin resistance (Table 5) and some mutations that are not associated with resistance[42, 145]. The resistance-associated mutations are the same for both fluoroquinolones, however the *gyrB* E501D mutation is more confidently associated with moxifloxacin resistance than levofloxacin, which is unsurprising as there is evidence of a difference in the amount of resistance conferred to levofloxacin compared to moxifloxacin by this mutation[203, 204]. All 14 of the resistance associated mutations are in the target genes *gyrA* and *gyrB*, are located within the proposed QRDR regions (bar *gyrB* A504V) and are close to the fluoroquinolone binding site (Figure 8), suggesting disruption of drug-binding is the primary resistance mechanism. Genome wide association studies have uncovered few other potential genes implicated in fluoroquinolone resistance[205], and *in vitro* studies suggest that efflux may play a role as various efflux pump inhibitors decreased resistance levels to fluoroquinolones[206-208].

Mutation	*Association with levofloxacin resistance*	*Association with moxifloxacin resistance*
gyrA_D94G	Associated with resistance	Associated with resistance
gyrA_A90V	Associated with resistance	Associated with resistance
gyrA_D94N	Associated with resistance	Associated with resistance
gyrA_D94A	Associated with resistance	Associated with resistance
gyrA_S91P	Associated with resistance	Associated with resistance
gyrA_D94Y	Associated with resistance	Associated with resistance
gyrA_G88C	Associated with resistance	Associated with resistance
gyrA_D94H	Associated with resistance	Associated with resistance
gyrB_E501D	Associated with resistance	Associated with resistance - interim
gyrB_D461N	Associated with resistance - interim	Associated with resistance - interim
gyrB_A504V	Associated with resistance - interim	Associated with resistance - interim
gyrA_G88A	Associated with resistance - interim	Associated with resistance - interim
gyrB_N499D	Associated with resistance - interim	Associated with resistance - interim
gyrB_E501V	Associated with resistance - interim	Associated with resistance - interim

Table 5 WHO catalogue mutations associated with resistance to fluoroquinolones[42, 145]. Interim reflects that there is some uncertainty in the association with resistance, for example if the association is based on phenotypic tests that have not been fully validated.

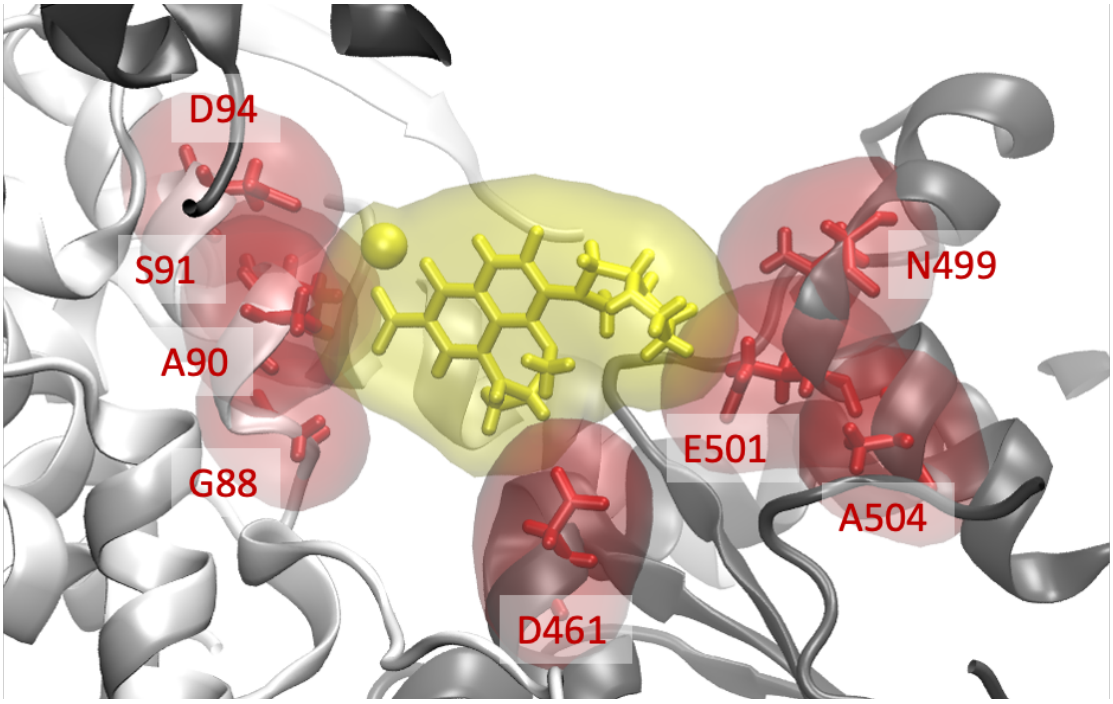

Figure 8 Positions of WHO catalogue mutations associated with fluoroquinolone resistance relative to the DNA gyrase fluoroquinolone binding site. Moxifloxacin and its coordinated Mg^{2+} ion are shown in yellow, gyrA chains are in white and gyrB chains are in grey. Image created in VMD[187] from PDB structure 5BS8[184].

To treat fluoroquinolone resistant TB, assuming the isolate is MDR, one should use the 6-month BPaLM treatment regimen without moxifloxacin[109] (Table 4). However, if this is not appropriate for the patient, a longer 18 to 20 month individualised longer treatment program is required (Table 4) [109]. There is evidence that high dose moxifloxacin could be used to treat low level fluoroquinolone resistance[209, 210], but a recent clinical trial in India showed no difference in the treatment outcome[211].

1.5 Thesis outline

In this thesis I aim to build upon our knowledge of fluoroquinolone resistance and test new methods to predict fluoroquinolone resistance from WGS data. These findings will be important to improve the performance of sequencing-based methods for fluoroquinolone DST in comparison to the current gold-standard but time-consuming phenotypic methods.

I will begin in chapter 2 by introducing the predictive methods I have selected to test, machine learning and free energy calculation. In chapter 3 I will introduce the CRyPTIC Consortium's dataset of WGS *M. tuberculosis* isolates and matched phenotypic DST results that form the underlying dataset for my work. Here I particularly focus on exploring the diversity of the isolates present in the dataset, identifying important resistance patterns and examining the limitations of the data. In chapter 4 I will use the CRyPTIC isolates to uncover and untangle genetic and geographical associations with phenotypic fluoroquinolone resistance and susceptibility. I will then evaluate the reliability of the assumptions of current sequence-based diagnostics, PCR-based molecular diagnostic tests and WGS with catalogue-based prediction, and assess their expected performance for detecting fluoroquinolone

resistance in the CRyPTIC isolates. In chapters 5 and 6 I will apply and evaluate machine learning algorithms and free energy calculations as rapid tools for predicting fluoroquinolone resistance based on structural changes associated with DNA gyrase mutations. Finally, in chapter 7, I will summarize my key findings and discuss their significance, limitations, and suggestions for further work.

2 Chapter 2: Predictive methods

2.1 Introduction

The quick and reliable prediction of drug resistance and susceptibility is important if the proposed next generation sequencing (NGS) approaches are to replace traditional culture-based phenotypic drug susceptibility testing (DST) for TB[144, 212]. A switch to using NGS could have a multitude of benefits as discussed in Section 1.3.3, but most importantly, using NGS in place of phenotypic DST would significantly decrease the time taken to return DST results[147], and therefore patients could be prescribed an effective TB treatment regimen more quickly. However, the current catalogue-based approach recommended to predict resistance from NGS data is not exhaustive[145, 158] and is unable to account for rare or novel resistance signatures until they have been seen a sufficient number of times in phenotypically resistant isolates[42, 145]. Rapid new computational predictive approaches are therefore essential to flag up resistance that may be missed by the current NGS predictive pipeline.

There are a range of computational predictive methods that can be considered and these can be divided into two contrasting groups; methods that are inference-based, versus methods that are theoretically exact. This chapter describes the predictive methods used in this thesis, machine learning (an inference-based method) and free energy calculations (a theoretically exact approach), and introduces the relevant theoretical background and rationale for particular methodological choices.

2.2 Machine learning

Artificial intelligence encompasses a range of technologies, including machine learning, that enable computers to simulate human behaviour to solve complex problems. Machine learning uses algorithms to learn patterns from data without being given specific instructions. For example, in *supervised* learning, a machine learning algorithm is used to train a statistical model using underlying patterns specific to a labelled dataset, and the resultant model can then be used to make predictions for new data. In the most basic form, a machine learning model predicts a label (y) and a function of predictors (x) that are also referred to as features:

$$y = f(x_i)$$

The machine learning algorithm is used to learn a statistical function of x_i using underlying patterns in the dataset so that y can be predicted for new data. The learning process is referred to as 'supervised' because throughout the training of the model, the labels for the dataset are provided to ensure the algorithm is learning the data patterns relevant to the labels.

The supervised machine learning approach is well suited to predicting antibiotic resistance and susceptibility as there are many complex underlying patterns that could be, and are, associated with resistance. Indeed, a range of machine learning models have been used to successfully predict antibiotic resistance in previous studies[213]. *Unsupervised* learning does not learn associations specific to labels, but rather learns patterns in unlabelled data and is therefore not appropriate for the goal of this thesis. Standard statistical modelling, upon which machine learning has its foundation, is also a useful inferential method that can be

used for prediction; however, the emphasis is on uncovering significant relationships and patterns within data, compared to the machine learning approach that concentrates more on the performance of the predictive aspect.

To evaluate the performance of a machine learning model, the dataset is first split into training and testing sets (usually 80% of the data is for training and 20% for testing). The 'performance' of a model can be assessed in multiple ways depending upon the desired characteristics of the model and this will be discussed later in the chapter. Once a model is trained using 80% of the data, it is used to predict the labels of the 20% of data that has not been seen by the model during training to evaluate how the model generalizes to new data. If the performance of a model is stronger on the training dataset than the test dataset, then the model can be considered overfit, meaning that it is too specific to the patterns seen in the training data and will not generalise well to make predictions for new data. This problem is common to all machine learning algorithms and needs to be considered and mitigated during training.

The feasibility and success of machine learning is ultimately reliant on the amount and quality of data used to train the algorithms, i.e. a large amount of accurate data are required. Thanks to the increased accuracy and decreased cost of NGS[214], it is now viable to sequence large numbers of clinical isolates and, in doing so, collect a wealth of genetic data. This has enabled major efforts by the CRyPTIC Consortium to collect whole genome sequencing (WGS) data for over 60,000 *M. tuberculosis* isolates from diverse backgrounds. The consortium has also collected matched phenotypic information in the form of minimum inhibitory concentration measurements (MIC) for 13 antibiotics for a subset of these

isolates. The WGS and phenotypic data of this subset were collected according to a standard operating procedure to ensure the quality and consistency of sample information. The resulting dataset is unparalleled in terms of its size and the wealth of information available. This makes it an excellent source of data with which to train machine learning models to predict antibiotic resistance in TB.

There are a range of supervised machine learning algorithms that can be considered for training models to predict antibiotic resistance or susceptibility (binary classifiers) or the MIC to a particular antibiotic, using the CRyPTIC dataset. However, it is important that the models considered for such resistance prediction tasks are easy to interpret, in that the rationale used for prediction should be understood, to build trust for use in clinical settings[213].

2.2.1 Binary classification

Binary classification is the task of predicting one of two categorical variables, and is therefore appropriate for predicting antibiotic resistance and susceptibility. There are many different machine learning algorithms that can be used for binary classification, including neural networks and support vector machines, but I chose to use forest-based models founded on decision trees, and logistic regression in this thesis due to their higher degree of interpretability.

2.2.1.1 Decision Trees

A decision tree describes a series of logical conditions that can be used to classify a dataset. An example of a three-layer tree classifying data as belonging to either group A or group B is shown in Figure 9. A decision tree machine learning algorithm iteratively constructs a tree that can then be used to make classifications. All the data instances are set at the root node of the tree and at this node, for each of the predictors, the data is partitioned based on a conditional value of the predictor. The information gain from the split is computed (i.e. how well the group A and group B data are split) for each predictor and the one that gives the largest information gain forms the logical condition at this node. If one of the branches from this criterion only contains group A or group B, the child node at the end of the branch is assigned as a leaf. If the branch contains both group A and group B labels then the child node at the end of the branch can be treated in the same way as the root node to further separate the group A and group B data in that branch. The tree continues to grow in layers this way until all the branches end in group A or group B leaves (i.e. all the data is partitioned) or until a user defined maximum tree depth is reached. When one has a new datapoint with an unknown label then one follows the tree until a leaf is reached, predicting the label.

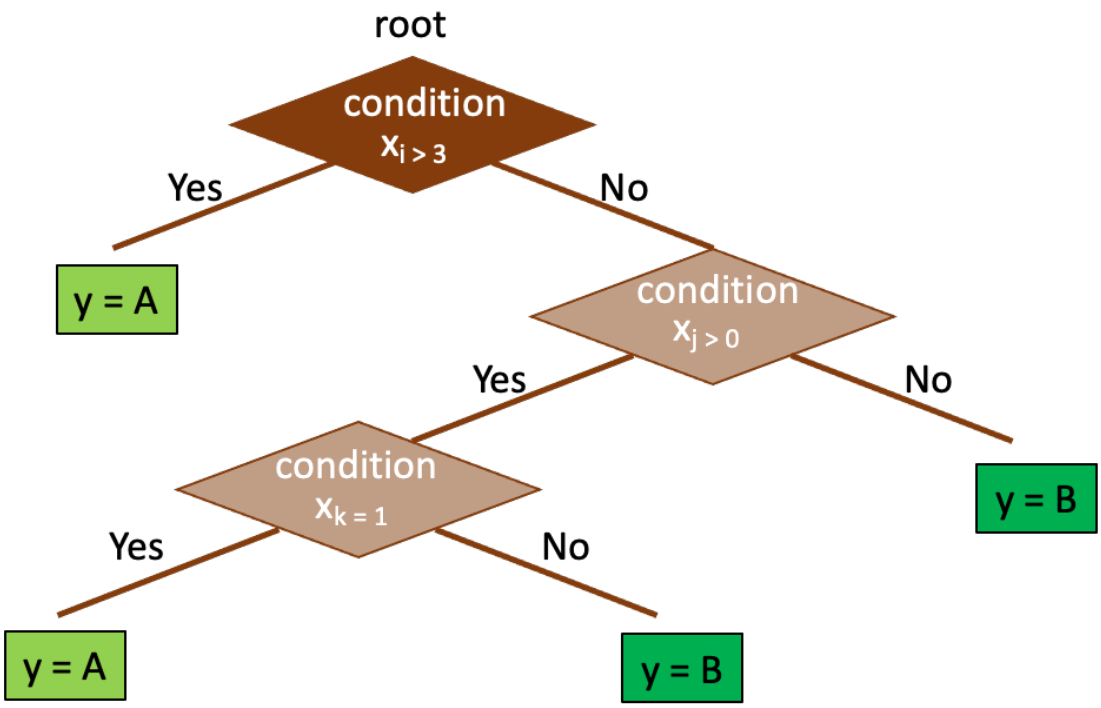

Figure 9 Three layered decision tree to classify whether a data entry is group A or B. The root node is depicted by a brown diamond, child nodes by light brown diamonds, leaf nodes as green squares and branches as brown lines.

A major advantage of decision trees is that, because a condition can occur in one branch but not another, it is possible to model interactions between different predictors that is not possible with other methods such as logistic regression. However, the method has the disadvantage in that it is prone to overfitting and so often does not generalise well when making new predictions.

2.2.1.2 Random Forest

Random forests were first introduced by Breiman[215] and comprise an ensemble of independent decision trees (see Section 2.2.1.1), each of which makes a class prediction for a datapoint, and the winning class prediction is the one with the majority vote from the

trees in the forest (Figure 10). By using many independent trees, the predictions are less dependent on errors from individual trees and therefore are less prone to overfitting. In order to make sure that the trees are independent the models are trained using 'bagging'; the original training data is randomly sampled and some data are replaced with other data in the training set (i.e. some datapoints are not used for a particular tree and others may be used twice). To further increase diversity among the trees, instead of considering all predictors when partitioning the data at a node, each tree is only allowed to use a random subset of predictors in the data.

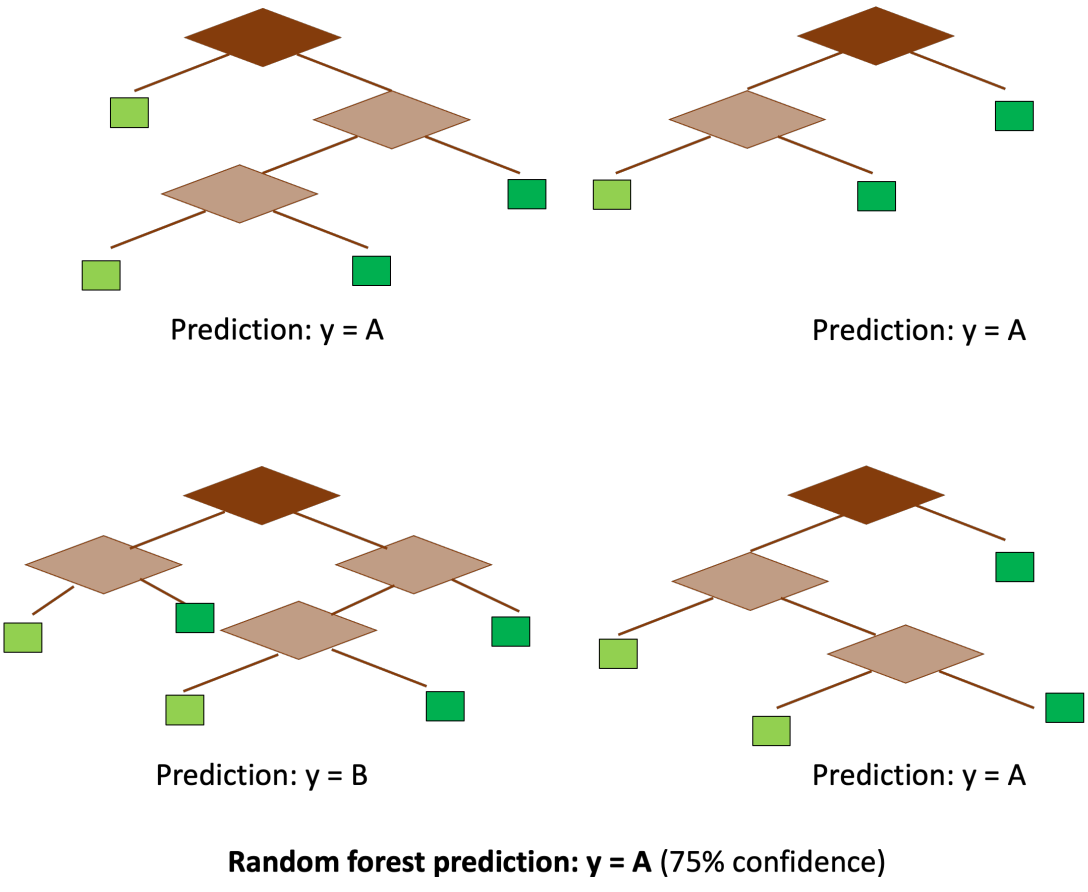

Figure 10 A prediction made by a random forest comprising four independent decision trees. The prediction of a random forest model is the majority vote from the independent trees, the proportion of the vote can be interpreted as a confidence score for the overall prediction.

Importantly, forest-based methods are interpretable; one can evaluate how important each of the predictors used is for making predictions. Each tree calculates the importance of each of the predictors used according to how well it increases the purity of the groups in the leaves. A more important feature will give a higher purity in the leaves. The average score for each predictor is taken from across the trees and is normalized to 1, so that the sum of all the predictor importance scores is 1.

2.2.1.3 XGBoost

XGBoost is an extension of the random forest model which uses gradient boosting to improve performance by creating new trees to predict the errors of prior models in an additive manner[216]. As this can result in overfitting, regularization factors such as shrinkage are also employed in the model. Shrinkage scales newly added weights from each step of tree boosting i.e. newly added trees are given less importance.

2.2.1.4 Logistic Regression

Logistic regression, although not a classification algorithm, can also be used for binary classification problems. Of the two groups, A and B, one is assigned the value 1 and the other 0. The log odds of y having a value of 1 can be calculated using the logistic regression model:

$$\log \frac{\mathrm{P}(x_i)}{1 - \mathrm{P}(x_i)} = \beta_0 + x_i \cdot \beta_i$$

Where β_i are coefficients (weights) and x_i the independent variables (predictors). From the logistic regression equation, the probability (*P*) that y takes on a value of 1 can then be calculated and a threshold can then be used to classify the datapoint as either 1 or 0. Generally, observations with a probability greater than or equal to 0.5 will be classified as 1 and those less than 0.5 as 0.

The task of the machine learning algorithm is to fit a model by learning the best weights for predicting which datapoints are '1', with the least error in predicting the groups. There are many different cost or loss functions (referred to as 'solvers') that can be used to estimate weights of the predictors, the most widely used are coordinate descent algorithms, such as LIBLINEAR[217], and algorithms based on Newton's method, such as the Limited-memory Broyden–Fletcher–Goldfarb–Shanno algorithm[218-221].

A regularisation term can be added to the equation to prevent overfitting of the model. There a many different regularization terms that can be used, but the two most common regularization terms, L1 and L2, penalize high coefficients in the model. If a feature occurs only in one class it will be assigned a very high coefficient by the logistic regression algorithm and thus the model will learn the training set too perfectly. The amount of regularization can be tuned using a hyperparameter '*C*' which increases or decreases the amount of penalization and therefore dictates how closely the model fits the training data; high values of *C* will return a more highly fitted model to the training data and low values will result in a more general model.

To interpret the logistic regression model, one considers the size of the assigned weightings (β coefficients) associated with each predictor (x), as this represents the expected change in log odds of y being 1, per unit change of x.

2.2.2 Minimum inhibitory concentration (MIC) prediction

MICs are conventionally measured using doubling dilutions and are therefore better described as a categorical variable rather than continuous. A more informative description of MIC is a series of ordered intervals. For example, supposing a bacterial population grows at a MIC of 1 mg/L but not at 2 mg/L, the true MIC is greater than 1 and less than or equal to 2 mg/L. A range of algorithms can be considered to predict MIC, and I again chose to test models that are extensions to the random forest and logistic regression approaches described in section 2.2.1, due to their interpretability.

2.2.2.1 Multiclass classification with random forests

Multinomial classification with random forest is an extension of the binary approach, where decision tree nodes are split until isolates are classified into one of any number of classes rather than just two. Again, the resultant class prediction is the one with the most votes across all the individual trees. This model does not consider that the MIC classes have order and therefore treats the problem as truly multiclass.

2.2.2.2 Ordinal methods

Ordinal methods such as 'all threshold'[222], 'immediate threshold'[222] and 'squared error'[223] consider ordering of classes. This is important as, for example, predicting an MIC of 0.5 mg/L when the real MIC is 4 mg/L would worse than predicting an MIC of 2 mg/L. These models are extensions of logistic regression, in that the probability P is split into a number of sections using thresholds which distinguish between group predictions.

The difference between 'all threshold', 'immediate threshold' and 'squared error' methods is how the models are penalised for making incorrect predictions by the loss function. For the 'all threshold' model, the amount of penalization increases over each threshold there is between the prediction and the true class. Therefore solutions (β weights) that minimise the number of thresholds that are crossed are encouraged and the method therefore minimises the mean absolute error between truth and prediction. The 'squared error' approach takes this a step further, as penalization increases logistically over each threshold crossed, and the method minimises mean squared error between truth and prediction. The 'immediate threshold' does not penalise for the number of thresholds between the true value and the predicted value, but rather has a blanket penalty for crossing any threshold.

Another method, ordinal ridge[223], works differently; a linear least squares model is fit with L2 regularisation, thereby treating the MIC as a continuous variable. Once an MIC is predicted it is rounded to the nearest group.

The amount of regularization can be tuned for all ordinal models using a hyperparameter α, where high values of α will return a model more highly fitted to the training data and low values will result in a more general model.

2.2.3 Evaluation of machine learning models

The 'performance' of a model can be assessed in multiple ways depending upon the desired characteristics of the model. For example, one could assess the performance of a binary classification algorithm by accuracy; simply the proportion of predictions that are correct. However, when using machine learning for medical applications, the performance of predicting one class may be more important than the other. For example, when predicting whether a patient has a disease it is most important to minimise the number of predictions of 'no disease' when a patient has disease. One therefore needs to consider the performance of a model in terms of sensitivity (the true positive rate) and specificity (1 – the false positive rate). To evaluate the overall performance in terms of both sensitivity and specificity one can use a receiver operating characteristic curve (ROC) and the probabilities of the class predictions. The curve begins at a point where all samples are classed as the negative class (0) and ends at a point corresponding to all samples are predicted as the positive class (1). The curve is then computed using different decision thresholds between 0 and 1, calculating the resultant true positive and false positive rates. The area under the curve (AUC) can be calculated to give the ROC AUC score, for which a value of 0.50 indicates that a model that performs no better than chance and a value of 1.0 indicates perfect performance.

To help prevent overfitting it can also be useful to evaluate the model performance during training. One can split the training set into a training and a validation set that is used as a mimic of the test set to tune models by identifying how well the model might generalize. However, the validation set is a relatively small sample compared to the rest of the dataset and therefore may not be an accurate representation of the population due to sampling bias. A further limitation is that splitting the training data to create a validation set decreases the amount of data one uses to train the model. An alternative approach is to use cross validation, where several 'folds' are created from the training data. In the case of five-fold cross validation, the data is split so that 20% of samples form a validation set. Each fold uses a different 20% of the training data as its validation set. Machine learning models are trained for each fold, using the remaining 80% of data and the models are evaluated on that fold's validation set. The best model can be selected as the one that has the highest mean performance across all the folds. In this way we can train and fine tune machine learning algorithms using all the available training data. Figure 11 shows an example process for how a dataset may be split for training and evaluating machine learning algorithms.

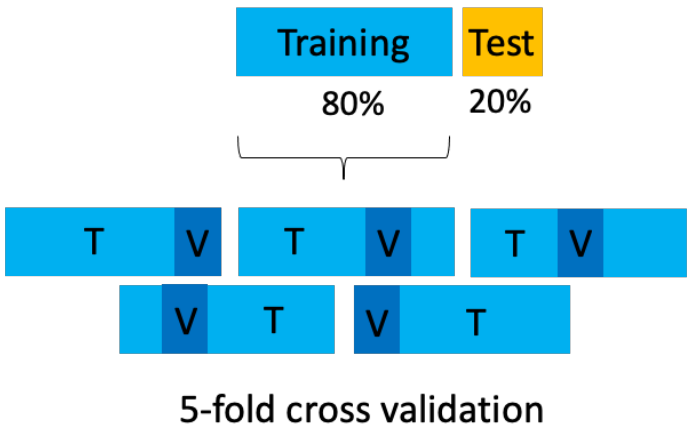

Figure 11 Splitting of a dataset to evaluate machine learning algorithms. In this example, the full dataset is split into training and test sets, 20% of samples are randomly assigned as the test group. The training set is further split into training (T) and validation sets (V) to facilitate model tuning using 5-fold cross validation.

2.2.4 Methods to improve model performance

To improve the performance of a machine learning model without impacting the interpretability, the simplest methods (asides from the obvious addition of more data with which to train the model) include feature selection and hyperparameter tuning.

2.2.4.1 Feature selection

Feature selection is a process used to reduce the number of features used to train a machine learning algorithm by removing irrelevant, redundant or noisy features. Reducing the number of features can reduce the likelihood of model overfitting, thereby improving performance on independent test sets compared to models trained without feature selection[224, 225]. Feature selection can also decrease the computational cost of training models by reducing dimensionality.

Although there are many approaches that can be used to reduce the number of features[226, 227], the method used in this thesis is based on recursive feature elimination (RFE)[228]. RFE is an example of a 'wrapper' method which is implemented using the desired machine learner. This is important because optimal features have been shown to depend on the particular biases of the machine learning model[225]. RFE performs backward feature elimination where the model is first fit using all the possible predictive features and then the least significant features are removed sequentially, until a desired number of features is reached. Backwards feature elimination has the advantage over a forward selection approach (where one starts with small set of features and adds more if they contribute to the prediction) in that it is more likely to identify features that interact with each other[228].

A problem with the RFE approach as described is that it is difficult to know upfront what the optimal number of features to use is. In this case, one can combine RFE with cross validation to automatically select the number of features to use. For every iteration where a feature is removed, the score of the resultant model trained is calculated on the validation set. The number of features left at the iteration which gives the maximum score on the validation set is taken as the optimal number of features.

2.2.4.2 Hyperparameter tuning

A hyperparameter is a value that can be used help improve the performance of machine learning models. There are many different hyperparameters that can be tuned to optimize models and reduce overfitting. For example, different values of *C* can be used in logistic regression to increase or decrease the level of regularisation and for forest-based models one can consider how the number of trees or the depth the trees are allowed to reach can influence how the models fit the training data.

Hyperparameter tuning also employs cross validation to train and evaluate a range of models each with different combinations of hyperparameter values. There are two main approaches that can be considered to find the best combination of hyperparameters, grid search[229] and random search[230]. For grid search, a user imputes a series of defined values for each hyperparameter whilst the grid search algorithm methodically tests all possible combinations of hyperparameter values on the training and validation sets. The grid search strategy may end up missing optimal models due to the constraints on the values of hyperparameters it can test and, as some hyperparameters may be more important than

others, grid search can unnecessarily explore hyperparameters that have little effect on the model performance. For random search one instead imputes a range between which values are randomly selected for each hyperparameter and defines the number of different random combinations of hyperparameters to try. Therefore where a larger number of hyperparameters are to be considered, a random search approach is favourable and has been shown to find models with equal or even improved performance using less compute time than the grid search approach[230].

2.3 Free energy calculations

The change in Gibbs free energy, ΔG, tells us whether a reaction is feasible or not in an isobaric-isothermal ensemble. It is defined in terms of the enthalpy (H), entropy (S) and temperature (T) of a system:

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

If it is negative, then the reaction will proceed, unless it is kinetically constrained. This property can be calculated for any reaction in a closed system, including the free energy associated with a binding process i.e. the binding affinity. The affinity can be affected by a range of factors including the number and strength of bonds formed between the ligand and the protein, the displacement of binding site waters, the flexibility of the ligand in the binding site and the ligand induced effects on protein conformation or flexibility. If one assumes that a protein mutation that causes resistance decreases the affinity of the drug for that protein, then one can use the binding free energy of the drug-ligand interaction to predict resistance.

There are several methods that can allow us to estimate the binding free energy of a protein-ligand interaction. These include molecular docking, endpoint methods, and alchemical free energy methods. Molecular docking is a two-step process which first docks a ligand into the binding site and then uses a scoring function to predict the binding affinity. There are a range of different scoring functions available, however they are unlikely to be accurate enough for this application due to their treatment of solvent as a continuous medium, limited ability to consider protein flexibility and in some cases an inability to account for entropy[231, 232]. Endpoint methods such as molecular mechanics/Poisson–Boltzmann surface area are more accurate than docking scoring functions and separately calculate enthalpic and entropic contributions to free energy using samples of the final states of a system[233]. However these methods typically make several approximations, especially when calculating the entropic contribution, and therefore are also unlikely to be accurate enough able to distinguish between relatively small differences in systems[234]. Some amino acid mutations constitute very small perturbations to a system, for example the resistance conferring mutation *gyrA* A90V[184, 199-201] only results in the addition of two methyl groups in place of two hydrogen atoms (Figure 12a,b). I assume that the molecular docking and endpoint approaches will not have sufficient accuracy to distinguish between the binding free energies of the ligand and a wild type compared to a mutated protein. I will therefore use alchemical free energy methods to predict binding affinities, as these are the most theoretically accurate methods, being rooted in classical statistical mechanics.

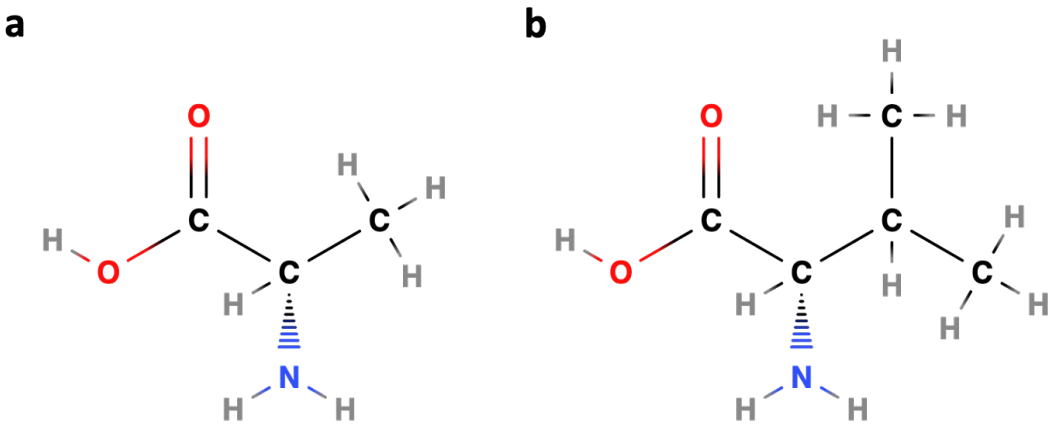

Figure 12 Structures of Alanine (a) and Valine (b)

2.3.1 Alchemical free energy methods

Assuming that a ligand forms a reversible noncovalent complex with the protein at thermodynamic equilibrium, the free energy difference between the bound complex in solution and the unbound ligand and protein in solution can be estimated as follows:

$$\Delta G = -RT\ ln\ K$$

where *R* is the gas constant, *T* is temperature and *K* is the equilibrium constant, which is the ratio of product (in this case the protein-ligand complex) to reactants (the protein and the ligand) in solution. Whilst the equilibrium constant, and therefore ΔG, can be calculated experimentally using isothermal titration calorimetry, it is more challenging to calculate computationally.

A macroscopic thermodynamic property, such as ΔG, is an average over an ensemble of many different microstates of the system. In our case the microscopic system comprises one protein molecule and one ligand molecule either bound or unbound in solution. A given

microstate refers to a unique positional configuration of the atoms in the system, and the ensemble of microstates that are averaged over must obey a probability density proportional to Boltzmann's factor (guaranteeing that higher energy states are less sampled). The energies of each microstate can be computed and the sum of all the states can be used to derive an ensemble average of the absolute binding free energy[235].

$$\Delta G = G_{bound} - G_{unbound}$$

$$\Delta G = < \Delta G >$$

$$= -RTln\left(\frac{Z_{bound}}{Z_{unbound}}\right)$$

where

$$Z = \sum_{i_1}^{i_n} e^{-\frac{E_i}{k_B T}}$$

where *i* are individual microstates, E_i is the energy of the given microstate, k_B is Boltzmann's constant and *T* is temperature.

Calculating the absolute binding free energy of a ligand binding to a wild-type protein and a ligand binding to a mutated protein in this way (ΔG_1 and ΔG_2, Figure 13) can be time consuming and inaccurate as there is a large difference between starting and end states (i.e. the ligand solvated in water and the ligand bound to the protein) and thus the calculations can take a long time to converge. In this thesis I will therefore use relative binding free energy calculations; since one is only interested in whether a mutation increases or decreases the antibiotic's affinity for the target, we need only calculate the difference in binding free energy between the wild type and mutant systems ($\Delta\Delta G_{binding}$, Figure 13).

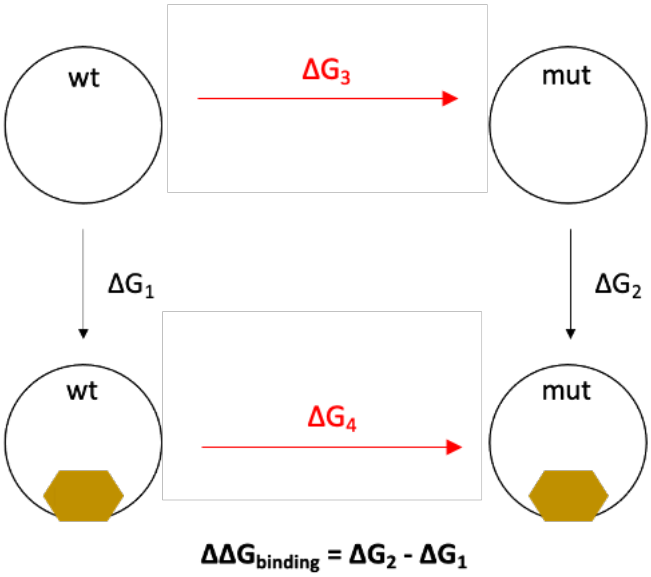

Figure 13 Free energy cycle of a ligand binding a wild type and mutated protein in a closed isothermal-isobaric system. As the system is closed and in NPT ensemble (constant temperature, pressure and number of particles), the total ΔG of the cycle is zero. Red lines represent the alchemical pathways.

Travelling around a free energy cycle of the closed system in an NPT ensemble (where there is no change in the number of particles (N) and a constant pressure (P) and temperature (T)), one ends at the same state that has the same free energy no matter the path taken, and therefore the sum of all the energies in the cycle is zero (Figure 13). Taking advantage of the path independence we can show that the Gibbs free energy of the wild-type protein transmuting into the mutant protein in the apo state (ΔG_3, Figure 13) and the free energy of the wild-type protein transmuting into the mutant protein with the ligand bound (ΔG_4, Figure 13) can be calculated to estimate the difference in relative binding free energy (RBFE) between the states:

$$\Delta G_3 + \Delta G_2 - \Delta G_4 - \Delta G_1 = 0$$

$$\Delta G_2 - \Delta G_1 = \Delta G_4 - \Delta G_3$$

$$\Delta\Delta G_{binding} = \Delta G_4 - \Delta G_3.$$

Calculating ΔG_3 and ΔG_4 instead of ΔG_1 and ΔG_2 to estimate $\Delta\Delta G_{binding}$ is more tractable. Assuming the ligand remains bound to the protein throughout the alchemical pathway ΔG_4 (Figure 13), we avoid having to wait for ligand binding (or dissociation) events as one must do if trying to calculate ΔG_1 or ΔG_2 directly. Further, the smaller difference in the end states for ΔG_3 and ΔG_4 (where one amino acid is different) compared to ΔG_1 and ΔG_2 (where a ligand is solvated or bound) means that the time taken for calculations to converge is much shorter.

To transmute one amino acid into another along an alchemical pathway, a three-step process is considered best practice to avoid inaccuracy in the free energy estimate[236, 237]. One starts by assuming that it is favourable to perturb as few atoms as possible to minimise the error introduced during the alchemical transformation, maximising the similarity between the states. I will take the assumed most favourable transmutation of Serine to Threonine as an example to illustrate the three-step process (Figure 14). Firstly, the partial charges on the Serine hydrogen atom to be perturbed are removed in step 1 (qoff step). Then in step 2, the van der Waals (vdW) radii of the additional methyl group of Threonine is gradually phased in as the Serine specific hydrogen vdW radii at that position in phased out (vdW step). Finally in step 3 the partial charges of the Threonine molecule are gradually phased in (qon step). The ordering of the steps is important because if the partial charges were retained while the vdW radii of the atoms were removed, the charges would become exposed and two positive or two negative charges could move toward one another resulting in huge electrostatic forces and instability[237].

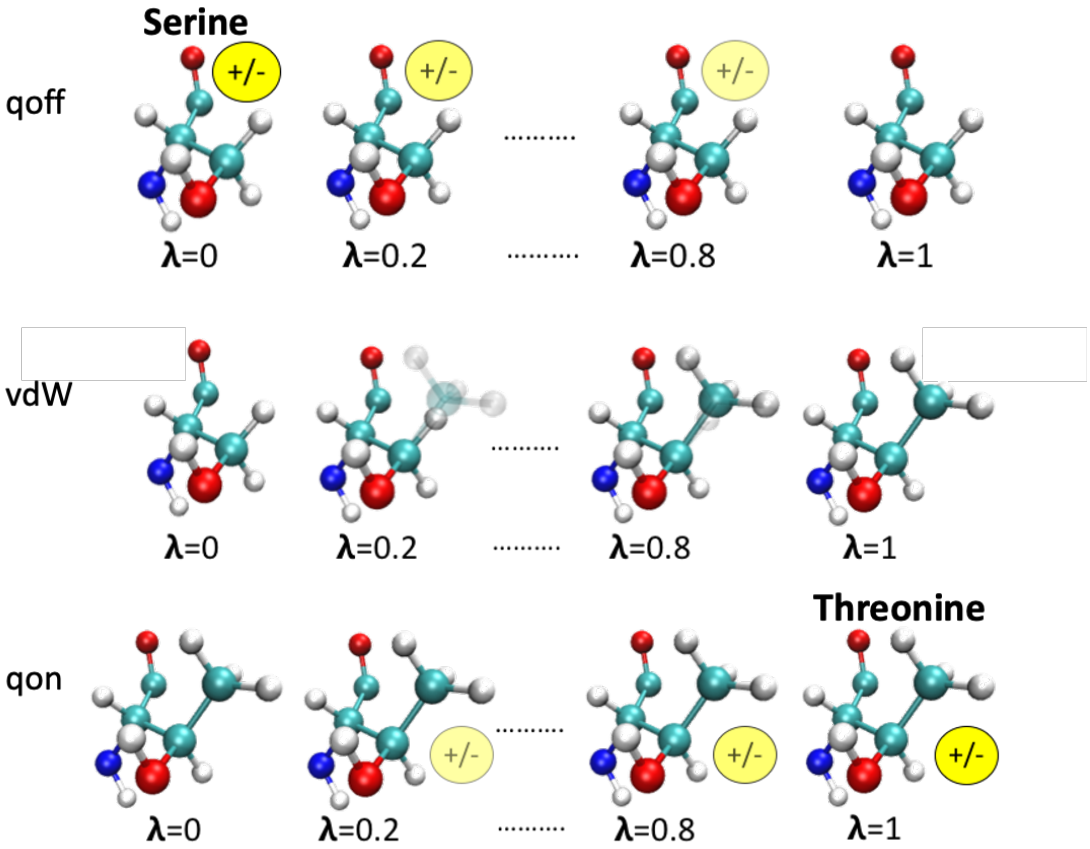

Figure 14 Alchemical transmutation of Serine to Threonine in three steps. Each step (qoff, vdW, qon) is performed gradually along a progress coordinate, λ.

2.3.2 Thermodynamic integration

A thermodynamic integration approach can be used to calculate the free energies ΔG_3 and ΔG_4. There are alternative methods that can be used to calculate the relative binding free energies, namely free energy perturbation (FEP)[238] and methods based on Bennett's acceptance ratio (BAR)[239]. The thermodynamic integration approach was used in this thesis, and I chose this method due to its relative ease of implementation and understanding compared to FEP and BAR methods.

Firstly, one defines a progress coordinate, λ, for the ΔG_3 and ΔG_4 alchemical transmutation reactions where 0 represents the wild-type amino acid state (A) and 1 represents the mutant amino acid state (B). Each state has a potential energy, *U* (internal energy of the system), and by constructing a pathway of intermediate states (λ-windows) between A and B that have values of the progress coordinate between λ = 0 and λ = 1, the free energy difference between the two states can be calculated as thus:

$$\Delta G = \int_0^1 \langle \frac{\partial U(\lambda)}{\partial \lambda} \rangle_\lambda \, d\lambda \, .$$

The values of the derivative $\partial U/\partial \lambda$ at each value of λ are calculated and stored by the molecular dynamics software at each timestep, permitting the ensemble average $\partial U/\partial \lambda$ of each λ-window to be calculated. The integral of these averages with respect to λ can then be approximated using an integration method, such as the trapezoidal rule, to give the alchemical free energy change (ΔG) between A (λ = 0) and B states (λ = 1). As such, the accuracy of the calculated free energy is dependent on the λ-window points chosen and the smoothness of the integrand. Thermodynamic integration has the benefit of allowing the user to flexibly place extra λ-windows where needed as each λ-window is treated independently. As the alchemical transformation should be completed in three steps (see section 2.3.1), the free energy contribution from each the qoff, vdW and qon steps can be estimated separately using thermodynamic integration and then added together to give ΔG_3 and ΔG_4 (Figure 15).

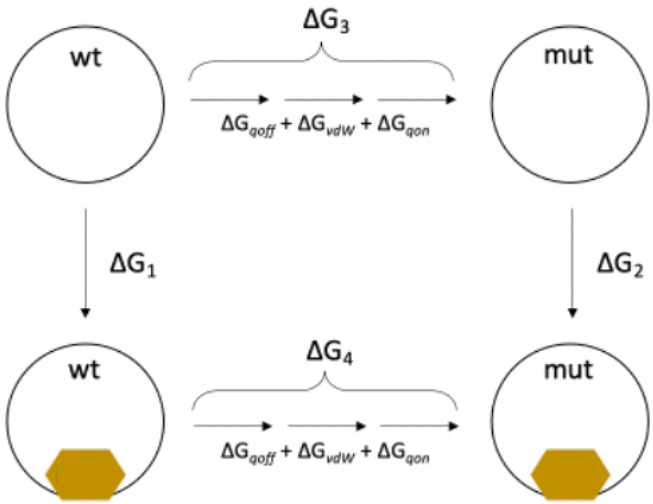

Figure 15 Calculation of ΔG_3 and ΔG_4 using the three step best practice alchemical method and thermodynamic integration

2.3.2.1 The singularity problem

A drawback of the thermodynamic integration approach is encountered when scaling the potential energy of vdW interactions between atoms (V) by λ. The vdW interactions are modelled using a Lennard-Jones potential:

$$V(r) = \frac{4\varepsilon\sigma^{12}}{r^{12}} - \frac{4\varepsilon\sigma^{6}}{r^{6}},$$

where r is the atomic distance between a pair of atoms, ε is the well depth at the optimal potential energy (how strongly the atoms attract one another) and σ is a measure of how close the two atoms can get. The part of the equation modelling the attractive potential is in blue and the repulsive part in red. A graphical representation of the potential helps show how, when one scales the potential by λ, a singularity can occur due to the removal or addition of atoms (Figure 16). Singularities can make the integration difficult to estimate and breaks thermodynamic integration's assumption that $\partial U/\partial\lambda$ is a continuous function.

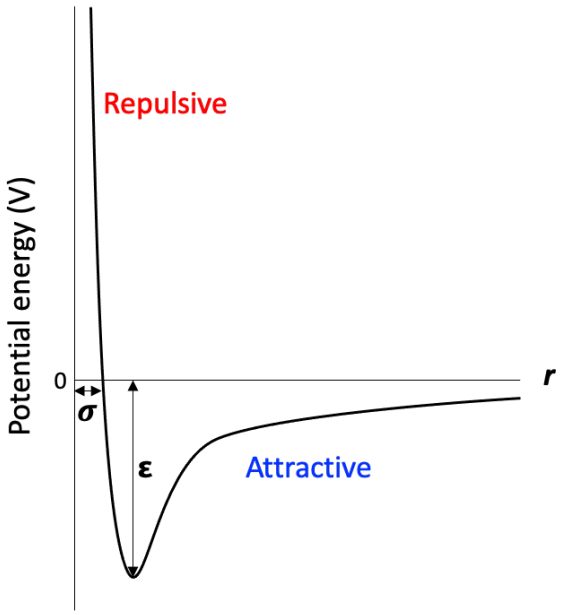

Figure 16 Graphical representation of the Lennard-Jones potential. r is atomic distance between a pair of interacting atoms. ε is the well depth at the optimal potential energy which measures how strongly the atoms attract one another and σ is a measure of how close the two atoms can get (i.e. the vdw radii).

This problem can be addressed using 'soft-core' potentials that keep the energies finite for all values of λ[240]. The soft-core potential takes advantage of free energy being a state function by modifying the Lennard-Jones potential for appearing and disappearing atoms so they are unphysical but finite at all values of λ except at λ=*0* and λ=*1* where they collapse into the standard Lennard Jones form. The soft-core potential used in this thesis[241] is:

$$\mathrm{V_{sc}}(r) = \lambda\left(\frac{4\varepsilon\sigma^{12}}{\mathrm{A}^{12}} - \frac{4\varepsilon\sigma^{6}}{\mathrm{A}^{6}}\right) + (1-\lambda)\left(\frac{4\varepsilon\sigma^{12}}{\mathrm{B}^{12}} - \frac{4\varepsilon\sigma^{6}}{\mathrm{B}^{6}}\right)$$

where

$$\mathrm{A} = (\alpha\sigma^{6}(1-\lambda)^{\mathrm{p}} + \mathrm{r}^{6})^{\frac{1}{6}}$$
$$\mathrm{B} = (\alpha\sigma^{6}\lambda^{\mathrm{p}} + \mathrm{r}^{6})^{\frac{1}{6}}$$

where α is the soft-core parameter (a constant, default value of 0.5 used), p is the soft-core power (a constant, default value of 1 used), σ is the interaction radius and *r* is the atomic distance.

2.3.2.2 Replica exchange

In order to accelerate the sampling of microstates to achieve accurate free energy predictions using thermodynamic integration, one may choose to use Hamiltonian replica exchange[242]. At each λ window the system has a different total energy because of how the λ value scales the forcefield's parameterisation of atoms which in turn affects the dynamics of the system. After a specified number of time steps the total energy of each pair of conformations is compared to the total energy after swapping the values of λ. If the energy is reduced, the swap is made permanent. If it the energy is higher, the swap is accepted with a probability proportional to the Boltzmann factor[243]. If a swap is made each λ window has the opportunity to sample more microstates. The testing and swapping process is repeated periodically to increase sampling at the different λ windows. However, if choosing to run the calculations using replica exchange, the λ windows cannot be flexibly placed and must be chosen up front because each λ-window is no longer independent of the others.

2.3.3 Molecular dynamics simulations

In order to generate an ensemble of microstates for free energy calculations using thermodynamic integration one needs to simulate the dynamics of the system in atomistic detail. This can be done using classical molecular dynamics (MD) [244]. In an MD simulation, all forces acting on all atoms by all the other atoms in the system are calculated at a point in

time and Newton's laws of motion are then applied to calculate the resultant forces and thence the velocities and accelerations of all the atoms, allowing their positions to be updated and time to be advanced. The process is repeated many times allowing the dynamics of the system to evolve, building up a simulation trajectory. In order to conserve energy the timesteps must be smaller than the fastest motion in a system, which for proteins are the vibrations of bonds involving hydrogen, which constrains the timestep to not be more than 1 fs. Importantly, over the MD trajectory, the distribution of energy states sampled is inherently proportional to the Boltzmann distribution (higher energy states are less likely to be observed) and hence the simulation naturally samples from the correct ensemble.

Several codes have been developed to run MD simulations; these include GROMACS[245], AMBER[246], NAMD[247] and CHARMM[248]. Due to the large number of force calculations (dominated by electrostatics, which naively requires N^2 calculations, where N is the number of atoms) that must be performed at each timestep, the production of MD trajectories is time-consuming and computationally expensive, which prohibited their use to study large systems at their inception[244, 249]. However, improvements in parallel computing, the continued increase in CPU speeds, optimisation of MD codes, the adoption of particle-mesh Ewald methods for calculating electrostatics and the use of graphical processing units (GPUs) have continually reduced the wall clock time to run simulations, which has led to increased popularity of the method in recent years[250].

2.3.3.1 *Forcefields*

Forcefields describe the potential energies and parameters of all the atoms in the system. A forcefield includes the potential energies for interactions between covalently bonded atoms (including bond stretching, angle flexing between two adjacent bonds, and the dihedral torsion) plus the interactions between atoms that are not bonded to one another such as van der Waals and electrostatic interactions (Figure 17).

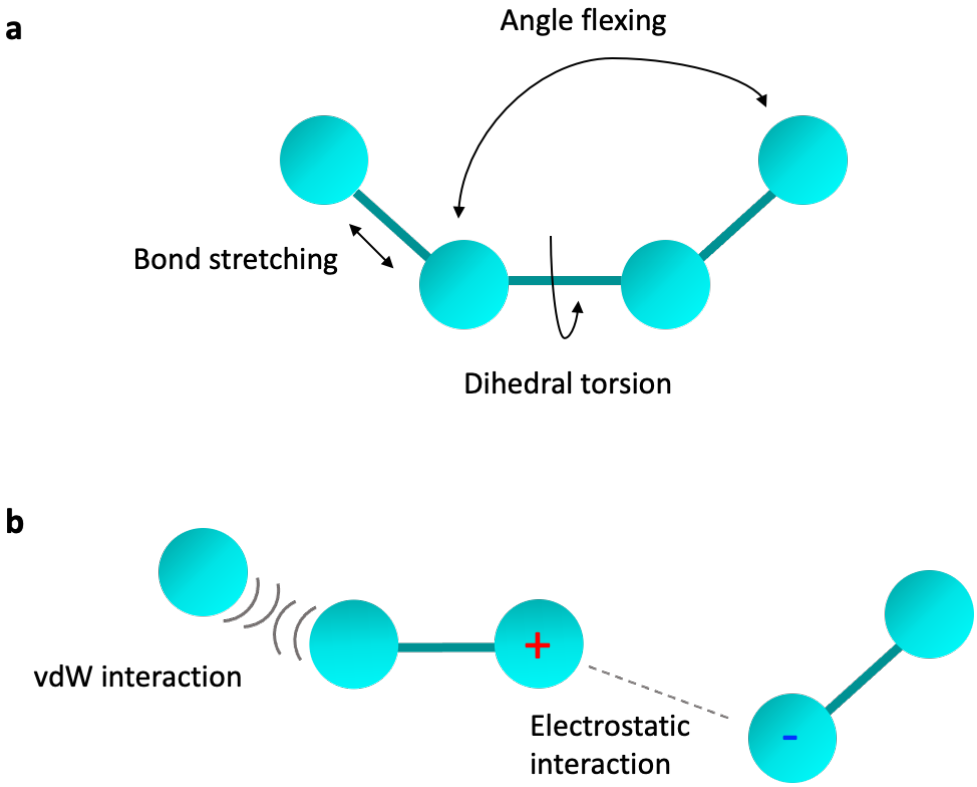

Figure 17 Bonded (a) and non-bonded (b) interactions modelled by forcefields

The accuracy of all thermodynamic properties calculated from ensembles generated by molecular dynamics is dependent on the accuracy of parameters in the forcefield. The parameters used to describe bonds between atoms often are calculated experimentally, for

example, bond equilibrium lengths can be measured from crystallography studies and bond spring constants can be determined using infrared spectroscopy whereby the strength of a bond corresponds to the vibrational frequency. For non-bonded interactions, to model the distribution of electronic charge over a molecule at the atomistic level, a partial charge approximation is employed and quantum mechanical calculations are used to assign a point charge, usually to each atomic centre. There are many different forcefields that can be used for parameterisation of small molecules and macromolecules such as proteins and nucleic acids including AMBER[251, 252], CHARMM[253, 254] and GROMOS[255] and these differ in their parameterisation of the different bonded or non-bonded interactions[251-253, 255-257]. The form of the AMBER forcefield used in this thesis is described:

$$U(R) = \sum_{bonds} K_r\left(r - r_{eq}\right)^2 + \sum_{angles} K_\theta\left(\theta - \theta_{eq}\right)^2 + \sum_{dihedrals} \frac{V_n}{2}(1 + cos[n\phi - \gamma])^2 + \sum_{i<j}^{atoms} \frac{A_{ij}}{R_{ij}^{12}} - \frac{B_{ij}}{R_{ij}^{6}} + \sum_{i<j}^{atoms} \frac{q_i q_j}{\varepsilon R_{ij}}$$

Bonded atoms are modelled using springs to represent bond lengths (blue), where *r* is the actual bond length, r_{eq} is the empirical bond length and K_r is a stretching force constant. Similarly, the angles between two bonds with a common central atom (green) are modelled using springs where θ is the actual bond angle, θ_{eq} is the empirical bond angle and K_θ is a bending force constant. Interactions between dihedral atoms are modelled using a function to approximate the differences in energy between the different possible confirmations (red) where V_n is the barrier to free rotation of the bond, *n* is the 360° rotational periodicity, ϕ is the torsion angle and γ is the angle where the potential energy is at its minimum value.

For the non-bonded interactions, the vdW interactions between pairs of atoms *i* and *j* are modelled using the 6-12 Lennard-Jones potential (purple). R_{ij} represents the atomic distance between *i* and *j* and A and B are constants that represent a measure of how strongly the atoms attract each other and the vdW radii. The electrostatic potential energy of all pairs of atoms in the system is modelled using Coulomb's law (orange), where R_{ij} represents the atomic distance between atoms *i* and *j*, q_i and q_j are the point charges on the atoms and ε is the permittivity of free space.

2.3.3.2 Integrators

To advance the timestep of an MD simulation, given the atomic positions and velocities at time *t* one can use Newtons second law of motion to find the force at time *t*, ***F(t)*** and then integrate to find the positions and velocities at the next timestep, $t + \Delta t$. A number of integrators exist which have different advantages and disadvantages; they are all obtained by starting from the Taylor expansions of the position, ***r***, and velocity, ***v***, where *m* is mass and ***F*** is force:

$$\mathbf{r}(t + \Delta t) = \mathbf{r}(t) + \mathbf{v}(t)\Delta t + \frac{\mathbf{F}(t)}{2m}\Delta t^2 + \cdots,$$

$$\mathbf{v}(t + \Delta t) = \mathbf{v}(t) + \frac{\mathbf{F}(t)}{m}\Delta t + \cdots.$$

The simplest is the Euler algorithm which truncates the expansions so that position and velocity at the next timestep are calculated as follows,

$$\mathbf{r}(t + \Delta t) = \mathbf{r}(t) + \mathbf{v}(t)\Delta t + \frac{\mathbf{F}(t)}{2m}\Delta t^2 ,$$

$$\mathbf{v}(t + \Delta t) = \mathbf{v}(t) + \frac{\mathbf{F}(t)}{m} \Delta t.$$

In practice, the Euler algorithm is inappropriate as it does not conserve energy and is not time reversible.

Commonly used MD integrators such as Verlet[258], velocity Verlet[259] and leap-frog[260] conserve energy and truncate the Taylor expansion less, however some truncation error will remain and there are other factors to consider when choosing the integrator. A Verlet algorithm combines the Taylor expansions for the positions at time *t* and at the previous timestep, $t - \Delta t$, and can be used to calculate the positions at the next timestep:

$$\mathbf{r}(t + \Delta t) = 2\mathbf{r}(t) - \mathbf{r}(t - \Delta t) + \frac{\mathbf{F}(t)}{2m}\Delta t^2 .$$

Although this does not include velocity, it can also be calculated:

$$\mathbf{v}(t + \Delta t) = \frac{\mathbf{r}(t + \Delta t) - \mathbf{r}(t - \Delta t)}{2\Delta t},$$

however this is dependent on first having calculated **r**(t + Δt) as above. One disadvantage of the Verlet algorithm is the velocity is calculated by subtracting two numbers which, if similar, can introduce numerical problems. The velocity Verlet algorithm is more commonly used and it is mathematically identical to the original Verlet algorithm presented here, but incorporates velocity explicitly to allow calculation of ***r*** and ***v*** without requiring the previous timestep, however you can only calculate the new velocities after the new positions and thence forces have been calculated,

$$\mathbf{r}(t + \Delta t) = \mathbf{r}(t) + \mathbf{v}(t)\Delta t + \frac{\mathbf{F}(t)}{2m}\Delta t^2$$

$$\mathbf{v}(t + \Delta t) = \mathbf{v}(t) + \frac{\mathbf{F}(t + \Delta t) + \mathbf{F}(t)}{2m} \Delta t.$$

Another way to calculate velocity at the new timestep without first having to calculate the new timestep positions, is the leap-frog algorithm where velocities are calculated a half step before the positions:

$$\mathbf{v}(t + \tfrac{\Delta t}{2}) = \mathbf{v}(t - \tfrac{\Delta t}{2}) + \frac{\mathbf{F}(t)}{m}\Delta t,$$

$$\mathbf{r}(t + \Delta t) = \mathbf{r}(t) + \mathbf{v}(t + \tfrac{\Delta t}{2})\Delta t.$$

2.3.3.3 Maintaining constant temperature and pressure

In order to estimate the Gibbs free energy (ΔG) of reactions at physiological conditions, the microstates generated by molecular dynamics simulations must belong to the NPT ensemble: the system should be closed with no change in the number of particles (N) and be isothermal-isobaric (at constant pressure (P) and temperature (T)). Thermostatic algorithms, such as those based on Langevin dynamics, can be used to keep the temperature constant by modifying Newton's laws of motion[261]. To keep the pressure constant, the system can be coupled to a pressure bath using a barostat such as the Parrinello-Rahman approach, which also involves a modification of Newton's laws[262].

2.3.3.4 LINCS

When running a MD simulation, the time step used needs to be slower than the fastest movement in a system in order to allow the dynamics to be accurately simulated. For

biological molecules the time step is therefore limited by high frequency covalent bond oscillations involving hydrogen atoms which have minimal effect on the dynamics of the overall system. There are several algorithms that can be used to constrain these bonds but the linear constraint solver (LINCS) is commonly used in GROMACS, which is our MD engine of choice. This algorithm can be used to reset bond lengths involving hydrogen atoms to the 'correct' length in the forcefield after an unconstrained update to the system, thereby in effect constraining the length of all bonds involving a hydrogen atom[263]. This allows a longer timestep of up to 2 fs to be used when evolving the dynamics, which reduces the number of timesteps needed to create a trajectory of a defined length and therefore the number of calculations to be solved, cutting down on the computational resources required to complete the MD simulation.

3 Chapter 3: Description of the CRyPTIC dataset

3.1 Introduction

The collection of *M. tuberculosis* isolates with matched genetic and phenotypic drug susceptibility testing (DST) information is particularly important for antibiotic resistance prediction from sequencing data. A matched dataset is necessary to associate individual mutations or genetic signatures with phenotypic effects. To date, several datasets, each larger than its predecessors, have facilitated the creation of catalogues of genetic mutations which can successfully predict resistance and susceptibility to several first and second line antitubercular drugs from sequencing data[143-145]. The success of genetic sequencing and catalogue-based resistance prediction relies on the production of comprehensive catalogues of resistance conferring mutations, however rare and novel resistance conferring mutations may not be identified or included in catalogues if they do not meet sufficient statistical confidence criteria[145].

Whole genome sequencing (WGS) data, as opposed to targeted sequencing of specific genomic regions, can help address these problems. WGS data has facilitated genome wide association studies[85] and the training of machine learning models[264-266] to identify rare, and predict putative, resistance conferring mutations throughout the genome. Identified mutations could be included in catalogues after confirmatory studies or flagged as putative resistance conferring mutations in bioinformatic pipelines. Further, the collection of WGS data, as opposed to other forms of sequencing data, has the added benefit of improving TB surveillance via analysis of global movement and transmission[149, 150].

To ensure that catalogues or predictive models are as generalisable as possible, it is important that the *M. tuberculosis* isolates on which they are based are diverse, as certain resistance or hyper-susceptibility conferring mutations can be associated with distinct geographic locations, lineages, sub-lineages, and phenotypic backgrounds. For example, (i) there is a high prevalence of the rifampicin resistance conferring mutation *rpoB* I491F in *M. tuberculosis* isolates in Southern Africa[132, 133], (ii) the majority of Lineage 2 isolates have a frameshift mutation in *tap* which may result in hyper-susceptibility to streptomycin even in the presence of other resistance conferring mutations[267, 268], (iii) the capreomycin resistance conferring mutation *tlyA* N236K is associated with Lineage 4.6.2 strains[269], and (iv) there are differences in the resistance conferring mutations seen in multi drug resistant (MDR) isolates versus those with rifampicin and pyrazinamide mono-resistance [270, 271]. It is therefore important that the biases and limitations of any data employed for these purposes are explored and acknowledged.

This chapter introduces the Comprehensive Resistance Prediction for Tuberculosis: an International Consortium (CRyPTIC) compendium, the largest, globally diverse, matched genotypic-phenotypic dataset of *M. tuberculosis* isolates to date[1]. The CRyPTIC consortium is a collaboration of 51 institutions from around the world that have the collective aim of improving TB control by achieving accurate prediction of resistance to anti tubercular drugs from WGS data. Ultimately the goal is to replace phenotypic DST with WGS, enabling a more rapid turn-around time for diagnosis, hence allowing patients to be quickly started on effective treatment in addition to advancing towards the goal of universal DST. In order to achieve these aims, the CRyPTIC consortium has collected a globally diverse dataset that contains 15,211 *M. tuberculosis* isolates. The whole genome of each isolate was sequenced and the minimum inhibitory concentration (MIC) to 13 antitubercular drugs was measured.

The 13 drugs include three of the four first line drugs; rifampicin, isoniazid and ethambutol, second line drugs; fluoroquinolones (levofloxacin and moxifloxacin), aminoglycosides (amikacin and kanamycin), ethionamide, rifabutin, two repurposed drugs; clofazimine and linezolid and two newer antibiotics; bedaquiline and delamanid.

My aims in this chapter are to highlight the diversity of isolates present in this dataset, identify the important resistance patterns and their implications, and recognise any biases and limitations of the data. This chapter has been published[1] and, unless otherwise stated, is my own work.

3.2 Methods

In this section I will outline how the CRyPTIC consortium collected and processed the 15,211 *M. tuberculosis* samples. Unless explicitly stated, all the work described was carried out by other members of the CRyPTIC consortium.

3.2.1 Sample collection and data processing overview

A full description of the methods for CRyPTIC data collection and processing has been published[272]. A brief overview is depicted (Figure 18) and outlined here, and details with relevance for this thesis are described in detail. As the dataset is publicly accessible, all files and data tables used for the analyses presented in this thesis, and are accessible via a file transfer protocol site (available at http://ftp.ebi.ac.uk/pub/databases/cryptic/reuse/).

The CRyPTIC consortium aimed to oversample for *M. tuberculosis* isolates with drug resistance and multi-drug resistance, however participating collection centres varied in their isolate collection approaches and timescales. For example, some sites used longitudinal sampling, rolling patient visits or biobank stocks. Metadata about each isolate (including country of origin and processing laboratory) was recorded and a summary table of the information was created. Each *M. tuberculosis* isolate was cultured, sequenced using an Illumina machine, and inoculated onto a 96-well broth microdilution plate to measure the MICs of 13 antitubercular drugs. After quality control procedures, phenotypic MIC data for 2,922 isolates were removed due to plate inoculation problems, labelling errors or contamination. The dataset therefore contains 15,211 isolates with WGS data, and 12,289 isolates with matched WGS and phenotypic data.

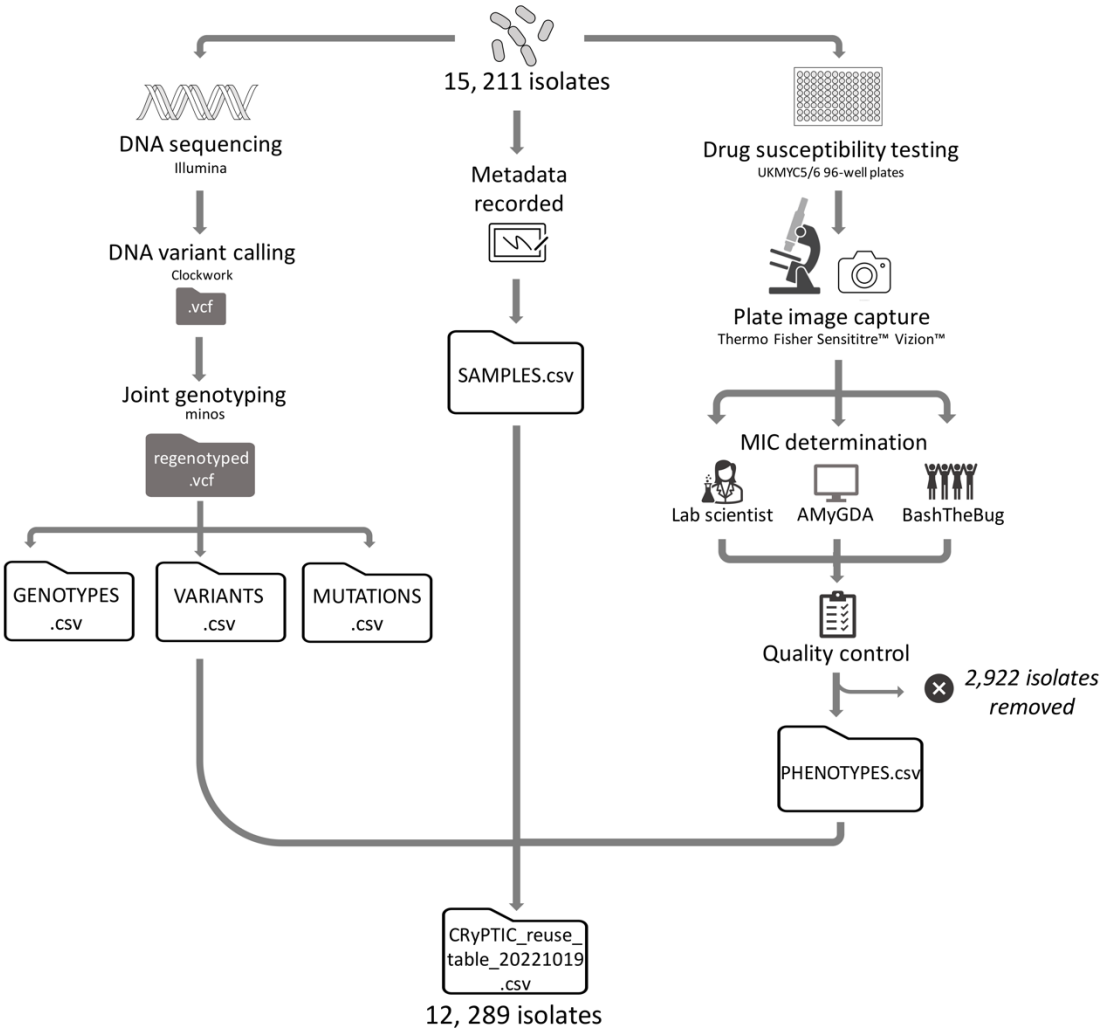

Figure 18 Collection and processing of sequencing and MIC data for 15, 211 global *M. tuberculosis* isolates. Briefly: Each isolate was DNA sequenced using an Illumina machine and plated onto 96 well plates containing 5-10x doubling dilutions of 13 antitubercular drugs for drug susceptibility testing. Associated metadata (including country of origin and processing laboratory) was recorded. DNA variant calling and analysis was performed using Clockwork and minos. After 14 days, MIC measurements were measured by a trained scientist using Vizion, and the plate was photographed to measure the MIC using the automated AMyGDA software and citizen scientists from BashTheBug. After quality control procedures, phenotypic MIC data for 2,922 isolates were removed. The dataset contains 15, 211 isolates with WGS data, 12,289 of which have matched phenotypic data, which are presented in "CRyPTIC_reuse_table_20211019.csv" on the FTP site (see Section 3.2.1)). Note: figure reproduced from The CRyPTIC Consortium[1].

3.2.2 Sequence data processing

Short reads produced by Illumina sequencing were first processed using the Clockwork variant caller using its default filters (available at https://github.com/iqbal-lab-org/clockwork). The pipeline removes common contaminants, maps reads to the H37Rv

reference genome (version 3)[142], applies both Cortex and SAMtools to independently call insertions/deletions and SNPs, and, via Minos[273], merges both sets of calls to produce a consensus set. The output from Clockwork is a variant call format (.vcf) file for each sample which lists all the positions where there is evidence that the genome may differ from the H37Rv reference genome[142]. In addition to single nucleotide variants, this can include insertions and deletions and positions where there is no or insufficient evidence to make a definitive assignment.

The samples next underwent a process called 'joint genotyping' by the Consortium; this aims to capture more of the genetic variation in each sample by simultaneously considering all genetic loci where there is evidence of variation in any sample in the entire dataset. This was done using Minos[273] and produces more comprehensive variant call files the Consortium calls 'regenotyped' .vcfs. These have an entry at all variable positions. As part of the joint genotyping process, several filters were applied to ensure only high confidence variants were called. These comprised a minimum read depth of two, a maximum read depth of less than the mean read depth plus three standard deviations and a minimum fraction of read support for the called allele of >= 90%. Finally, a minimum genotype confidence percentile filter was applied to remove low confidence calls using a statistical measure described in Hunt *et al.*[273] . As described elsewhere[1], a genome mask was applied to remove untrustworthy loci, such as highly repetitive regions, and positions where fewer than 90% of the isolates passed default clockwork or Minos filters were removed from the regenotyped.vcfs. Pass or fail information for each of the filters are included as columns in the summary tables VARIANTS.csv and MUTATIONS.csv which are produced from the regenotyped .vcf files using gumpy (available at https://github.com/oxfordmmm/gumpy). These tables contain all single nucleotide polymorphisms and insertion or deletion variants

in genes and intergenic regions of interest for antibiotic resistance. The tables only include variants that are not called as wild type compared to version 3 of the H37Rv ATCC 27294 reference genome[142] and this means that the fraction read support for the wild-type allele must be < 50% (this is important later). Lineages were assigned by Mykrobe[274] and are recorded in GENOTYPES.csv.

3.2.3 Minimum inhibitory concentration measurement

M. tuberculosis isolates were cultured on solid media and then suspended in Middlebrook 7H9 broth according to a standard operating procedure[275]. A 100-fold dilution of suspension was prepared and 100 µl was used to inoculate each well of either a UKMYC5 or UKMYC6 96 well plate. These plates were designed by the CRyPTIC project and are based on the MYCOTB plate that is commercially available from Thermo Fisher. The UKMYC6 plate design is a slight modification of the original UKMYC5 design based on the results of a validation study[275]. Both plates include 5-10 doubling dilutions of 13 antitubercular drugs: rifampicin (RIF), isoniazid (INH), ethambutol (EMB), levofloxacin (LEV), moxifloxacin (MXF), amikacin (AMI), kanamycin (KAN), bedaquiline (BDQ), clofazimine (CFZ), delamanid (DLM), linezolid (LZD), ethionamide (ETH) and rifabutin (RFB) that are freeze dried to the base of the wells[272] (Figure 19a,b). Pyrazinamide was not included on either plate due to poor performance because the broth was not sufficiently acidic, and para-aminosalicylic acid was not included on the UKMYC6 plate because the growth on the UKMYC5 plate was not reproducible[275].

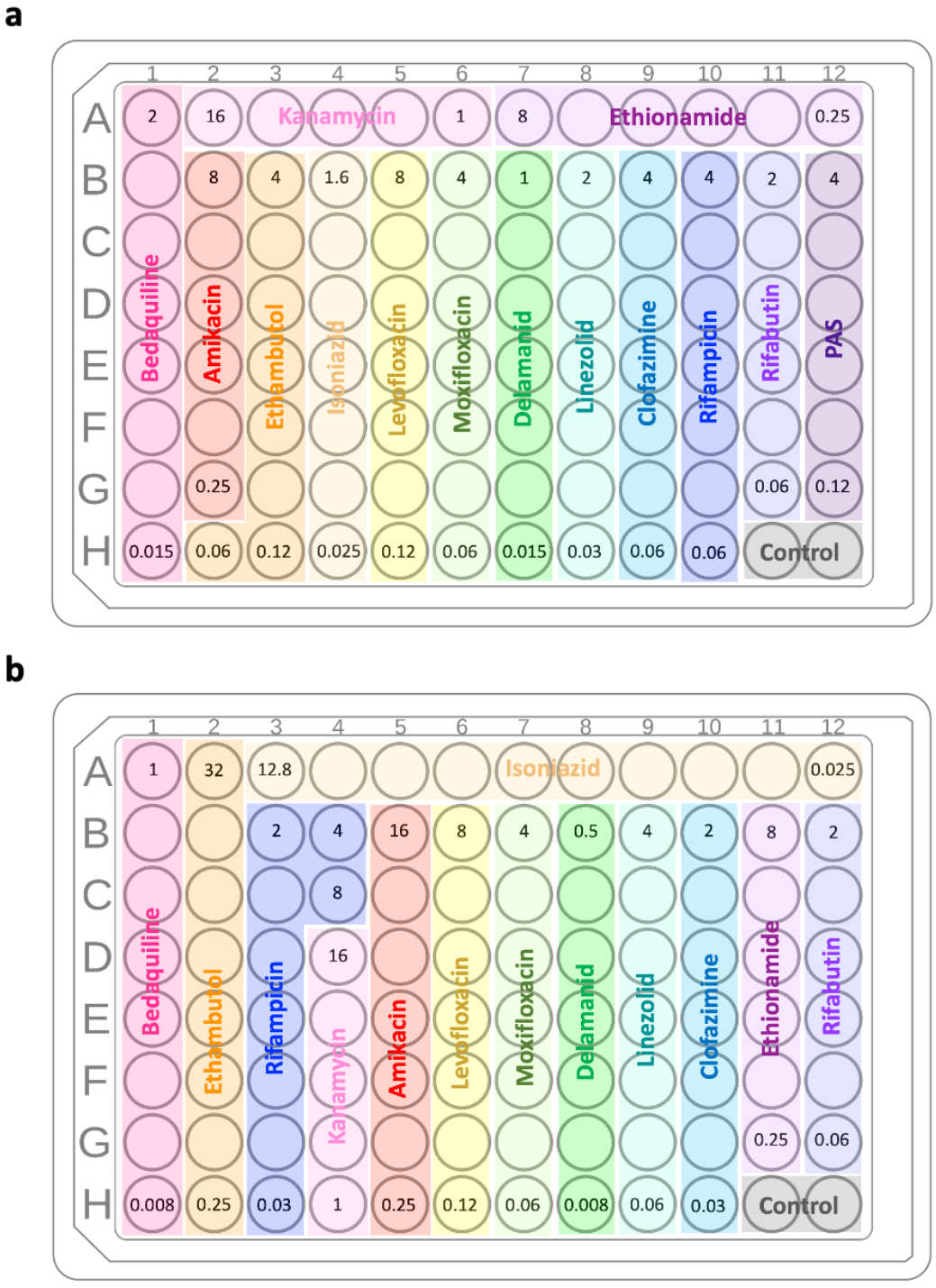

Figure 19 UKMYC5 (a) and UKMYC6 (b) 96 well plate designs. The first and last concentrations (mg/L) in the series are shown, the wells in between have intermediate sequential doubling dilutions. The wells labelled 'control' are positive control wells containing no antibiotic. PAS = para-aminosalicylic acid. Note: reproduced from the CRyPTIC Consortium[272].

After 14 days of incubation, the MICs were measured by a trained scientist using a Vizion™ digital viewing system, which also photographed the plate. All photographs were uploaded to a central database via a web browser. Since assessing the growth of *M. tuberculosis* is a difficult and subjective task, CRyPTIC chose to also measure the MICs using two alternative approaches; the automated AMyGDA software[276] and consensus measurements derived from growth classifications made by the citizen scientists of the BashTheBug project[277]. After quality control procedures, phenotypic MIC data for 2,922 isolates were removed.

Of the 12,289 remaining isolates, 88.1% returned an MIC reading for all 13 drugs on the plate. For each drug, the number of isolates with an MIC measurement, and the associated quality of the reading, is presented in Table 6. The quality of an MIC reading is classified based on concurrence between the different approaches used to measure MIC. MICs were classified as high quality if at least two methods concurred on the MIC and low if all three methods differed. A MIC measurement was assigned medium quality if the laboratory scientist recorded the Vizion measurement but did not capture a plate image, or if the Vizion and AMyGDA MIC measurement disagreed and there was no BashTheBug measurement.

	MIC MEASUREMENTS	HIGH QUALITY	MEDIUM QUALITY	LOW QUALITY
INH	12,070	9,519	1,351	1,200
RIF	12,099	8,955	1,356	1,788
EMB	12,158	7,506	1,355	3,297
LEV	12,163	7,774	1,354	3,035
MXF	12,194	6,785	1,353	4,056
AMI	12,072	8,973	1,350	1,749
KAN	12,130	9,333	1,355	1,442
BDQ	12,068	8,536	1,355	2,177
CFZ	12,049	7,763	1,352	2,934
DLM	11,927	8,095	1,349	2,483
LZD	12,189	7,141	1,355	3,693
ETH	12,132	8,821	1,355	1,956
RFB	12,150	10,042	1,352	756
TOTAL	157,401	109,243	17,592	30,566

Table 6 Quality metrics for phenotype data. Stated for each drug is the total number of MIC measurements stratified into "high" quality (at least two MIC measurement methods agree), "medium" quality (either Vizion and AMyGDA disagree, or there is no plate picture) or "low" quality (all three MIC measurements methods disagree) phenotype classifications as described in Methods.

3.2.4 Binarisation of MIC measurements into resistant and susceptible

Binary resistant and susceptible phenotypes were assigned to the isolates from their MICs by applying epidemiological cut-off (ECOFF) values which were proposed by the CRyPTIC consortium (Table 7)[272]. The ECOFF is defined as the MIC that encompasses 99% of phenotypically wild-type isolates. Isolates with MICs at or below the ECOFF are wild-type by definition and therefore are susceptible to the drug in question and isolates with MICs above the ECOFF are considered to be non-wild type, having acquired resistance to the drug[278].

DRUG	ECOFF (mg/L)
Isoniazid	0.1
Rifampicin	0.5
Rifabutin	0.12
Ethambutol	4
Ethionamide	4
Levofloxacin	1
Moxifloxacin	1
Amikacin	1
Kanamycin	4
Bedaquiline	0.25
Clofazimine	0.25
Delamanid	0.12
Linezolid	1

Table 7 Epidemiological cut-off values (ECOFFs) used to binarize MIC measurements into resistant and susceptible. Isolates with an MIC above the cut-off are considered resistant and those at or below the cut-off as susceptible. These ECOFFs were proposed by the CRyPTIC Consortium[272].

The distribution of MICs for each drug for each of the UKMYC5 and UKMYC6 plates is shown in Figure 20a-b. The MIC measurements of drugs such as isoniazid, rifampicin, rifabutin, amikacin and kanamycin follow a bimodal distribution which easily separates into distinct resistant and susceptible distributions upon application of the ECOFF. For ethambutol, ethionamide, levofloxacin and moxifloxacin the distributions are more complex, making identifying resistant samples more difficult, and the new and repurposed drugs (bedaquiline, clofazimine, delamanid and linezolid) have very few isolates with high MIC.

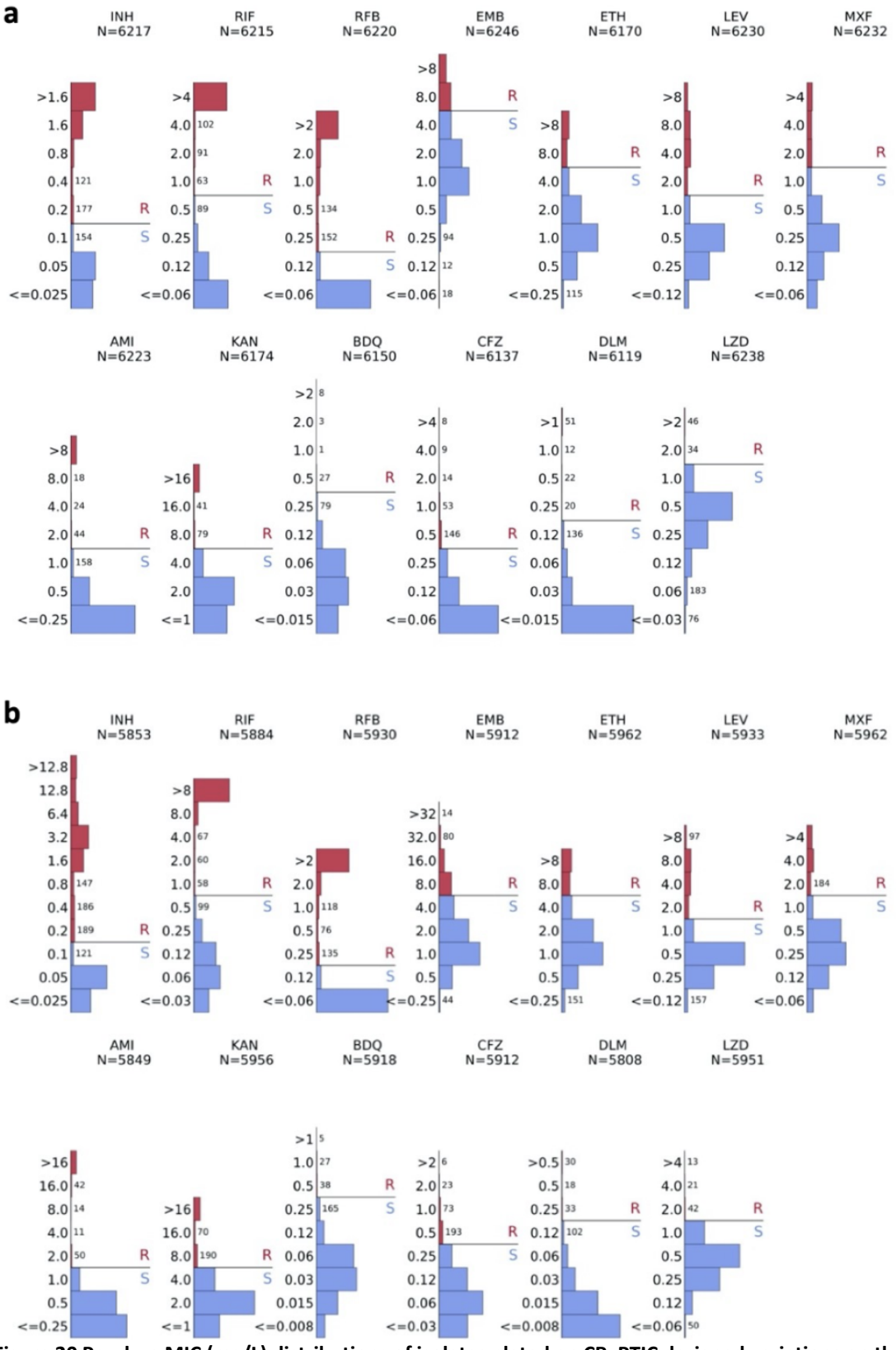

Figure 20 Per drug MIC (mg/L) distributions of isolates plated on CRyPTIC designed variations on the Thermo Fischer Sensititre MYCOTB MIC plates; UKMYC5 (a) and UKMYC6 (b). The solid black line represents the epidemiological resistance cut-off (ECOFF) for each drug as determined by The CRyPTIC Consoritum[272]. Isolates with an MIC above the cut-off are considered resistant. N denotes the total number of isolates tested on each plate that returned a phenotype for each drug. Where a bar represents less than 200 isolates, the number of isolates is labelled.

3.2.5 Data Analysis

Unless otherwise stated, I performed all data analysis in the remainder of this chapter. All data analyses were conducted and graphs were prepared using Python juypter notebooks and packages. For statistical tests, 95% confidence intervals were used throughout. Two proportion z-tests were used when comparing two proportions, and Wilson's method was used for estimating 95% confidence intervals for individual percentages[279]. Please see https://github.com/alicebrankin/thesis_notebooks for the codebase to reproduce the analyses and figures in this chapter.

3.3 Results

3.3.1 12,289 Mycobacterium tuberculosis Isolates

The 12,289 isolates with matched phenotypic and genotypic data originate from 20 countries across Africa, Asia, Europe, and South America (Figure 21). Almost two thirds (63.4%) of the isolates are from five countries: Peru (2,638 isolates), South Africa (1,641 isolates), India (1,468 isolates), China (1,107 isolates), and Vietnam (935 isolates). Just under a fifth (2,369, 19.3%) of the isolates had no country of origin recorded, which could arise as some isolates came from freezer stocks or because metadata was incompletely recorded at processing laboratories.

Just over half of the isolates in the dataset are from Lineage 4 (50.4%), with the largest contributor of these isolates being Peru. The next most common lineages in the dataset are Lineage 2 (35.2% of isolates), of which the largest contributor was China, Lineage 3 (8.6% of isolates), of which the largest contributor was India, and Lineage 1 (5.6% of isolates), of

which the largest contributor was India. No Lineage 5 isolates are present in the dataset and only six isolates in the dataset are from Lineage 6; these originated from Burkina Faso and Germany. The remaining 14 isolates were determined to be *Mycobacterium bovis*, an animal restricted pathogenic mycobacterium which can cause disease in humans[280]. There was a significant association between country of origin and lineage (Pearson's chi-squared test, X-squared = 7707.9, p = 0.0).

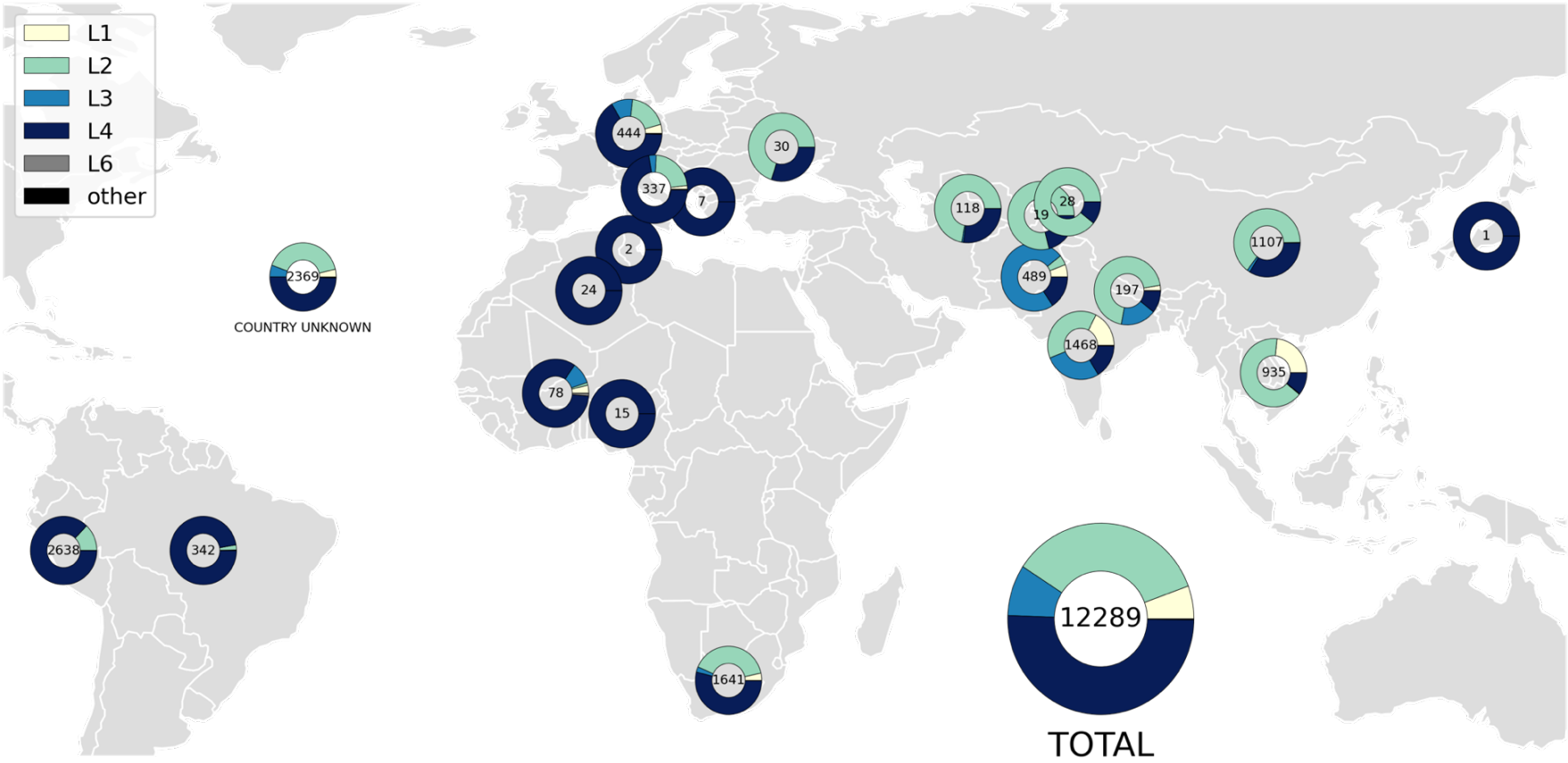

Figure 21 Country of origin and lineage distribution of 12,289 CRyPTIC isolates with matched genotypic and phenotypic data. Pies show the proportions of isolates of different lineages in each contributing country, the number of isolates originating from that country is shown in the centre of the pie. The TOTAL pie shows the proportion of each lineage present in the total dataset. The COUNTRY UNKNOWN pie shows the isolates for which the country of origin was unknown (e.g. if metadata was incompletely recorded or the isolate came from a freezer stock). L1 = Lineage 1, L2 = Lineage 2, L3 = Lineage 3, L4 = Lineage 4, L6 = Lineage 6, other = animal-restricted mycobacteria. There is a significant association between country of origin and lineage (Pearson's chi-squared test, X-squared = 7, 707.9, p = 0.0).

3.3.2 Resistance to 13 antitubercular drugs

Resistance to each of the 13 antitubercular drugs is represented within the dataset, which was expected given the size and bias towards collection of resistant isolates (Figure 22). Isoniazid had the highest percentage of resistant isolates (49.0%), followed by the other first line drugs rifampicin (38.7%) and ethambutol (18.6%). Although not a first-line drug, since it belongs to the same drug class as rifampicin, rifabutin also has a high prevalence of resistance (36.7%). The second line drugs generally had lower levels of resistance, with levofloxacin having the highest proportion of resistant isolates (17.6%), followed by ethionamide (14.2%), moxifloxacin (14.1%) and then the injectable drugs kanamycin (9.2%) and amikacin (7.3%). Reassuringly, a relatively small of isolates were assessed to be resistant to the new and repurposed drugs clofazimine (4.4%), delamanid (1.6%), linezolid (1.3%) and bedaquiline (0.9%).

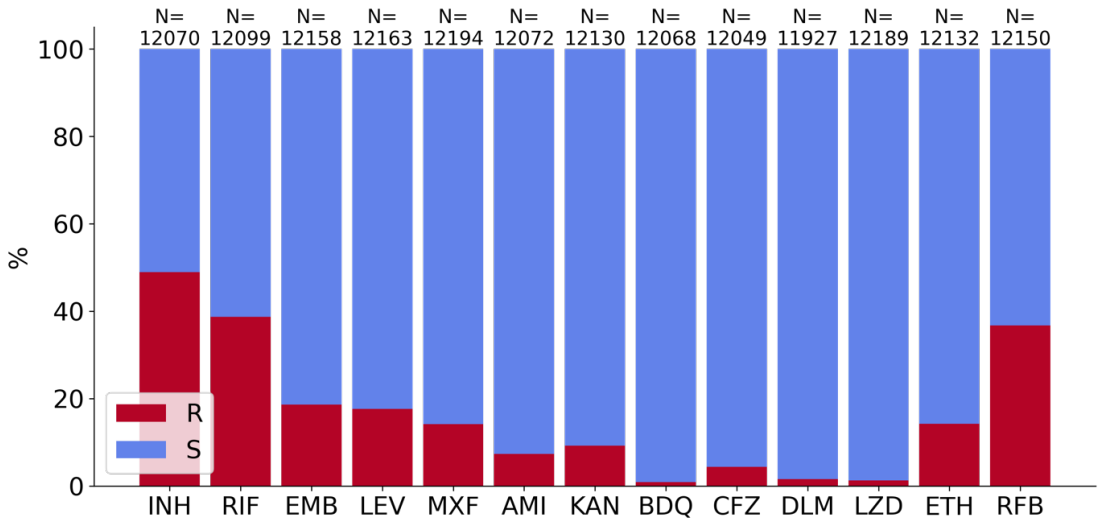

Figure 22 Prevalence of resistance to each of 13 drugs in the CRyPTIC dataset. N shows the total number of isolates with an MIC measurement (of any quality) for the corresponding drug.

Because all isolates with resistance to each of the drugs have corresponding whole genome sequencing data, I can survey the prevalence of known resistance conferring mutations present in the resistant isolates (Table 8). As resistance conferring mutations are not yet well characterised for the new and repurposed drugs (NRDs), bedaquiline, delamanid, clofazimine and linezolid, these have been omitted. For rifampicin, isoniazid, amikacin and kanamycin the most common resistance conferring mutation was present in over 50% of the resistant isolates. As expected, for isolates resistant to rifampicin or isoniazid, *rpoB* S450L and *katG* S315T were the most prevalent resistance conferring mutations respectively and more than half of isolates resistant to aminoglycosides contained the *rrs* a1401g mutation. Although a specific mutation was not present in the majority of ethambutol resistant isolates, greater than 50% of the isolates contained mutations at the *embB* M306 locus. For both fluoroquinolones, the most prevalent resistance conferring mutation was *gyrA* D94G followed by *gyrA* A90V.

DRUG	N	GENE	MUTATION	%
RIF	4,685	*rpoB*	S450L	62.2
			D435V	10.8
			H445D	4.3
			H445Y	3.1
			D435Y	2.4
INH	5,909	*katG*	S315T	77.5
		fabG1	c-15t	19.0
			g-17t	3.4
			t-8c	2.8
		inhA	I194T	1.0
EMB	2,261	*embB*	M306V	36.9
			M306I	20.9
			Q497R	14.1
			G406A	2.9
			G406D	1.5
KAN	1,121	*rrs*	a1401g	57.8
		eis	c-14t	5.1
			g-10a	2.8
AMI	883	*rrs*	a1401g	71.0
			g1484t	0.6
LEV	2,146	*gyrA*	D94G	35.1
			A90V	21.6
			D94N	7.1
			D94A	5.5
			S91P	4.1
MXF	1,724	*gyrA*	D94G	40.4
			A90V	18.3
			D94N	8.6
			D94A	5.0
			D94Y	3.7
ETH	1,727	*fabG1*	c-15t	49.4
			L203L	6.8

Table 8 The most prevalent mutations associated with phenotypic drug resistance in the CRyPTIC dataset. Depicted is a survey of the resistance-associated mutations present in CRyPTIC isolates [143, 144]. GENE: genic region of interest in which resistance conferring mutations can be found; MUTATION: most common resistance conferring mutations to each drug as seen in previous mutation catalogues [143, 144]. Non-synonymous amino acid mutations are denoted by upper case letters while nucleotide substitutions for non-coding sequences are denoted by lower case letters. Negative numbers denote substitutions in promoter regions; %: percentage of total phenotypically resistant isolates with the mutation, total number of resistant isolates to each drug is shown by N.

For each isolate, phenotypic measurements were taken for all 13 drugs in parallel and I could therefore examine co-occurrence of drug resistance, and whether resistance to a particular drug increases the likelihood of resistance to another within the dataset. Isolates containing all possible two-drug resistant combinations were present within the compendium (Figure 23). The most strongly associated resistant combination was ethambutol resistance with isoniazid resistance, where 98.5% of ethambutol resistant isolates were also resistant to isoniazid. The least frequent combination was isoniazid resistance with bedaquiline resistance – only 1.5% of isoniazid resistant isolates in the dataset had resistance to bedaquiline.

Resistance to any of the drugs was strongly associated with resistance to the first line drugs isoniazid and rifampicin (Figure 23). A higher proportion of rifampicin resistant isolates were resistant to isoniazid than isoniazid resistant isolates that were resistant to rifampicin. Isoniazid resistance was the most strongly associated with resistance to each of the other drugs bar rifabutin and moxifloxacin, where drugs from the same class are also being compared in these cases. These findings were expected as isoniazid resistance typically evolves prior to other drug resistance[92, 95].

Resistance to both drugs in the rifamycin class was common in the dataset; 96.8% of rifabutin resistant isolates were also resistant to rifampicin although significantly fewer rifampicin resistant isolates were resistant to rifabutin (91.3%, $p < 0.00001$) (Figure 23). In a similar fashion for the aminoglycosides, a smaller proportion of kanamycin resistant isolates were resistant to amikacin than amikacin resistant isolates that were resistant to kanamycin (72.0%, 90.4%, $p < 0.00001$). There were further differences between the two

aminoglycosides in that amikacin resistance was less commonly seen as a second resistant phenotype than kanamycin except for linezolid and delamanid resistant isolates which were more likely to have amikacin resistance. For the fluoroquinolones, a smaller proportion of levofloxacin resistant isolates were resistant to moxifloxacin than moxifloxacin resistant isolates that were resistant to levofloxacin (78.5%, 97.6%, $p < 0.00001$) and moxifloxacin resistance was less commonly seen as a second resistance phenotype than levofloxacin for all other drugs.

Of the second line drugs, levofloxacin and moxifloxacin were more commonly seen as a second resistant phenotype than the injectable drugs kanamycin and amikacin (Figure 23). Besides isoniazid, rifampicin and rifabutin, levofloxacin resistance was most strongly associated with resistance to each of the other second line agents and NRDs, even more so than the other first line drug ethambutol.

Isolates resistant to the NRDs (bedaquiline, clofazimine, delamanid and linezolid) were most likely to also be resistant to isoniazid, followed by rifampicin and rifabutin (Figure 23). The NRDs were less commonly seen as a second resistance phenotype, but within the NRDs co-occurrence of resistance was proportionally higher; bedaquiline, linezolid and delamanid resistance was commonly seen with clofazimine resistance (52.4%, 34.2% and 26.3% of isolates had co-resistance with clofazimine respectively). Most concerningly, resistance to the two new drugs, delamanid and bedaquiline, was seen in combination despite WHO recommendations not to use the drugs in combination to prevent development co-resistance[281] (12.9% of bedaquiline resistant isolates were resistant to delamanid and 7.1% of delamanid resistant isolates were resistant to bedaquiline). Compared to the other NRDs,

a particularly high proportion of linezolid resistant isolates were resistant to second line injectable drugs and high proportions of linezolid and bedaquiline resistant isolates had fluoroquinolone resistance.

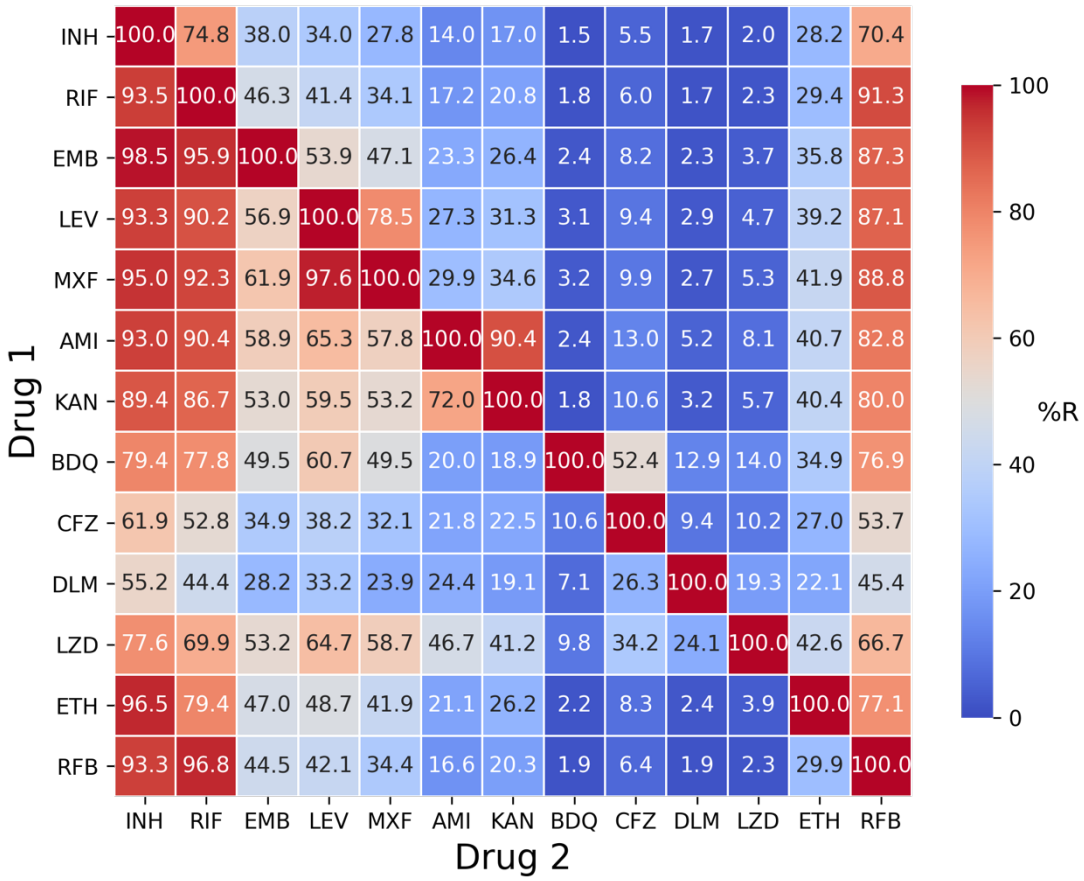

Figure 23 Co-occurrence of antibiotic resistance in CRyPTIC M. tuberculosis isolates. The heatmap shows the probability of an isolate being resistant to Drug 2 if it is resistant to Drug 1. Only samples with phenotypes for both Drug 1 and Drug 2 are included.

3.3.3 Clinically important resistance phenotypes

For the purpose of describing the prevalence of clinically important resistance categories (e.g. MDR, XDR) present in the dataset I assumed that all MICs that could not be read had susceptible phenotypes. Consequently, the calculated prevalence of R (resistant to at least

one drug), MDR, XDR etc. in the dataset are likely underestimates. Within the compendium, 55.4% of isolates had resistance to at least one of the 13 drugs, with the remainder assumed totally drug susceptible (Figure 24). Of these 6,184 drug resistant isolates, 68.8% were either RR or MDR. Of the RR/MDR isolates, 38.8% were pre-XDR and 3.0% were XDR. The proportion of pre-XDR is surprisingly high given that the WHO estimates RR/MDR cases with fluoroquinolone resistance to be around 20%[110]. One hypothesis is that a large number of resistant samples were contributed by India, an area associated with high levels of fluoroquinolone resistance (Figure 25)[28]. However, even without including samples of Indian or Nepalese (another country associated with high levels of fluoroquinolone resistance[282]) origin, the proportion of RR/MDR isolates in the dataset that had fluoroquinolone resistance was 33.0% (95% CI 31.5-34.6%) suggesting that the WHO estimate may be too low.

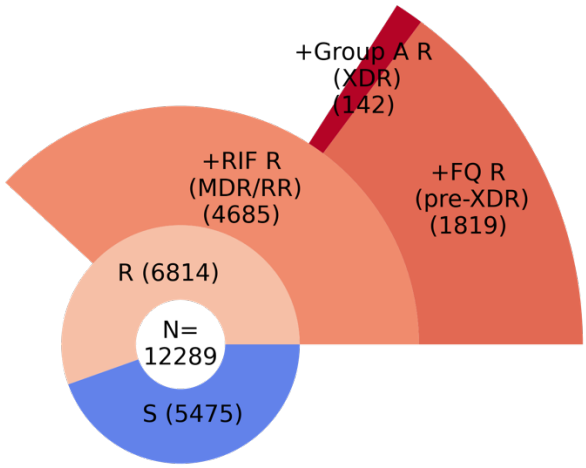

Figure 24 Phenotypes of 12, 289 CRyPTIC isolates with a binary phenotype for at least one drug. Where an isolate did not have a phenotypic reading for a particular drug the isolate was assigned susceptible to that drug for this analysis. R = resistant to at least one drug.

Two of the XDR isolates returned a resistant phenotype to *all* 13 of the drugs assayed and therefore can be described as totally drug resistant (TDR). One TDR isolate belonged to

Lineage 4 and was contributed by South Africa, and the other belonged to Lineage 2 with an unknown country of origin contributed by Sweden (Table 9). Given that the two isolates are from different lineages, the TDR isolates cannot be part of the same outbreak.

UNIQUEID	COUNTRY OF ORIGIN	LINEAGE
site.11.subj.XTB-18-224.lab.XTB-18-224.iso.1	UNKNOWN, processed in Sweden	Lineage 2
site.10.subj.YA00026182.lab.YA00026182.iso.1	South Africa	Lineage 4

Table 9 Sample information for isolates classified as resistant to all 13 CRyPTIC drugs tested.

The resistant isolates originated from 18 out of the 20 countries that contributed samples, although 818 of the resistant isolates had an unknown country of origin. Peru contributed the most resistant isolates (n = 1260), followed by South Africa (n = 988) and India (n = 913) (Figure 25). The proportion of resistant isolates in each country's contribution varied between 2.6% and 100%, but this reflects the different sampling strategies used in different countries rather than the prevalence of resistance. RR/MDR, pre-XDR and XDR isolates were found in isolates from all countries that contributed > 100 resistant isolates bar Peru, Vietnam and Nepal where no XDR isolates were collected (Figure 25). The relative proportions of RR/MDR, pre-XDR and XDR isolates are also likely reflective of sampling strategy, for example Vietnam and Brazil sampled a high proportion of non-MDR/RR resistant phenotypes compared to other countries; 73.9% and 55.1% of resistant isolates contributed by Vietnam and Brazil, respectively, were neither MDR nor RR, whereas the proportion for other countries varied between 1.4% and 40.3%. A high proportion of RR/MDR resistant isolates sampled by Nepal and India were either pre-XDR or XDR (92.9% and 69.8% of RR/MDR isolates), which is consistent with previous observations of a high prevalence of fluoroquinolone resistance in these countries[193, 282].

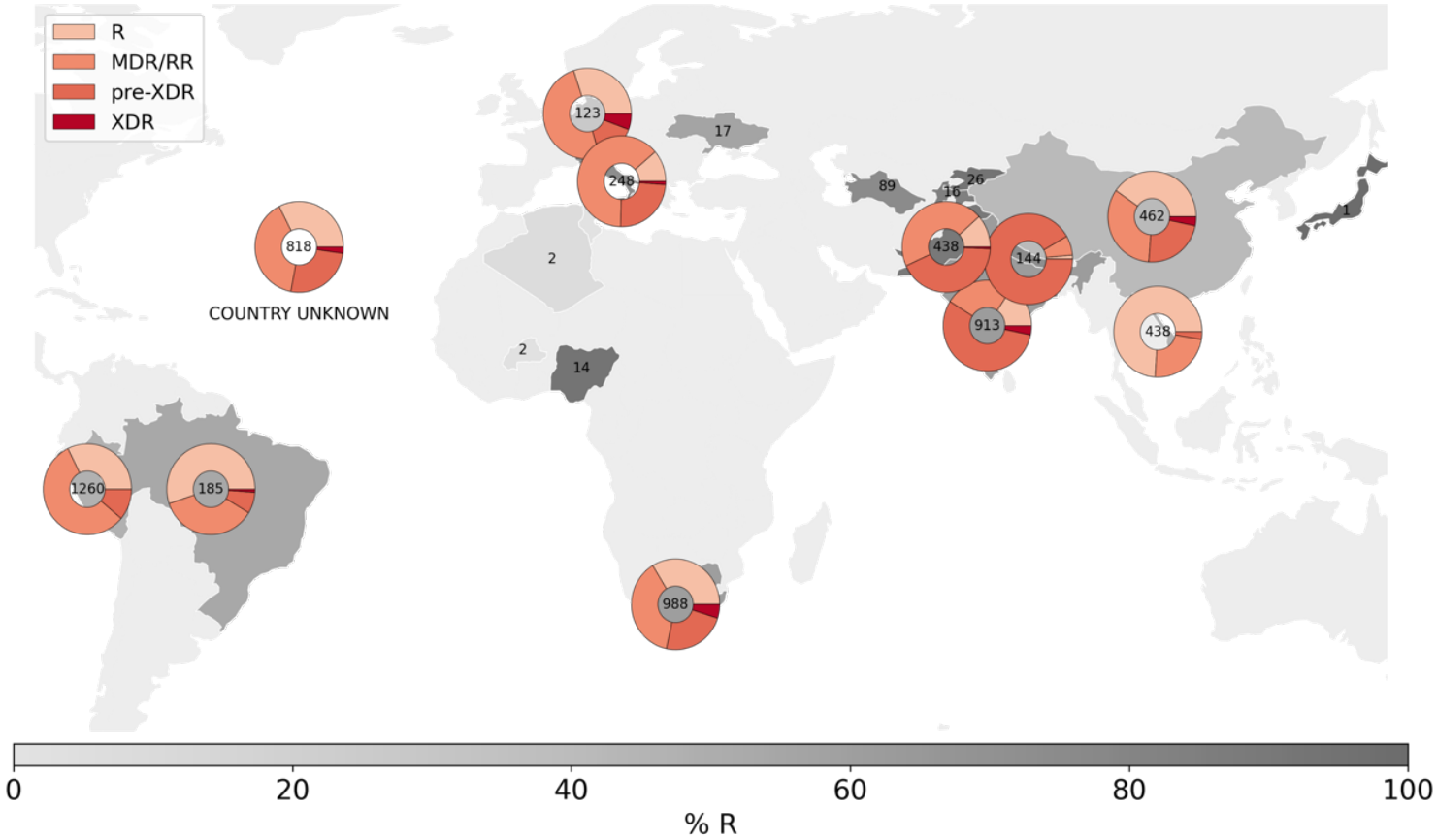

Figure 25 Geographical distribution of 6, 184 drug resistant CRyPTIC isolates and distribution of clinically important resistance phenotypes. Numbers represent the number of drug resistant isolates contributed by each country. Intensity of grey shows the percentage of isolates that were resistant in contributions from each country. Donut plots show the proportions of clinically important resistant phenotypes (RR (rifampicin resistant) /multi-drug resistant (MDR), pre-extensively drug resistant (pre-XDR) and XDR) identified in countries contributing > 100 drug resistant isolates.

Isolates with resistance to at least one drug were present within the four major lineages in the dataset; 66.7% of Lineage 3 isolates, 59.1% of Lineage 2 isolates, 49.1% of Lineage 4 isolates and 35.6% of Lineage 1 isolates had resistance to at least one drug. The highest number of resistant isolates had a Lineage 4 background (3,043 isolates), followed by Lineage 2 (2,886), Lineage 3 (625) and Lineage 1 (256). RR/MDR, pre-XDR and XDR resistance phenotypes were well represented in the four major *M. tuberculosis* lineages (Figure 26). The relative proportions of resistance categories will have been influenced by different resistance sampling approaches in different countries as the lineages are geographically distinct (see Section 3.3.1). Bearing this in mind, within the compendium, Lineage 3 isolates contained the most MDR/RR isolates as a proportion of resistant isolates (77.6%), Lineage 2 isolates contained the most pre-XDR isolates as a proportion of MDR/RR isolates (54.2%) and Lineage 2 contained the most XDR isolates as a proportion of MDR/RR isolates (4.7%) (Figure 26).

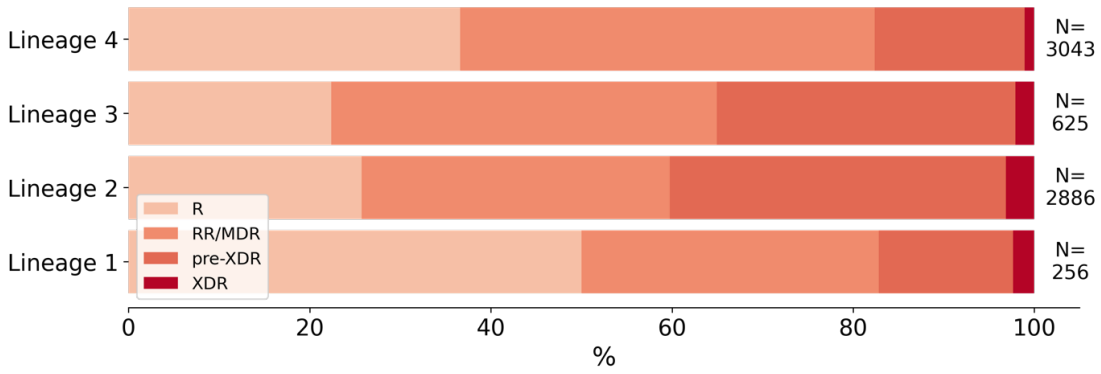

Figure 26 Proportions of resistance phenotypes in the 4 major *M. tuberculosis* lineages. N is the number of isolates of the lineage called resistant to at least one of the 13 drugs. Acronyms: R = resistant (to at least one of the 13 drugs but not rifampicin), MDR = multi drug resistant, RR = rifampicin resistant, XDR = extensively drug resistant.

I next investigated the levels of resistance to additional drugs in the different phenotypic backgrounds including rifampicin susceptible, RR/MDR, pre-XDR and XDR.

Of the isolates that were rifampicin susceptible, 20.3% were resistant to isoniazid and therefore could be described as isoniazid mono-resistant; this phenotype is seen in 1,470 isolates, representing a significant portion of resistance in the dataset (Figure 27a). The next most prevalent resistance in a rifampicin susceptible background was ethionamide resistance, which is somewhat expected as resistance to ethionamide can occur via the same mutations in the *inhA* promoter as isoniazid resistance[283, 284]. Concerningly, resistance to all second line and NRD drugs was seen in the rifampicin susceptible background suggesting that resistance to secondary agents can occur prior to development of RR/MDR. Further, clofazimine, levofloxacin, kanamycin and moxifloxacin resistance was significantly more prevalent than resistance to the first line drug ethambutol in this phenotypic background. Clofazimine and levofloxacin resistance (3.4% and 2.8% respectively) was significantly higher than all the other drugs bar isoniazid and rifampicin, and resistance to group A agents, bedaquiline and linezolid, was significantly lower than any other drug (0.3% and 0.6% respectively).

For rifampicin resistant isolates, 93.5% were resistant to isoniazid (Figure 27b), meaning a total of 4,353 isolates in the dataset satisfy the definition of MDR and 302 of these isolates are rifampicin mono-resistant (6.5%). This is higher than the WHO's estimated global prevalence of 1%[110], although a large number of resistant samples in this study were from South Africa, where there is a high prevalence of rifampicin mono-resistance[285]. After removing the samples of South African origin and repeating the analysis, the proportion of

rifampicin mono-resistant samples in the dataset was 4.7%, suggesting the WHO estimate may be too low. Asides from rifabutin resistance, the next most common resistance was to the other first line drug ethambutol (46.3%) and significantly more rifampicin resistant isolates were resistant to ethambutol than any fluoroquinolone, aminoglycoside or NRD. Significantly more of the rifampicin resistant isolates were resistant to either of the fluoroquinolones than either of the aminoglycosides and there was significantly more levofloxacin resistance present (41.4%) than moxifloxacin resistance (34.1%) and significantly more kanamycin resistance was present (20.8%) than amikacin resistance (17.2%). The prevalence of resistance to any of the NRDs in this background was lower than for any second line drug. Of the NRDs, clofazimine resistance was seen at significantly higher prevalence (6.0%) than linezolid (2.3%), delamanid (1.7%) and bedaquiline (1.8%) for which there was no significant difference in resistance prevalence to the three drugs.

For pre-XDR isolates, more were resistant to levofloxacin (98.3%) than moxifloxacin (80.1%) and nearly all (99.0%, CI 98.5-99.6%) of the isolates were resistant to isoniazid (Figure 27c). As seen with RR isolates (Figure 27b), significantly more kanamycin resistance was present (33.8%) than amikacin resistance (29.1%) and the prevalence of resistance to any of the NRDs in pre-XDR isolates was lower than for any second line drug. Within the NRDs, clofazimine resistance was significantly more prevalent (9.3%) than resistance to any other NRD.

In XDR isolates, significantly more were resistant to linezolid (65.7%) than bedaquiline (44.6%) and 11.1% of isolates were resistant to both the group A drugs (Figure 27d). Nearly all the XDR isolates were isoniazid resistant (97.9%), and it is possible that all isolates were

truly isoniazid resistant as the confidence interval included 100%. As with the pre-XDR isolates, a greater proportion of XDR isolates were resistant to levofloxacin (97.9%) than moxifloxacin (86.4%), but the difference was not significant at $p < 0.05$. The prevalence of resistance to NRDs was not significantly lower than for second line injectable drugs, as was seen for pre-XDR isolates (Figure 27c), except for delamanid resistance which was significantly lower than for any other drug at 18.8%.

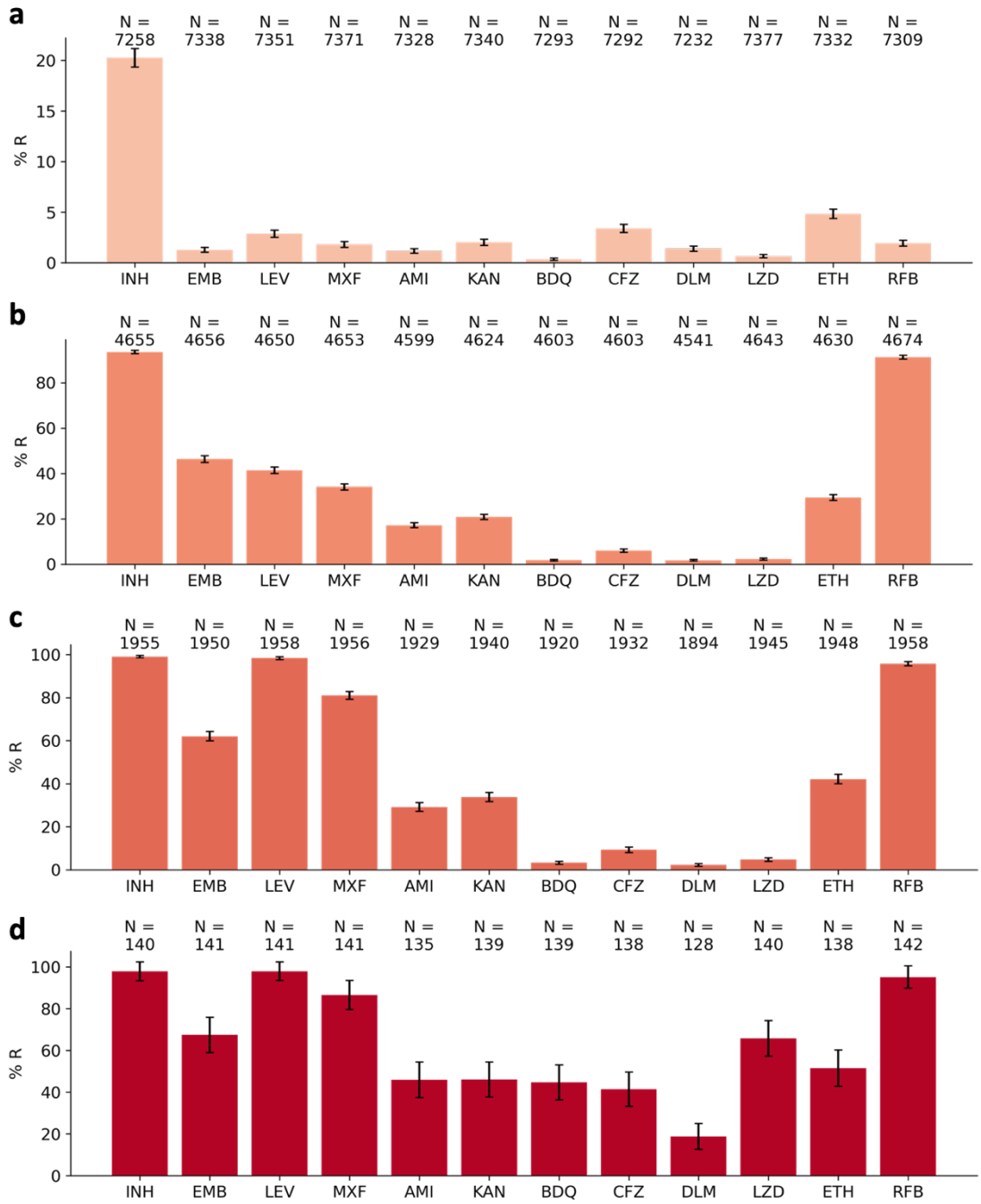

Figure 27 Percentage of isolates that are resistant to additional drugs in a background of (a) rifampicin susceptible, (b) rifampicin resistant (RR/MDR), (c) rifampicin resistant and resistant to a fluoroquinolone (pre-XDR), (d) rifampicin resistant and resistant to a fluoroquinolone and resistant to either bedaquiline or linezolid (XDR). Only samples with phenotypes for rifampicin and the additional drug are included, and this number is indicated by N. Error bars show 95% confidence intervals calculated using the method of WIlson[279]

3.4 Discussion

Through the scale of the project, the collection of isolates from across four continents and 23 countries, and the emphasis on oversampling for resistance, the CRyPTIC compendium provides an invaluable and diverse reservoir of information regarding resistance to the 13 antitubercular drugs studied. Via the simultaneous collection of phenotypic and genetic information for all 13 drugs in parallel, this resource also offers a uniquely detailed view into combinations of drug resistance and susceptibility.

A particular strength of the CRyPTIC compendium is the data collated for second-line drugs, for which global data from the past 15 years is less substantial than for first line drugs, due to limited drug susceptibility testing[110]. The dataset shows that a greater proportion of RR/MDR isolates were additionally resistant to a fluoroquinolone than any other second line drug with 41.4% and 34.1% of RR/MDR isolates being resistant to levofloxacin and moxifloxacin respectively (Figure 27b). Comparatively, the WHO estimated RR/MDR cases with fluoroquinolone resistance to be around 20%; the proportions present in the compendium may be higher due to the large number of resistant samples contributed by areas associated with a high level of fluoroquinolone resistance. However when excluding samples of such orgins from the dataset the proportion of RR/MDR isolates in the dataset that had fluoroquinolone resistance was still signifcantly higher than the 20% estimate (see Section 3.3.3).

It is also concerning that in a rifampicin susceptible background, resistance to both fluoroquinolones and second line injectable drugs was seen significantly more than resistance to another first line drug, ethambutol (Figure 27a). This suggests that there is a significant level of resistance to second-line drugs that pre-dates the evolution of RR/MDR, as corroborated by a systematic review that found patients previously prescribed fluoroquinolones (even if this was for a non-mycobacterial infection) were three times more likely to have fluoroquinolone resistant TB[190]. I therefore hypothesise that these patients had latent TB and when using fluoroquinolone treatment for another (non-mycobacterial) infection, any actively replicating bacilli outside of granulomas[15-18] or slow replicating bacilli within granulomas acquired resistance, then once the latent TB progresses to active disease the bacilli are resistant.

Isolates resistant to any individual drug were more likely to be resistant to fluoroquinolones than second line injectable drugs (Figure 23). This could be due to more widespread use of fluoroquinolones as they are recommended over injectable drugs because of their comparative ease of administration[110]; fluoroquinolones are recommended for most RR/MDR TB treatement regimens (Table 4). The careful stewardship, study and surveillance of fluoroquinolone resistance in both TB and other infectious diseases will be paramount for the success of TB treatment.

When building an MDR treatment regime, the appropriate selection of second-line drugs from each drug class could improve treatment outcomes, as there were differences in resistance between drugs of the same class (Figure 23). There was more resistance to levofloxacin and kanamycin than moxifloxacin and amikacin respectively in all phenotypic

backgrounds (Figure 27a-d) which suggests that amikacin and moxifloxacin may be the most appropriate drugs to recommend as part of treatment regimens. This also supports, from a resistance standpoint, recommendations for switching from kanamycin to amikacin when treating MDR TB patients[286]. Interestingly, 28.0% of kanamycin resistant isolates were not resistant to amikacin and 21.5% of levofloxacin resistant isolates were not resistant to moxifloxacin and therefore these isolates could still be treatable with the respective drug class. However, these conclusions are dependent on the cut-off used to infer resistance (see section 3.2.4).

This dataset is the first global survey of resistance to the NRDs bedaquiline, clofazimine, delamanid and linezolid. Reassuringly, fewer isolates were resistant to these drugs than for first and second line drugs in the dataset both as a whole and in RR/MDR and pre-XDR backgrounds (Figure 22, Figure 27b,c). However, similarly to the second-line drugs there may be some propensity to the development of resistance to clofazimine and delamanid prior to some first-line drugs; there was a significantly higher level of resistance to these drugs compared to ethambutol in a rifampicin susceptible background (Figure 27a). One hedges that this result is dependent on the ECOFFs used to infer resistance, which may be less reliable for the NRDs as comparatively fewer resistant samples were collected[272].

Co-resistance between the NRDs was also common in the dataset, and the link previously observed between bedaquiline and clofazimine resistance[287] is evident; 52.4% of bedaquiline resistant isolates were also resistant to clofazimine. Of particular concern is that co-resistance between the two new drug classes was seen despite recommendations against use of the drugs in combination specifically to prevent development of co-resistance[281];

12.9% of bedaquiline resistant isolates had delamanid resistance and 7.1% of delamanid resistant isolates had bedaquiline resistance (Figure 23). As the rate of spontaneous evolution of resistance to bedaquiline and delamanid is comparable to the first line drugs rifampicin and isoniazid respectively[288], vigilant stewardship and genetic and phenotypic surveillance of these compounds is of utmost importance to ensure their clinical utility lasts for as many years as possible.

Molecular diagnostic tests have increased global capabilities to detect resistance, these tests work by identifying the most common genetic determinants for resistance to different drugs (see Section 1.3.2). The survey of common resistance conferring mutations present in resistant isolates lends support to the likely effectiveness of molecular diagnostic tests; for most first and second line drugs > 50% of the isolates the carried the most common resistance conferring mutation to the corresponding drug (Table 6).

Many molecular diagnostic tests infer isoniazid resistance upon detection of rifampicin resistance, and although this a good proxy (93.5% of rifampicin resistant isolates in the dataset were isoniazid resistant (Figure 23)), reliance solely on these tests would mean RR/MDR cases become inseparable. The tests may be particularly unreliable in regions with high prevalence of rifampicin mono-resistance[289] where such patients could be denied effective isoniazid treatment. Furthermore, there are implications for resistance surveillance as studies have shown that these isolates are qualitatively different; it has been shown that there are differences in the genetic determinants of rifampicin mono-resistance and MDR[1, 270]. In this dataset, 302 of the RR/MDR isolates (6.5%) could be described as rifampicin mono-resistant (Figure 27b).

The emphasis on rifampicin resistance testing could also mean the detection and surveillance of isoniazid mono-resistant cases is overlooked; these cases are important as they can rapidly evolve into MDR cases[290]. Of the rifampicin susceptible isolates in this study, 20.3% were resistant to isoniazid (Figure 27a) and therefore the compendium contains 1470 isoniazid mono-resistant isolates, significantly more than rifampicin mono-resistant isolates (n = 302) which is consistent with a global prevalence survey[291].

The CRyPTIC dataset is biased in several ways. Firstly, the dataset is enriched for resistance and therefore prevalence of resistance to all drugs in the dataset and specific resistance phenotypes will be much higher than in reality (Figure 22, Figure 24). Secondly, not all countries contributed the same number of isolates, therefore the dataset will be more reflective of some countries than others and the extent of the bias will be difficult to unravel as country metadata for 2,369 samples was not provided (Figure 21). Furthermore, as some countries oversampled for resistant isolates more than others, the dataset will be more reflective of the resistance present in some countries than others (Figure 25). Thirdly, different countries sampled resistance differently, with some prioritising MDR or mono-resistance, there is therefore limited ability to make compare the types of resistance observed in different countries (Figure 25). Finally, lineages were not collected representatively with Lineage 4 and Lineage 2 isolates dominating the dataset (Figure 21). As lineages are geographically distinct (see Section 3.3.1), the same limitations apply for making comparisons between lineages as for countries in this dataset.

In addition to bias, the dataset has several other limitations. For example, the ECOFFs used to define resistance could be wrong as it is difficult to culture and read growth of *M.*

tuberculosis on 96 well plates (although CRyPTIC did use three different reading techniques to try and minimise errors, see Section 3.2.3). Further, because the plates have limited numbers of wells, the MIC dilutions are truncated (Figure 19), and therefore isolates that have an MIC over the limit of detection will not be accurately measured. In addition, although some samples with known labelling errors were removed during quality control, others will remain and are difficult to identify. Another limitation is the comparatively restricted sample metadata, for instance the consortium was unable to measure MICs for the first line drug pyrazinamide, and other useful information, for example regarding any previous TB treatment, was not able to be recorded.

Despite its limitations, this dataset provides the TB community with a greater insight into global resistance patterns and has the potential to facilitate and inspire numerous future studies through the wealth of both genetic and phenotypic data provided. Although the dataset presents exciting opportunities to better understand drug resistance in TB, the finding of two unrelated isolates that were resistant to all 13 drugs tested also provides a stark warning of the future if we do not successfully diagnose and treat drug resistant infections (Table 9).

4 Chapter 4: Fluoroquinolone resistance case study

4.1 Introduction

As I discussed in the previous chapter, the CRyPTIC dataset offers a wealth of information about resistance to second line antitubercular drugs (see Section 3.4), for which there is limited global data due to a paucity of drug susceptibility testing. Of particular importance is the study of fluoroquinolone resistance. Levofloxacin or moxifloxacin are recommended for all MDR TB treatment regimens (Table 4), and the WHO now estimates that around 20% of MDR TB isolates have fluoroquinolone resistance[47,12], and the CRyPTIC data suggests the true proportion could be significantly higher (see Section 3.3.3). Reliance on fluoroquinolone drugs for TB treatment may further increase as the WHO recommends the phasing out of other second line antibiotics (amikacin, kanamycin and streptomycin[68]) and fluoroquinolones may also form the backbone of a shorter treatment regimen for drug susceptible TB[69]. It is therefore imperative that fluoroquinolone resistance is diagnosed as early as possible to treat patients and prevent further spread of fluoroquinolone resistant strains. However worldwide, only 50% of MDR *M. tuberculosis* isolates were tested for fluoroquinolone resistance in 2020[12].

The WHO catalogue of resistance associated mutations has identified 14 DNA gyrase mutations which it recommends using for the detection of both levofloxacin and moxifloxacin resistance by WGS[42, 145] (Table 5). The success of catalogue-based prediction relies on the quality and comprehensiveness of the resistance catalogue and thus our understanding of the mechanisms of resistance. For levofloxacin and moxifloxacin resistance prediction the catalogue achieved sensitivities of 84.4% and 87.7% respectively and

specificities of 98.3% and 91.6% respectively on the dataset of 38,000 isolates used to produce it. Due to overfitting performance is likely an upper limit and so it is particularly noteworthy that more than 10% of fluoroquinolone resistance was not explained by the mutation catalogue, despite acquired fluoroquinolone resistance being well characterised as attributable to the *gyrA* and *gyrB* genes[292]. A recent genome wide association study of the CRyPTIC isolates found only one other gene containing a variant associated with fluoroquinolone resistance which was seen at very low frequency[205]. This is a paradox since it suggests that the determinants of fluoroquinolone resistance are still not fully understood but there are no obvious candidate genes or variants.

Although WGS technology offers a myriad of additional diagnostic benefits (see Section 1.3.3), it is not yet feasible in many areas of high TB burden. Commercially available molecular diagnostic tests that detect specific resistance conferring mutations could be a more appropriate tool to increase fluoroquinolone resistance diagnosis and surveillance in these regions (see Section 1.3.2). For rifampicin resistance detection, the GeneXpert MTB/RIF molecular diagnostic test has facilitated increased testing for RIF resistance in areas of high TB burden [110, 129]. A range of molecular diagnostic tests have been developed to diagnose fluoroquinolone resistance and some tests have now been approved by the WHO (Table 10).

Assay	*Manufacturer*	*Type*	*Wild type probes*	*Resistant mutation probes*	*Susceptible probes/ agnostic*	*References*
AID TB FQ/EMB Kit	Autoimmun Diagnostika	LPA	*gyrA* 90, 91, 94	*gyrA* A90V, S91P, D94A, D94N, D94Y, D94G		293
Anyplex II MTB/MDR/XDR	Seegene	Multiplex real-time PCR		*gyrA* A90V, S91P, D94A, D94G, D94H, D94N, D94Y		294
Genoscholar FQ+KM-TB II	Nipro	LPA	*gyrA* 88-97	*gyrA* A90V, D94A, D94G	*gyrA* S95T	295
GenoType MTBDRsl v1	**Hain Lifescience**	**LPA**	***gyrA* 85:96**	***gyrA* A90V, S91P, D94A, D94N/Y, D94G, D94H**	***gyrA* S95T**	**49, 135, 296**
GenoType MTBDRsl v2	**Hain Lifescience**	**LPA**	***gyrA* 85:96, *gyrB* 497:502**	***gyrA* A90V, S91P, D94A, D94N/Y, D94G, D94H *gyrB* N499D, E501V**	***gyrA* S95T**	**49, 135, 296-298**
MeltPro MTB/FQ	Zeesan Biotech	Probe-Based Melting Curve Analysis	*gyrA* 88:94			299
REBA MTB/XDR	YD Diagnostics	LPA	*gyrA* 79:98	*gyrA* G88A, G88C, A90V, S91P, D94A, D94G, D94H, D94N, D94Y		300
Xpert MTB/XDR	**Cepheid**	**Semi-quantitative nested PCR + high resolution melt technology**		***gyrA* G88A, G88C, A90V, S91P, D94A, D94G, D94H, D94N, D94Y *gyrB* N461D, N461V, N499T, E501D, E501V**	***gyrA* S95T**	**49, 131, 301, 302**

Table 10 PCR based molecular diagnostic tests for fluoroquinolone resistance and the DNA gyrase mutations that they can detect. Rows in bold are WHO approved tests. LPA = line probe assay.

Each test varies in, and is fundamentally limited by, the mutations that it detects. For example, GeneXpert MTB/RIF only detects mutations in codons 428 to 452 of the *rpoB* gene

and hence was unable to detect the *rpoB* I491F mutation behind an MDR outbreak in Eswatini[132]. Molecular diagnostic tests can also give false positive resistance diagnosis due to other mutations present in the region of interest; this could prevent patients being treated with the most effective drugs and result in either more toxic or last line antitubercular drugs being used unnecessarily. For instance, a mutation that does not confer resistance, *gyrA* A90G, prevents binding of a wild-type probe in the Genotype MTBDRsl v1 and v2 assays leading to a test interpretation of resistant to fluoroquinolones[135, 136]. Another mutation *gyrA* S95T, is not associated with resistance but is highly prevalent and located within the *gyrA* QRDR[145] and as such molecular diagnostics tests have been designed to either explicitly detect or be agnostic to this mutation[135, 301]. However, the presence of other mutations, including synonymous mutations within the QRDR could disrupt probe binding and impact the ability of molecular tests to correctly diagnose resistance or susceptibility.

Both WGS with catalogue-based resistance prediction and molecular diagnostic tests rely on certain assumptions and therefore may have limitations if these assumptions are not met. Firstly, one assumes that a resistance associated mutation always results in an isolate having resistance to the corresponding drug. This is not necessarily true for the fluoroquinolones; the combination of *gyrA* A90G and *gyrA* T80A mutations, found in an African *M. tuberculosis* sub-lineage, restores fluoroquinolone susceptibility in isolates with a resistance conferring *gyrA* D94N mutation[303].

By using the same catalogue or molecular diagnostic test in all geographic locations, one also assumes that the genetic background of *M. tuberculosis* isolates does not affect the prevalence or level of resistance conferred by a particular mutation. The local prevalence of

a particular mutation could be an important consideration when choosing a diagnostic tool, especially if a molecular diagnostic test does not detect that specific resistance conferring mutation. *In vitro*, the mutational profile and minimum inhibitory concentrations of fluoroquinolone resistant *M. tuberculosis* isolates has been shown to be dependent on genetic background[304]. Different geographic, lineage and phenotypic backgrounds may also have differing effects *in vivo*; for example a study of Indian clinical *M. tuberculosis* isolates found a significantly higher frequency of some genetic mutations in the Beijing lineage (Lineage 2) compared to other lineages in pre-XDR and XDR TB isolates[305]. On the other hand, a study of Chinese clinical isolates found that the Beijing lineage had no effect on mutation frequency compared to the other lineages tested[306].

Finally, we assume that clinical resistance is based entirely on the genetic code of the *M. tuberculosis* isolate; this is unlikely to be true as epigenetic regulation from nucleotide methylation has been shown to be associated with resistance[307, 308]. Although an important avenue for future research, this is beyond the scope of the CRyPTIC project as it stands, due to the nature of sequencing data that was collected.

Another important question to ask when using catalogue-based resistance prediction and molecular diagnostic tests for fluoroquinolone resistance diagnosis, is to what level can they detect resistance conferring alleles in minority populations? Mixed populations are common in *M. tuberculosis* infections and are particularly implicated in fluoroquinolone resistance[202, 309]. Relatively few studies report on the minor alleles detected in fluoroquinolone resistant *M. tuberculosis* isolates, but a recent systematic review found 12 studies from small,

localised sample datasets and estimated the prevalence of mixed populations containing minor resistance conferring alleles to be around 10% of fluoroquinolone resistant isolates[108].

Although WGS can provide information about mixed populations, this is dependent on the sequencing depth achieved and the variant callers and bioinformatic pipelines used because filters and thresholds are put into place to screen out sequencing errors[310]. Some bioinformatics pipelines may choose to omit minor alleles; indeed alleles seen at a fraction read support (FRS) for the major allele of less than 90% were not included in the compilation and evaluation of the WHO catalogue[145]. A previous study found that minor alleles were responsible for 11-38% of the true resistance predictions made for aminoglycoside and fluoroquinolone resistance in WGS *M. tuberculosis* isolates (although the confidence intervals were large due to small sample size)[311].

Molecular diagnostic tests have also been shown to detect fluoroquinolone resistance in minor populations, but the limit of detection varies between tests. For instance the line probe assays, GenoType MTBDRsl v1 and GenoScholar-FQ + KM TB II, have been shown to pick up *gyrA* D94G alleles present in a >= 5% minor population[312], yet Xpert MTB/XDR could only detect resistance conferring alleles in >=25% of the population[313].

The aim of this chapter is to use the CRyPTIC isolates to increase our knowledge about the genetic determinants of fluoroquinolone resistance. In doing so, I aim to evaluate how reliable the assumptions of sequence based molecular diagnostics are and assess their likely performance for fluoroquinolone resistance detection in the CRyPTIC isolates.

4.2 Methods

4.2.1 Dataset

All data used were from the CRyPTIC data compendium[1] described in Chapter 3: Description of the CRyPTIC dataset. Please see https://github.com/alicebrankin/thesis_notebooks for the codebase to reproduce the analyses and figures in this chapter.

4.2.2 Evaluation of the mutations used by catalogues or molecular diagnostic tests for identifying resistance in levofloxacin and moxifloxacin WGS isolates

From the WHO 2021 catalogue I extracted mutations that were categorised as 1) associated or 2) potentially associated with resistance for levofloxacin and moxifloxacin[42, 145]. For molecular diagnostic tests, specific mutations detected by mutant probes and regions where a mutation would result in disruption of binding of a wild-type probe were identified using test package inserts or literature (Table 10).

I evaluated the sensitivities and specificities of both the WHO catalogue and a range of molecular diagnostic tests for identifying resistance using CRyPTIC isolates, having excluded MIC measurements in which we have a low confidence for each drug. For the tests, where a genetic region is interrogated by a wild-type probe, any mutation within that region was considered detected by that test, and therefore called resistant, unless the test was agnostic to that mutation. Where the CRyPTIC bioinformatic pipeline reported no evidence (a so-called 'null' call) or when there was evidence of variation but it failed the statistical checks (a 'filter fail') at a position within the region covered by a probe, these mutations were not predicted resistant. Where tests are designed to be agnostic to certain mutations known not

to be associated with resistance, for example *gyrA* S95T (Table 10), they were also not predicted resistant. As some molecular tests can identify both specific mutations and the presence of any mutation within a genetic region, they can be interpreted differently. For example, resistance can be directly detected where a specific mutation is detected by a mutant probe, or resistance can be inferred where the test detects any mutation within a region covered by a wild-type probe. The sensitivity and specificity analyses were performed separately for the two possible interpretations of these tests.

4.2.3 Statistical analysis

When comparing prevalence in two populations a two-proportion z-test was used. Where 95% confidence intervals were required for individual percentages, these were calculated using the method of Wilson[277]. Multiple logistic and linear regression (ordinal least squares) models, and subsequent Bonferroni corrections , were implemented using the Python3 package statsmodels[314]. For regression models, categorical variables were dummy encoded, with the largest group in each variable being used as the reference. A statistical significance of $p < 0.05$ was used throughout.

4.2.4 Identification of mixed alleles

Throughout this chapter I considered any allele with a fraction read support (FRS) >= 0.9 as homogeneous and an allele with FRS < 0.9 and at least two reads supporting an alternative allele as one of two or more alleles in a mixed population. Assuming the error rate of Illumina sequencing is ~1%, if two or more reads support an alternative allele, it is unlikely that this is due to sequencing error.

Since the variant caller (Clockwork) used by the CRyPTIC project was setup conservatively, with only variants having an FRS >= 90% being called, variants with FRS < 90% were recorded as 'filter fails'. Therefore, for isolates that had a filter fail at a position in *gyrA* or *gyrB*, the regenotyped variant call format (.vcf) files were interrogated (as these have an entry at all variable positions). From the regenotyped .vcf file, FRS for each of the reference and all possible alternative alleles were calculated at that position. Please see Sections 3.2.1 and 3.2.2 for more information about 'regenotyped' .vcf files and processing of the CRyPTIC sequencing data. Where an alternative allele was present at a position with a minimum sequencing depth filter fail as flagged by Minos[273], the allele was not included and hence assumed wild type for analyses. The filter for minimum genotype confidence percentile was not used to exclude any variants because the filter is partially dependent on the FRS[273]. Due to the large file sizes, and subsequent compute time required to interrogate them, only regenotyped .vcf files for samples that had a phenotypic measurement for either moxifloxacin or levofloxacin (12,354 samples) were included for the analysis of mixed alleles in the DNA gyrase genes in Section 4.3.7.

4.3 Results

4.3.1 Overview of fluoroquinolone resistance in the CRyPTIC dataset

The CRyPTIC dataset contains a total of 2191 isolates that were resistant to a fluoroquinolone (as defined by the CRyPTIC consortium's proposed ECOFFs[272]), and the majority of these (76.6%) were resistant to both levofloxacin and moxifloxacin (Figure 28a). The 2191 isolates originate from 16 different countries; India contributed the highest number of fluoroquinolone resistant isolates, followed by South Africa, Pakistan, Peru, China and Nepal (Figure 28a). Each of the four major *M. tuberculosis* lineages are represented

within the fluoroquinolone resistant isolates, with Lineages 2 and 4 being the most prevalent (Figure 28b). The majority (88.4%) of fluoroquinolone resistant isolates had a phenotypic background of MDR, but other phenotypic backgrounds (isoniazid and rifampicin susceptible, isoniazid resistant and rifampicin susceptible or isoniazid and rifampicin resistant) are also represented (Figure 28c).

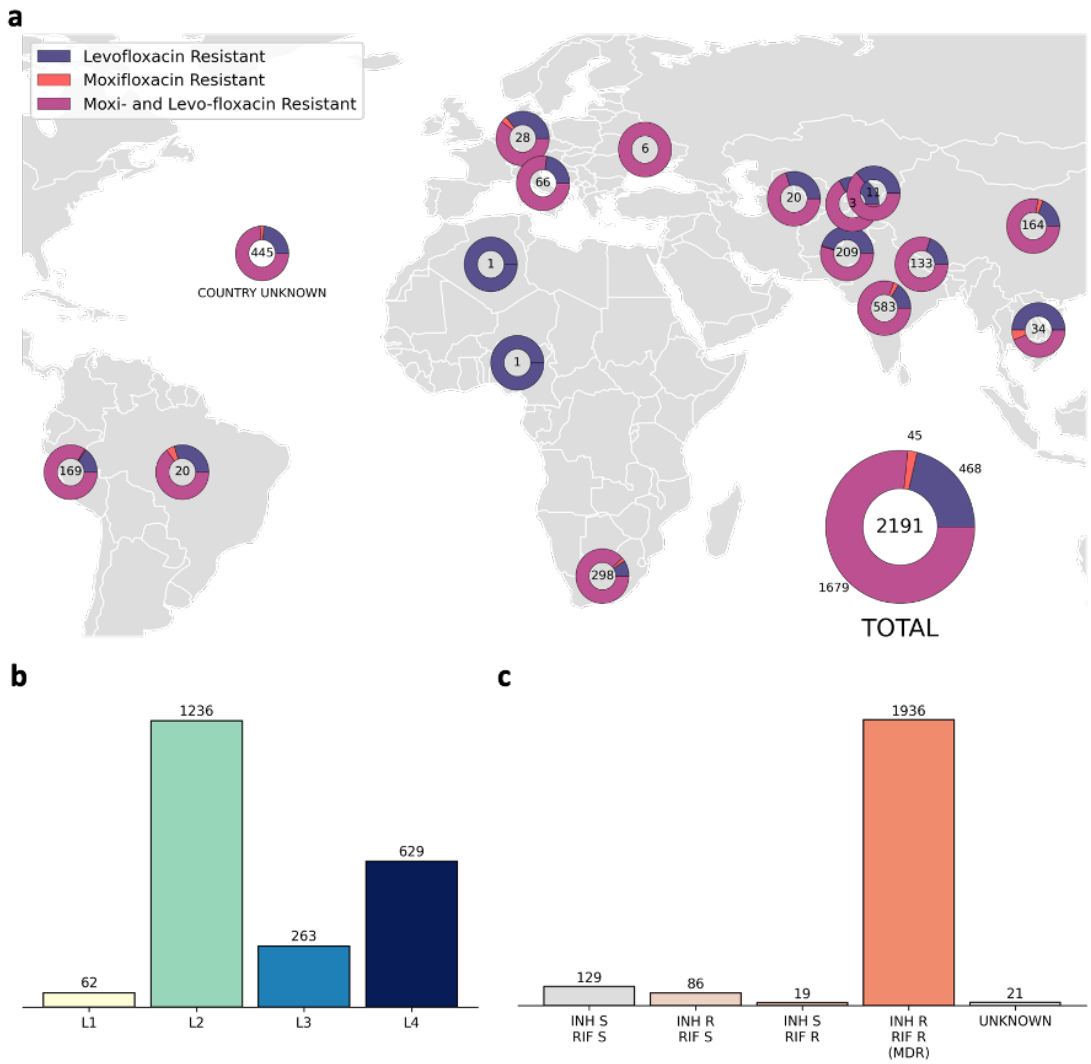

Figure 28 Overview of 2191 fluoroquinolone resistant isolates in the CRyPTIC compendium. a) Number of fluoroquinolone resistant isolates contributed by each country. Pie charts show the proportion of these isolates that were resistant to levofloxacin, moxifloxacin or both fluoroquinolones for each country, for isolates where the country of origin is unknown, and in total. b) Number of fluoroquinolone resistant isolates from *M. tuberculosis* Lineages 1, 2, 3 and 4. c) Number of fluoroquinolone resistant isolates in isolates with different phenotypic backgrounds: , INH S + RIF S = isolates with resistance to one or more anti-tubercular drugs but susceptible to both isoniazid and rifampicin, INH R + RIF S = isolates that are resistant to isoniazid but susceptible to rifampicin, INH S + RIF R = isolates that are resistant to rifampicin but susceptible to isoniazid, MDR = isolates that are resistant to both isoniazid and rifampicin, UNKNOWN = isolates where phenotypes for isoniazid or rifampicin resistance could not be determined.

Looking at the prevalence of fluoroquinolone resistance in different phenotypic backgrounds within the compendium isolates, over 40% of MDR isolates within the compendium had fluoroquinolone resistance, and significantly more were resistant to levofloxacin than moxifloxacin (Figure 29). This was significantly higher than the proportion of isoniazid and rifampicin susceptible isolates, isoniazid resistant and rifampicin susceptible isolates and isoniazid susceptible and rifampicin resistant isolates that were fluoroquinolone resistant. There was a significantly higher proportion of fluoroquinolone resistance in isolates that were resistant to either rifampicin or isoniazid than isolates that were susceptible to both drugs.

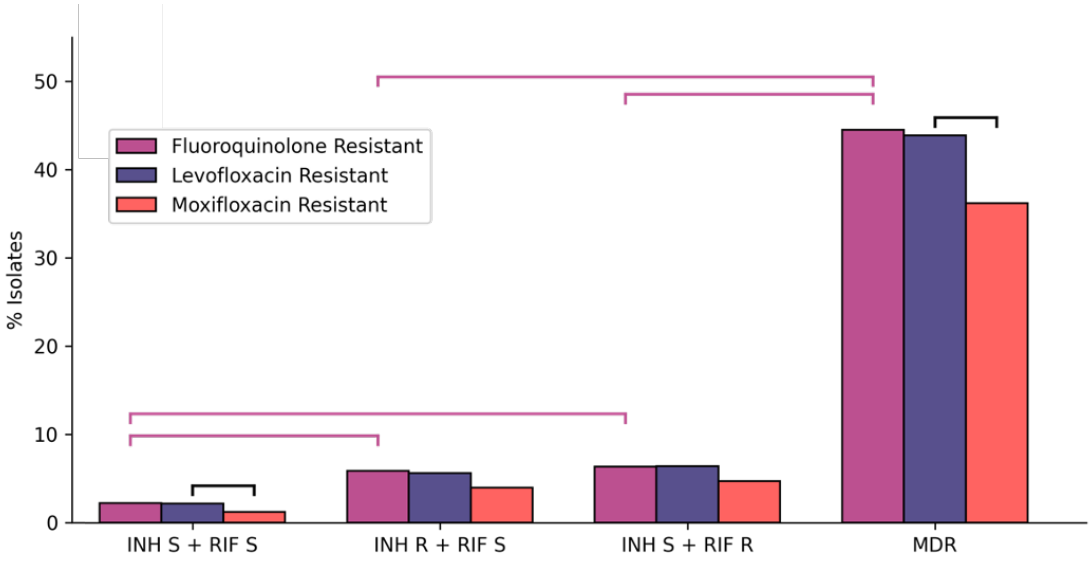

Figure 29 Proportion of isolates with fluoroquinolone resistance in different phenotypic backgrounds. ANY R = isolates with resistance to one or more anti-tubercular drugs, INH S + RIF S = isolates susceptible to both isoniazid and rifampicin, INH R + RIF S = isolates that are resistant to isoniazid but susceptible to rifampicin, INH S + RIF R = isolates that are resistant to rifampicin but susceptible to isoniazid, MDR = isolates that are resistant to both isoniazid and rifampicin. Pink brackets indicate a significant difference ($p < 0.05$) in the proportion of isolates that were resistant to a fluoroquinolone in different phenotypic backgrounds as calculated by a Z test (isolates in a background of any resistance were excluded as these could not be reasonably compared with the other backgrounds). Black brackets indicate a significant difference ($p<0.05$) in the proportion of isolates resistant to levofloxacin and moxifloxacin in a particular phenotypic background as calculated by Z test.

4.3.2 DNA gyrase mutations

Levofloxacin and moxifloxacin are both structurally similar members of the fluoroquinolone class, but they may have different genetic drivers of resistance. To see if there was a difference in the non-synonymous mutations seen in levofloxacin and moxifloxacin resistant isolates, I compared the prevalence of each *gyrA* or *gyrB* mutation seen in the two populations. Because genetic background may influence the mutations seen, I controlled for country of origin, lineage and rifampicin and isoniazid resistance background. Using the largest group of these three variables, Lineage 2 isolates of Indian origin that were MDR, there was no significant difference ($p < 0.05$) between the mutations seen and their prevalence in levofloxacin and moxifloxacin resistant isolates (Table 11). Therefore, for subsequent genetic analysis of resistant populations, moxifloxacin and levofloxacin resistant isolates were grouped.

Mutation	LEV_N	LEV_TOTAL	LEV_%	MXF_N	MXF_TOTAL	MXF_%	P
gyrA D94G	184	388	47.4	181	362	50.0	0.48
gyrA A90V	92	388	23.7	83	362	22.9	0.80
gyrA D94N	30	388	7.7	31	362	8.6	0.68
gyrA D94A	29	388	7.5	27	362	7.5	0.99
gyrA S91P	16	388	4.1	14	362	3.9	0.86
gyrA D94Y	14	388	3.6	13	362	3.6	0.99
gyrB N499T	5	388	1.3	4	362	1.1	0.82
gyrA D94H	4	388	1.0	3	362	0.8	0.77
gyrB E501D	3	388	0.8	5	362	1.4	0.42
gyrA G88C	2	388	0.5	2	362	0.6	0.94
gyrA I462V	2	388	0.5	1	362	0.3	0.60
gyrA H70R	2	388	0.5	1	362	0.3	0.60
gyrB T500N	2	388	0.5	2	362	0.6	0.94
gyrB I486L	1	388	0.3	1	362	0.3	0.96
gyrB E501V	1	388	0.3	0	362	0.0	0.33
gyrB V670F	1	388	0.3	1	362	0.3	0.96
gyrB D461N	1	388	0.3	0	362	0.0	0.33
gyrA R292G	1	388	0.3	1	362	0.3	0.96
gyrB R446H	1	388	0.3	0	362	0.0	0.33
gyrA D89G	1	388	0.3	1	362	0.3	0.96
gyrA R450S	1	388	0.3	1	362	0.3	0.96
gyrB A233P	1	388	0.3	1	362	0.3	0.96
gyrB T500A	1	388	0.3	1	362	0.3	0.96
gyrB A504T	1	388	0.3	1	362	0.3	0.96
gyrB N499S	1	388	0.3	1	362	0.3	0.96
gyrB S447F	1	388	0.3	1	362	0.3	0.96

Table 11 Prevalence of mutations seen in levofloxacin and moxifloxacin resistant Lineage 2 MDR *M. tuberculosis* isolates from India. P values indicate the significance of the difference between the percentage of levofloxacin and moxifloxacin resistant isolates with the mutation, as calculated using a two proportions Z test.

The majority (78.6%) of the 2191 fluoroquinolone resistant isolates had at least one non-synonymous mutation in the *gyrA* QRDR (Figure 30). A small proportion (3.1%) contained no mutation in the *gyrA* QRDR but did have a non-synonymous mutation in the *gyrB* QRDR and 2.2% contained a non-synonymous mutation elsewhere in either the *gyrA* or *gyrB* (above a background of known lineage specific mutations). A significant proportion of the fluoroquinolone resistant isolates (16.0%) contained no non-synonymous mutations in either *gyrA* or *gyrB* according to the CRyPTIC dataset.

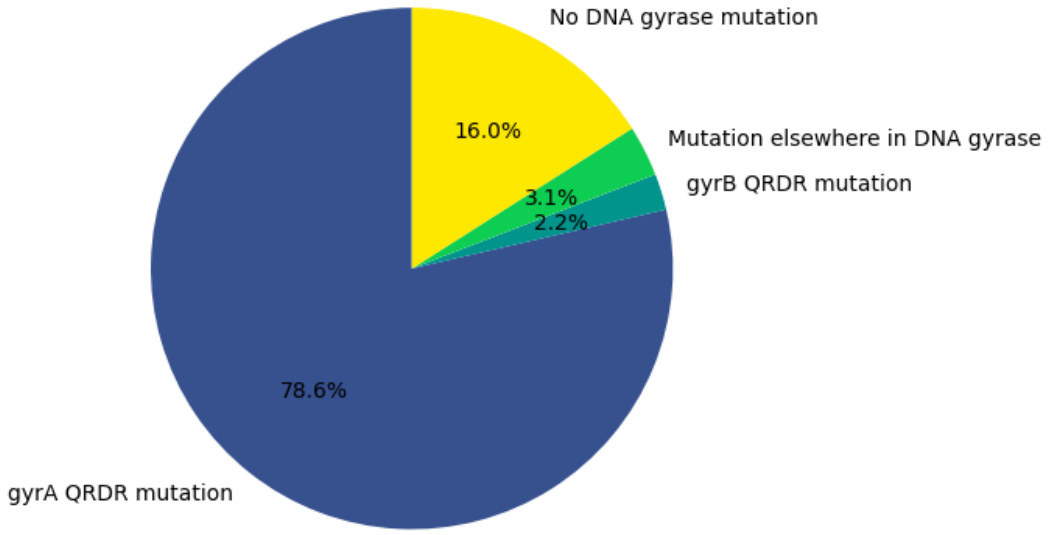

Figure 30 Proportion of 2191 fluoroquinolone resistant isolates with DNA gyrase mutations. '*gyrA* QRDR mutation' represents the proportion of isolates with a non-synonymous mutation in the *gyrA* QRDR. '*gyrB* QRDR mutation' represents the proportion of isolates with a *gyrB* QRDR mutation and no *gyrA* QRDR mutations. 'Mutation elsewhere in DNA gyrase' shows the proportion of isolates with no *gyrA* or *gyrB* QRDR mutations but at least one non-synonymous mutation elsewhere in either *gyrA* or *gyrB.* 'No DNA gyrase mutation' represents the proportion of fluoroquinolone resistant isolates that do not have a non-synonymous mutation in either *gyrA* or *gyrB.*

I then investigated the prevalence of *gyrA* or *gyrB* QRDR mutations in phenotypically fluoroquinolone resistant and susceptible isolates. All of the *gyrA* QRDR mutations seen in resistant isolates were also seen in fluoroquinolone susceptible isolates, bar *gyrA* D89G, A90E and D94F which were only seen in resistant isolates and at low prevalence (< 0.1%) (Figure 31). The *gyrA* QRDR mutations were enriched in fluoroquinolone resistant isolates compared to susceptible isolates, for example the resistance associated mutation *gyrA* D94G was seen in 34.8% of fluoroquinolone resistant isolates and 1.6% of fluoroquinolone susceptible isolates, and *gyrA* A90V was seen in 21.3% of resistant isolates and 2.1% of phenotypically susceptible isolates. The most variation was seen at the *gyrA* D94 codon, where 7 different non-synonymous mutations were seen. Some synonymous mutations were seen at positions known to have resistance conferring variants, for example A90A and

G88G were seen at low prevalence in fluoroquinolone susceptible isolates. Mutations outside of *gyrA* positions 88 - 94 were rare in both resistant and susceptible isolates - there were 17 different mutations (excluding known lineage specific mutations), 6 of which were synonymous, and all were seen in less than 0.3% of isolates.

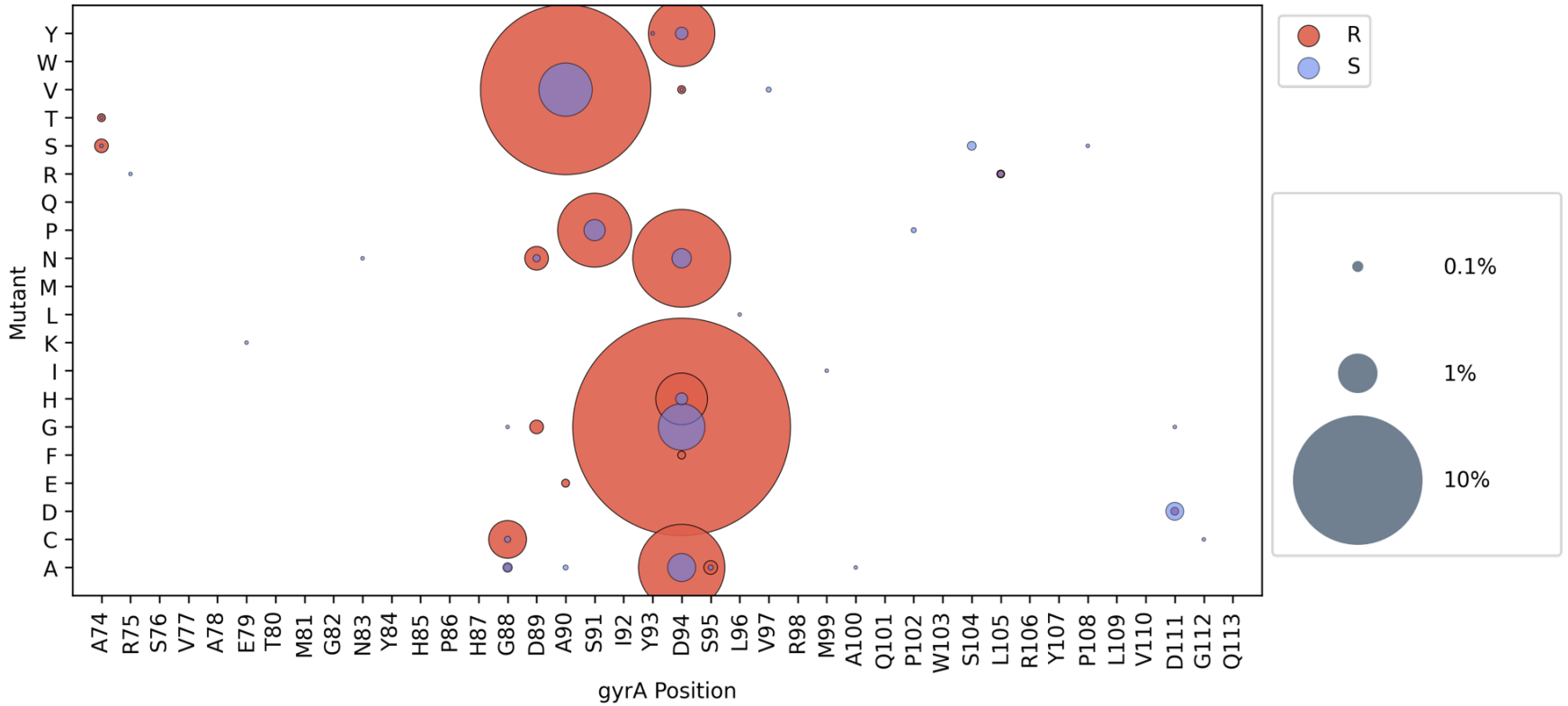

Figure 31 Mutations found in the QRDR of *gyrA* in fluoroquinolone resistant isolates and fluoroquinolone susceptible isolates. Presence of a coloured spot indicates that the mutation was found in an isolate and spot size corresponds to the proportion of fluoroquinolone resistant or susceptible isolates carrying that mutation.

For the *gyrB* QRDR, mutations were seen at lower prevalence than the *gyrA* QRDR and none of the mutations were present in more that 1% of fluoroquinolone resistant or susceptible isolates (Figure 32). As with the *gyrA* QRDR mutations, the majority of mutations seen in phenotypically resistant isolates occurred at low prevalence (<0.1%) in fluoroquinolone susceptible isolates. The most genetic variation was seen at codon N499, where five non-synonymous mutations and one synonymous mutation were observed. Twenty mutations were seen outside of codons known to have resistance conferring alleles (461, 499, 501 and 504), but these were rare – present in less than 0.1% of phenotypically resistant or susceptible isolates. Some mutations at *gyrB* QRDR positions known to have resistance conferring mutations were present in susceptible isolates only, for example *gyrB* D461A, D461V and the synonymous mutation N499N, but these were seen in less than 0.1% of the fluoroquinolone susceptible population.

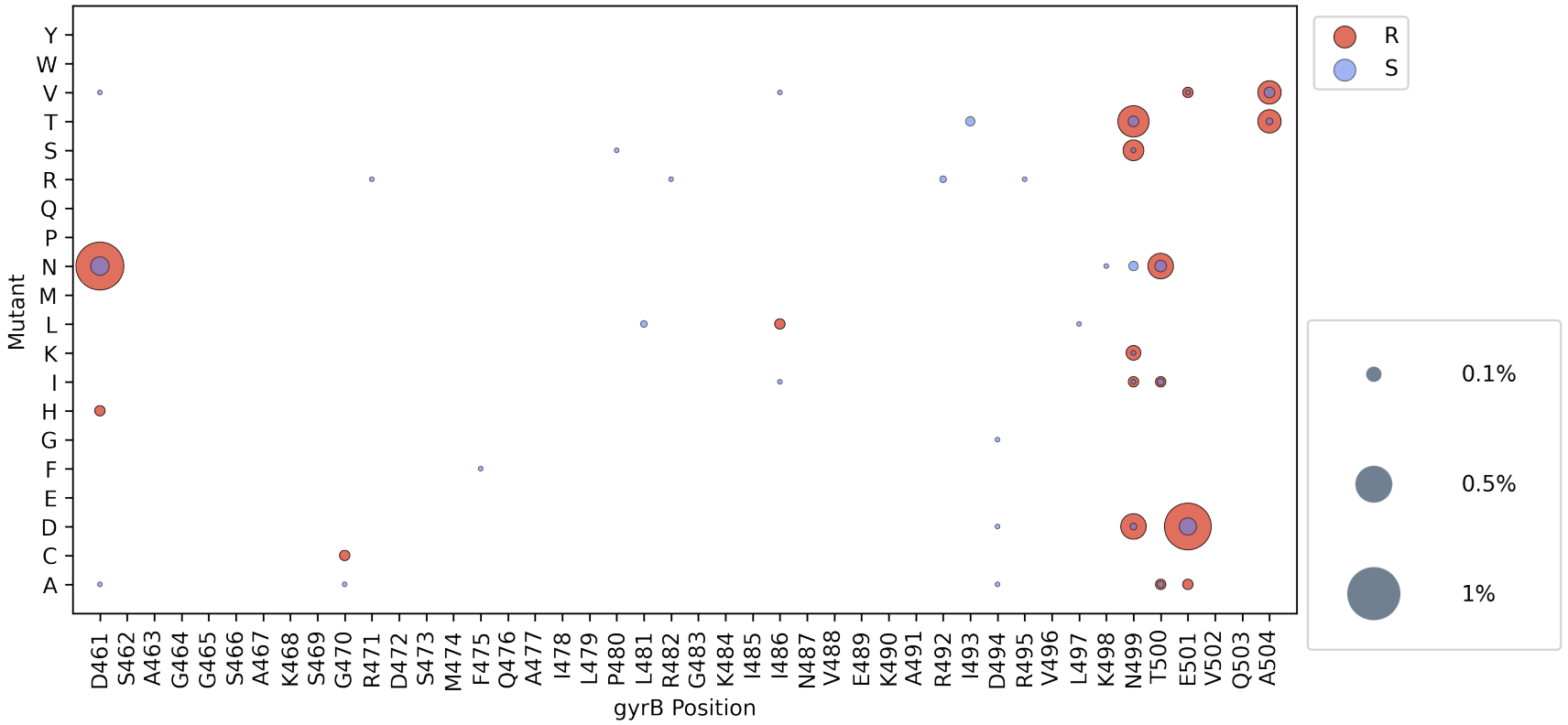

Figure 32 Mutations found in the QRDR of *gyrB* in fluoroquinolone resistant isolates and fluoroquinolone susceptible isolates. Presence of a coloured spot indicates that the mutation was found in an isolate and spot size corresponds to the proportion of fluoroquinolone resistant or susceptible isolates carrying that mutation.

4.3.3 Catalogue resistance associated mutations in fluoroquinolone resistant isolates

As there was no difference between the prevalence of different mutations in levofloxacin and moxifloxacin resistant isolates (Table 11), these were combined for the following genetic analyses. All 14 of the resistance associated mutations in the WHO 2021 catalogue are represented within the fluoroquinolone resistant isolates of the compendium, with the most prevalent resistance associated mutations being *gyrA* D94G and *gyrA* A90V (Figure 33). Overall, *gyrB* resistance associated mutations were less prevalent than *gyrA* mutations and mutations at *gyrA* G88 were less common that mutations at other *gyrA* positions.

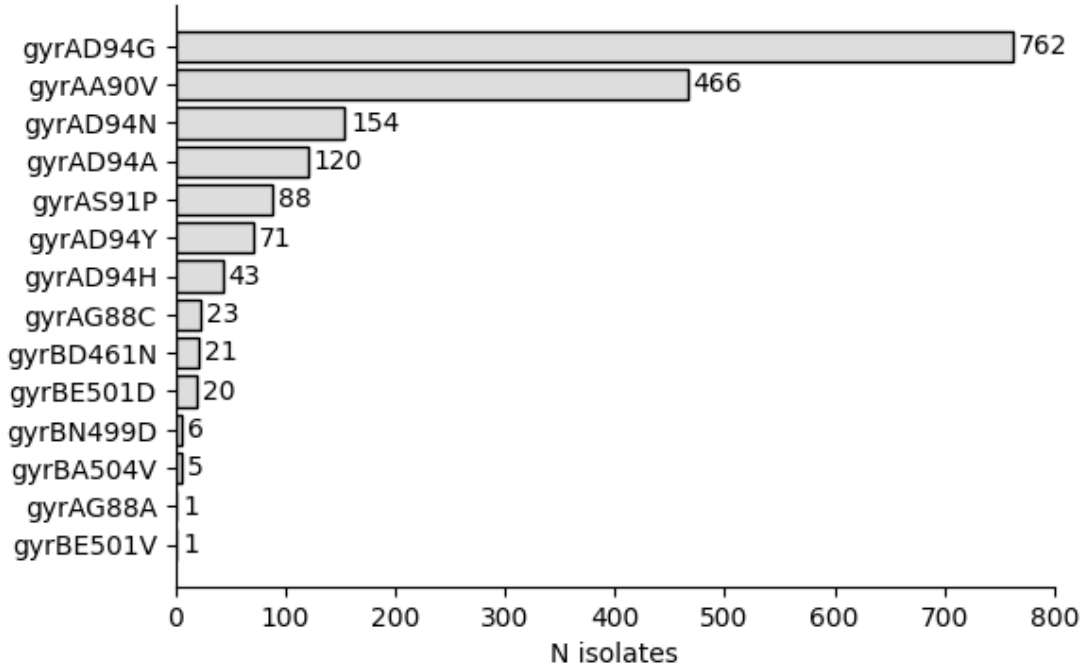

Figure 33 Number of isolates with WHO catalogue fluoroquinolone resistance associated mutations seen in the 2191 fluoroquinolone resistant isolates in the CRyPTIC dataset. Note that the numbers are not additive since an isolate may have more than one of the mutations.

Just 36 of the 2191 fluoroquinolone resistant isolates (1.6%) contained more than one resistance associated mutation and no isolates were found to contain more than two resistance associated mutations. The combinations of resistance associated mutations seen and the number of times they occurred within fluoroquinolone resistant isolates are shown in Table 12. The *gyrA* A90V mutation was seen in 29 of the 36 isolates with two resistance associated mutations, and was seen most frequently in combination with *gyrA* D94A as opposed to the more prevalent mutation *gyrA* D94G (Figure 33), which was only seen in 10 isolates with two resistance conferring mutations (Table 12).

COMBINATION OF RESISTANCE ASSOCIATED MUTATIONS	N_ISOLATES
gyrAA90V + gyrAD94A	10
gyrAA90V + gyrAD94G	8
gyrAA90V + gyrAS91P	4
gyrAA90V + gyrBD461N	4
gyrAD94A + gyrBE501D	3
gyrAA90V + gyrBA504V	2
gyrAD94G + gyrBA504V	1
gyrAA90V + gyrAD94N	1
gyrAD94G + gyrBE501D	1
gyrAD94N + gyrBE501D	1
gyrAD94A + gyrBA504V	1

Table 12 Combinations of resistance associated mutations and the number of fluoroquinolone resistant isolates they are seen in within the CRyPTIC compendium

To investigate whether the double mutations increased the magnitude of resistance compared to the individual mutations on their own, I compared the mean levofloxacin and moxifloxacin $\log_2$MIC of isolates with the double mutations to that of isolates with the individual mutations only (Figure 34a-d). For both levofloxacin and moxifloxacin, the $\log_2$MIC was significantly higher ($p < 0.01$) for isolates with the *gyrA* A90V + *gyrA* D94A double

mutation than for either of the mutations seen individually (Figure 34a-b). For the *gyrA* A90V + *gyrA* D94G double mutation, isolates had a significantly higher log_2MIC to both fluoroquinolones than for the *gyrA* A90V mutation alone (Figure 34c). The moxifloxacin log_2MIC of the *gyrA* D94G mutation was not significantly lower than for the *gyrA* A90V + *gyrA* D94G double mutation, but the levofloxacin log_2MIC was (Figure 34d). It is difficult to estimate the true increase in MIC that occurs with these double mutations, and whether the increase is additive, as the log_2MICs are at the limit of detection due to censoring of the concentrations in the plate design (Figure 19) and assumptions are made at the extremes of detection; the log_2MIC is taken from what would be the next doubling dilution, i.e. a measured MIC of > 4 mg/L is assumed to be 8 mg/L but could in fact be higher. Further, the genetic background of the isolates may have a confounding effect.

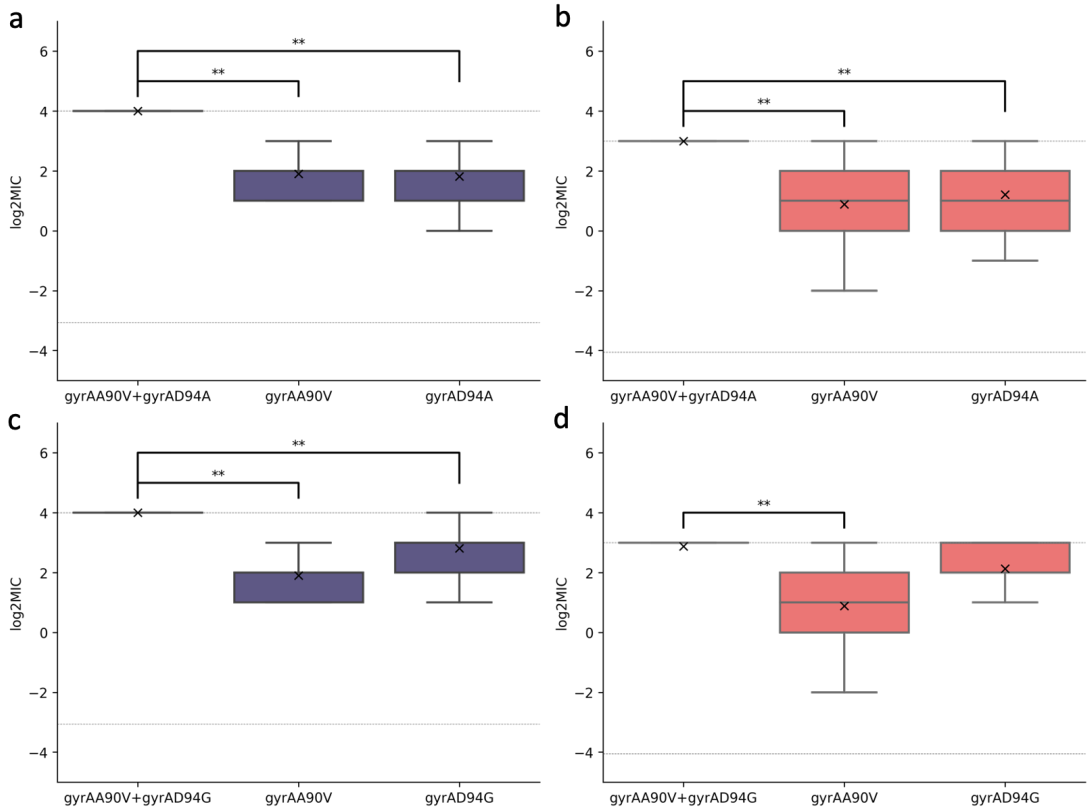

Figure 34 Box and whisker plots showing distribution of log_2MICs for isolates with combinations of resistance associated mutations compared to the mutations as individual. a) Difference in levofloxacin log_2MIC between *gyrA* A90V + *gyrA* D94A mutants and isolates with the individual mutations. b) Difference in moxifloxacin log_2MIC between *gyrA* A90V + *gyrA* D94A mutants and isolates with the individual mutations. c) Difference in levofloxacin log_2MIC between *gyrA* A90V + *gyrA* D94G mutants and isolates with the individual mutations. d) Difference in moxifloxacin log_2MIC between *gyrA* A90V + *gyrA* D94G mutants and isolates with the individual mutations. Mean log_2MICs are shown by a cross, the area between the two grey dashed lines show the limit of MIC detection using the CRyPTIC plates. Brackets with ** indicate a significant difference in means at $p < 0.01$ as determined by the Wilcoxon rank-sum test. Only high or medium quality MICs were included in this analysis (Table 6).

4.3.4 Associations between resistance conferring mutations and genetic background and sample origin

In order to test whether specific resistance conferring mutations are associated with the genetic background of an isolate, I constructed logistic regression models to describe the likelihood of finding *gyrA* D94G and A90V mutations based on the following variables: lineage, country of origin and rifampicin and isoniazid phenotypic resistance background –

all of which are confounded. Using only the most common resistance conferring mutations ensures that there will be sufficient samples for the models to be reliable. In order to further increase the number of samples used by the models, both levofloxacin and moxifloxacin isolates were included, as I previously found no difference in the mutations seen in levofloxacin and moxifloxacin resistant isolates (Table 11). To prevent small group sizes, samples from countries with fewer than 20 fluoroquinolone resistant isolates were not included and isolates with unknown rifampicin or isoniazid resistance backgrounds, rifampicin mono-resistant isolates and MDR isolates were grouped together as a single 'rifampicin resistant/MDR' category – as the unknown isolates were most likely to be MDR.

Compared to fluoroquinolone resistant isolates with a Lineage 2 background, Lineage 1 and Lineage 4 isolates were 76.9% and 43.0% less likely to contain a *gyrA* D94G mutation after controlling for other variables (country and isoniazid and rifampicin resistance background) (Table 13). Isolates that were susceptible to isoniazid and rifampicin had significantly lower odds of having a *gyrA* D94G mutation compared to rifampicin resistant isolates (either rifampicin mono-resistant or multidrug resistant). Compared to India, fluoroquinolone resistant isolates from China, Peru and Vietnam were 50.9%, 37.6% and 88.1% less likely to contain a *gyrA* D94G mutation after controlling for lineage and isoniazid and rifampicin resistance background. After applying a Bonferroni correction to correct for multiple testing, the effects of Lineage 1, Lineage 4 and China being the country of origin remained statistically significant at $p < 0.05$.

	Coef.	Adj. OR	Adj. [0.025	Adj. 0.975]	p>|z|	Corr. p
LINEAGE_Lineage 1	**-1.465**	**0.231**	**0.102**	**0.526**	**0.000**	**0.008**
LINEAGE_Lineage 2	Ref	Ref	Ref	Ref	Ref	Ref
LINEAGE_Lineage 3	-0.269	0.764	0.509	1.147	0.194	1.000
LINEAGE_Lineage 4	**-0.562**	**0.570**	**0.426**	**0.763**	**0.000**	**0.003**
COUNTRY_BRA	-1.393	0.248	0.056	1.106	0.068	1.000
COUNTRY_CHN	**-0.710**	**0.491**	**0.328**	**0.736**	**0.001**	**0.010**
COUNTRY_DEU	0.570	1.768	0.813	3.846	0.151	1.000
COUNTRY_IND	Ref	Ref	Ref	Ref	Ref	Ref
COUNTRY_ITA	-0.335	0.715	0.410	1.247	0.238	1.000
COUNTRY_KGZ	-0.766	0.465	0.121	1.778	0.263	1.000
COUNTRY_NPL	0.336	1.400	0.953	2.055	0.086	1.000
COUNTRY_PAK	-0.194	0.824	0.543	1.250	0.362	1.000
COUNTRY_PER	**-0.475**	**0.622**	**0.398**	**0.972**	**0.037**	**0.628**
COUNTRY_TKM	0.078	1.081	0.438	2.668	0.866	1.000
COUNTRY_VNM	**-2.129**	**0.119**	**0.027**	**0.515**	**0.004**	**0.075**
COUNTRY_ZAF	-0.196	0.822	0.611	1.107	0.197	1.000
BACKGROUND_INH_AND_RIF_S	**-0.645**	**0.525**	**0.304**	**0.905**	**0.020**	**0.346**
BACKGROUND_INH_MONOR	-0.496	0.609	0.296	1.255	0.179	1.000
BACKGROUND_RIF_R/MDR	Ref	Ref	Ref	Ref	Ref	Ref
Intercept	-0.170	0.844	0.703	1.013	0.069	1.000

Table 13 Association between lineage, country of origin or background and the odds of a *gyrA* D94G mutation being present in a fluoroquinolone resistant isolate. Rows in bold show the variables that are statistically significant at $p < 0.5$. Corr. p are the Bonferroni-corrected p values.

Compared against other lineage backgrounds, fluoroquinolone resistant isolates in Lineage 4 were 58.3% more likely to contain a *gyrA* A90V mutation compared to Lineage 2 isolates after controlling for country of origin and isoniazid/rifampicin resistance background (Table 14). Isolates from several countries (Brazil, Peru, Vietnam and South Africa) had lower odds of containing a *gyrA* A90V mutation compared to isolates collected in India. However, after applying the Bonferroni correction to correct for multiple testing, none of the associations were statistically significant at $p < 0.05$.

	Coef.	Adj. OR	Adj. [0.025	Adj. 0.975]	p>|z|	Corr. p
LINEAGE_Lineage 1	0.339	1.403	0.729	2.702	0.311	1.000
LINEAGE_Lineage 2	Ref	Ref	Ref	Ref	Ref	Ref
LINEAGE_Lineage 3	-0.128	0.880	0.553	1.400	0.589	1.000
LINEAGE_Lineage 4	**0.459**	**1.583**	**1.156**	**2.169**	**0.004**	**0.072**
COUNTRY_BRA	**-2.205**	**0.110**	**0.014**	**0.845**	**0.034**	**0.575**
COUNTRY_CHN	-0.444	0.642	0.404	1.019	0.060	1.000
COUNTRY_DEU	-0.199	0.820	0.324	2.073	0.675	1.000
COUNTRY_IND	Ref	Ref	Ref	Ref	Ref	Ref
COUNTRY_ITA	-0.548	0.578	0.297	1.126	0.107	1.000
COUNTRY_KGZ	-0.375	0.687	0.146	3.230	0.635	1.000
COUNTRY_NPL	0.262	1.300	0.850	1.986	0.226	1.000
COUNTRY_PAK	-0.047	0.954	0.607	1.499	0.837	1.000
COUNTRY_PER	**-0.651**	**0.521**	**0.321**	**0.847**	**0.009**	**0.145**
COUNTRY_TKM	-1.877	0.153	0.020	1.158	0.069	1.000
COUNTRY_VNM	**-1.626**	**0.197**	**0.045**	**0.860**	**0.031**	**0.523**
COUNTRY_ZAF	**-0.397**	**0.672**	**0.473**	**0.956**	**0.027**	**0.462**
BACKGROUND_INH_AND_RIF_S	-0.485	0.616	0.342	1.108	0.106	1.000
BACKGROUND_INH_MONOR	0.297	1.346	0.690	2.625	0.383	1.000
BACKGROUND_RIF_R/MDR	Ref	Ref	Ref	Ref	Ref	Ref
Intercept	-1.176	0.309	0.250	0.381	0.000	0.000

Table 14 Association between lineage, country of origin or background and the odds of a *gyrA* A90V mutation being present in a fluoroquinolone resistant isolate. Rows in bold show the variables that are statistically significant at $p < 0.5$. Corr. p are the Bonferroni-corrected p values.

4.3.5 Association between genetic background, country of origin and phenotypic background and the level of resistance conferred by resistance mutations

I next investigated whether the most common resistance conferring mutations, *gyrA* D94G and A90V, were associated with different magnitudes of resistance depending on the genetic background of the isolate. For this analysis I only included isolates from countries with over 50 fluoroquinolone resistant isolates. I also controlled for the effect of multiple resistance conferring mutations as these would be expected to increase the MIC more than a single non-synonymous mutation (Figure 34a-c). Indeed, the presence of additional resistance conferring mutations was positively associated with levofloxacin MIC for isolates with a *gyrA* D94G mutation (Table 15). I did not consider the effects of other DNA gyrase

mutations; there may be other resistance mutations in the gyrase genes that are not catalogued but it is expected these will be rare in the dataset. There can be effects on MIC associated with presence of non-resistance conferring mutations, for example the *gyrA* A90G + T80A neutralising effect discussed in section 4.1, but this mutation was only present in one isolate in the compendium, is lineage associated, and no other compensatory mutations have yet been found.

Compared to rifampicin resistant isolates (rifampicin mono-resistant or MDR) that had a *gyrA* D94G mutation, isolates with *gyrA* D94G that were susceptible to rifampicin (both isoniazid and rifampicin susceptible isolates or isoniazid mono-resistant isolates) were negatively associated with levofloxacin MIC (Table 15). Levofloxacin MIC was positively associated with isolates originating from China compared to India and there were no significant associations found between lineage and levofloxacin MIC at $p < 0.05$. After Bonferroni correction, the negative association between rifampicin susceptible backgrounds and the levofloxacin MIC of isolates with *gyrA* D94G remained significant (Table 15).

	Coef.	[0.025	0.975]	P>|t|	Corr. p
LINEAGE_Lineage 1	1.019	-0.144	2.182	0.086	1.000
LINEAGE_Lineage 2	Ref	Ref	Ref	Ref	Ref
LINEAGE_Lineage 3	0.486	-0.010	0.981	0.055	0.713
LINEAGE_Lineage 4	0.164	-0.239	0.567	0.424	1.000
COUNTRY_CHN	**0.627**	**0.103**	**1.150**	**0.019**	**0.248**
COUNTRY_IND	Ref	Ref	Ref	Ref	Ref
COUNTRY_ITA	0.024	-0.647	0.694	0.944	1.000
COUNTRY_NPL	-0.336	-0.748	0.077	0.110	1.000
COUNTRY_PAK	-0.333	-0.879	0.213	0.232	1.000
COUNTRY_PER	-0.061	-0.633	0.511	0.834	1.000
COUNTRY_ZAF	0.341	-0.026	0.707	0.069	0.893
MULT_MUT	**1.420**	**0.473**	**2.367**	**0.003**	**0.044**
BACKGROUND_INH_AND_RIF_S	**-1.229**	**-1.923**	**-0.535**	**0.001**	**0.007**
BACKGROUND_INH_MONOR	**-1.770**	**-2.579**	**-0.960**	**0.000**	**0.000**
BACKGROUND_RIF_R/MDR	Ref	Ref	Ref	Ref	Ref
Intercept	2.515	2.303	2.728	0.000	0.000

Table 15 Association between lineage, country of origin or phenotypic background and the levofloxacin log_2MIC of isolates with a *gyrA* D94G mutation, controlling for the presence of additional resistance conferring mutations. Rows in bold show the variables that are statistically significant at $p < 0.5$. Corr. p are the Bonferroni-corrected p values.

The presence of multiple mutations was not significantly associated with an increased moxifloxacin MIC in isolates with a *gyrA* D94G mutation (Table 16), which was expected as I previously observed no difference in MIC between *gyrA* D94G and *gyrA* A90V double mutations and the *gyrA* D94G mutation on its own (Figure 34d) - though this may arise from these MICs being at the limit of detection. As seen for levofloxacin MICs, compared to rifampicin resistant isolates (rifampicin mono-resistant or MDR), isolates that were susceptible to rifampicin (both isoniazid and rifampicin susceptible isolates or isoniazid mono-resistant isolates) that had a *gyrA* D94G mutation were negatively associated with moxifloxacin MIC (Table 16). Moxifloxacin MIC for isolates containing a *gyrA* D94G mutation was also significantly positively associated with isolates originating from China and South Africa compared to India. There were differences in associations for levofloxacin MIC and moxifloxacin MIC, in that moxifloxacin MIC was negatively associated with Lineage 3

compared to Lineage 2 (Table 16) but levofloxacin MIC was not significantly associated with any lineage in comparison to Lineage 2 (Table 15). After Bonferroni correction, the effects of Lineage 3, country of origin China, and rifampicin susceptible backgrounds on moxifloxacin MIC of isolates with the *gyrA* D94G mutation remained significant.

	Coef.	[0.025	0.975]	P>|t|	Corr. p
LINEAGE_Lineage 1	0.515	-0.581	1.611	0.356	1.000
LINEAGE_Lineage 2	Ref	Ref	Ref	Ref	Ref
LINEAGE_Lineage 3	**-0.949**	**-1.459**	**-0.438**	**0.000**	**0.004**
LINEAGE_Lineage 4	-0.110	-0.505	0.284	0.583	1.000
COUNTRY_CHN	**0.819**	**0.327**	**1.311**	**0.001**	**0.015**
COUNTRY_IND	Ref	Ref	Ref	Ref	Ref
COUNTRY_ITA	-0.340	-1.004	0.324	0.315	1.000
COUNTRY_NPL	-0.261	-0.686	0.165	0.229	1.000
COUNTRY_PAK	-0.476	-1.021	0.068	0.086	1.000
COUNTRY_PER	0.177	-0.391	0.746	0.540	1.000
COUNTRY_ZAF	0.415	0.070	0.760	0.019	0.241
MULT_MUT	0.806	-0.086	1.698	0.076	0.994
BACKGROUND_INH_AND_RIF_S	**-1.715**	**-2.406**	**-1.023**	**0.000**	**0.000**
BACKGROUND_INH_MONOR	**-1.945**	**-2.850**	**-1.041**	**0.000**	**0.000**
BACKGROUND_RIF_R/MDR	Ref	Ref	Ref	Ref	Ref
Intercept	2.063	1.867	2.259	0.000	0.000

Table 16 Association between lineage, country of origin or phenotypic background and the moxifloxacin log_2MIC of isolates with a *gyrA* D94G mutation, controlling for the presence of additional resistance conferring mutations. Rows in bold show the variables that are statistically significant at $p < 0.5$. Corr. p are the Bonferroni-corrected p values.

For isolates with a *gyrA* A90V mutation, the levofloxacin MIC was negatively associated with isoniazid and rifampicin susceptible isolates compared to rifampicin resistant/MDR isolates (Table 17), but unlike for isolates with *gyrA* D94G mutations there was no association with isoniazid mono-resistant phenotypic background (Table 15, Table 17). There was no significant difference in the effect of lineage, but for country of origin, Nepal and Pakistan were negatively associated with levofloxacin MIC compared to isolates with *gyrA* A90V

originating from India. After Bonferroni correction, the negative association between isoniazid and rifampicin susceptible isolates and levofloxacin MIC remained significant at $p < 0.05$ and the association of increased MIC where there were additional resistance conferring mutations was significant as expected (Table 17).

	Coef.	[0.025	0.975]	P>|t|	Corr. p
LINEAGE_Lineage 1	0.638	-0.253	1.529	0.160	1.000
LINEAGE_Lineage 2	Ref	Ref	Ref	Ref	Ref
LINEAGE_Lineage 3	0.146	-0.411	0.703	0.606	1.000
LINEAGE_Lineage 4	0.253	-0.163	0.669	0.232	1.000
COUNTRY_CHN	-0.274	-0.834	0.285	0.335	1.000
COUNTRY_IND	Ref	Ref	Ref	Ref	Ref
COUNTRY_ITA	-0.527	-1.207	0.153	0.128	1.000
COUNTRY_NPL	**-0.578**	**-1.103**	**-0.053**	**0.031**	**0.405**
COUNTRY_PAK	**-0.619**	**-1.151**	**-0.087**	**0.023**	**0.295**
COUNTRY_PER	0.373	-0.249	0.994	0.239	1.000
COUNTRY_ZAF	0.305	-0.177	0.787	0.214	1.000
MULT_MUT	**1.951**	**1.423**	**2.479**	**0.000**	**0.000**
BACKGROUND_INH_AND_RIF_S	**-1.501**	**-2.159**	**-0.843**	**0.000**	**0.000**
BACKGROUND_INH_MONOR	-0.468	-1.307	0.371	0.273	1.000
BACKGROUND_RIF_R/MDR	Ref	Ref	Ref	Ref	Ref
Intercept	1.715	1.409	2.022	0.000	0.000

Table 17 Association between lineage, country of origin or phenotypic background and the levofloxacin log_2MIC of isolates with a *gyrA* A90V mutation, controlling for the presence of additional resistance conferring mutations. Rows in bold show the variables that are statistically significant at $p < 0.5$. Corr. p are the Bonferroni-corrected p values.

Similarly to levofloxacin, the moxifloxacin MIC of isolates with *gyrA* A90V was negatively associated with isoniazid and rifampicin susceptible isolates compared to rifampicin resistant/MDR isolates, but not isoniazid mono-resistant isolates (Table 18). There was however a significant effect of lineage on moxifloxacin MIC; Lineage 3 isolates with *gyrA* A90V were more negatively associated with moxifloxacin MIC compared to Lineage 2. For country of origin, Italy and Pakistan were negatively associated with moxifloxacin MIC

compared to isolates with *gyrA* A90V originating from India (Table 18). After Bonferroni correction, the negative associations between isoniazid and rifampicin susceptibility, Lineage 3, and levofloxacin MIC remained significant at $p < 0.05$, and the association of increased MIC where there were additional resistance conferring mutations was significant (Table 18).

	Coef.	[0.025	0.975]	P>|t|	Corr. p
LINEAGE_Lineage 1	-0.384	-1.307	0.539	0.414	1.000
LINEAGE_Lineage 2	Ref	Ref	Ref	Ref	Ref
LINEAGE_Lineage 3	**-1.107**	**-1.736**	**-0.477**	**0.001**	**0.008**
LINEAGE_Lineage 4	-0.157	-0.625	0.310	0.508	1.000
COUNTRY_CHN	0.249	-0.407	0.905	0.456	1.000
COUNTRY_IND	Ref	Ref	Ref	Ref	Ref
COUNTRY_ITA	**-0.848**	**-1.617**	**-0.079**	**0.031**	**0.400**
COUNTRY_NPL	-0.491	-1.107	0.126	0.118	1.000
COUNTRY_PAK	**-0.720**	**-1.324**	**-0.117**	**0.019**	**0.253**
COUNTRY_PER	0.123	-0.579	0.826	0.730	1.000
COUNTRY_ZAF	0.154	-0.362	0.671	0.557	1.000
MULT_MUT	**1.669**	**1.070**	**2.268**	**0.000**	**0.000**
BACKGROUND_INH_AND_RIF_S	**-1.504**	**-2.360**	**-0.648**	**0.001**	**0.008**
BACKGROUND_INH_MONOR	-0.247	-1.157	0.662	0.593	1.000
BACKGROUND_RIF_R/MDR	Ref	Ref	Ref	Ref	Ref
Intercept	1.018	0.687	1.348	0.000	0.000

Table 18 Association between lineage, country of origin or phenotypic background and the moxifloxacin $\log_2$MIC of isolates with a *gyrA* A90V mutation, controlling for the presence of additional resistance conferring mutations. Rows in bold show the variables that are statistically significant at $p < 0.5$. Corr. p are the Bonferroni-corrected p values.

4.3.6 Resistance prediction using WHO catalogue mutations and *in silico* molecular diagnostic tests

To test how the catalogue and molecular diagnostic tests might perform despite these limitations, I used a subset of the CRyPTIC isolates, excluding those that had low confidence MICs to levofloxacin and moxifloxacin (Figure 35a-b).

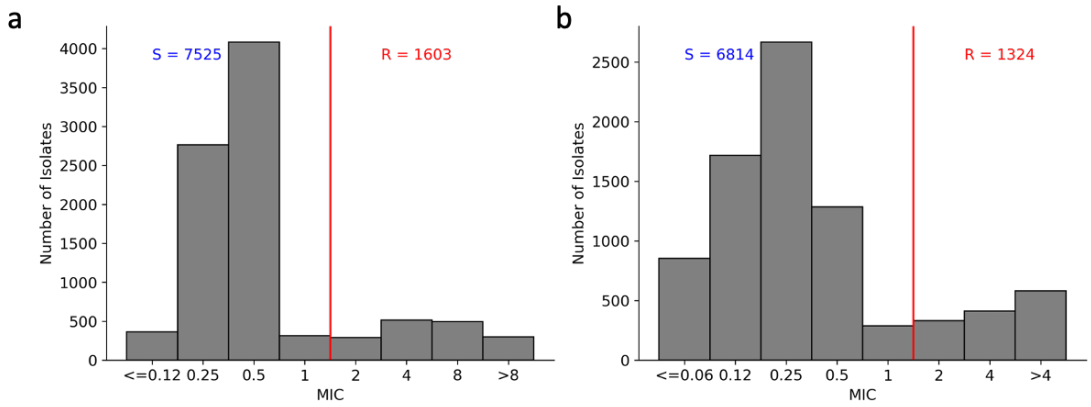

Figure 35 Distribution of *M. tuberculosis* isolates Minimum Inhibitory Concentrations to (a) levofloxacin and (b) moxifloxacin. The red line indicates the epidemiological cut off value (ECOFF) that was used to distinguish resistant and susceptible isolates.

I examined how many of these isolates contained the resistance conferring mutations listed in the catalogue or used by each molecular test in Table 10 and then calculated the sensitivity and specificity of these mutations for identifying levofloxacin and moxifloxacin resistance in the dataset (Figure 36). The catalogue mutations identified 83.1% of levofloxacin resistance and 85.4% of moxifloxacin resistance and relatively few susceptible isolates contained these resistance associated mutations; the catalogue mutations identified resistance with over 90% specificity for both levofloxacin and moxifloxacin. There was no significant difference in sensitivity between levofloxacin and moxifloxacin resistance prediction ($z = -1.717$, $p = 0.086$) but levofloxacin had significantly higher specificity ($z = 9.486$, $p < 0.0001$). In total, 271 levofloxacin resistant isolates and 193 moxifloxacin resistant isolates were not predicted by the catalogue; therefore 16.9% of levofloxacin resistance and 14.6% of moxifloxacin resistance in this dataset is not explainable by mutations in the WHO catalogue.

In general, the molecular diagnostic tests identified a smaller proportion of resistant isolates than the catalogue, bar the resistance inferred interpretation of Genotype MTBDRsl v2 which identified 83.4% of levofloxacin resistance and 86.3% of moxifloxacin resistance (Figure 36). However, after statistical analysis, the differences between Genotype MTBDRsl v2 and the catalogue were not significant at $p < 0.05$. Both the catalogue mutations and the resistance inferred interpretation of Genotype MTBDRsl v2 had significantly higher sensitivity than the resistance detected interpretations of Genoscholar FQ + KM – TB II and AID TB FQ/EMB Kit, which detect only three and six resistance associated mutations respectively. When considering a resistance detected test interpretation only, on increasing the number of mutations detected to 14, as for the Xpert MTB/XDR test, the sensitivity was significantly higher than the 3 or 6 mutations used for Genoscholar FQ + KM – TB II and AID TB FQ/EMB Kit for both levofloxacin and moxifloxacin resistance detection. For both levofloxacin and moxifloxacin resistance detection the sensitivity significantly increased when using the resistance inferred interpretation of Genoscholar FQ + KM – TB II compared to the resistance detected interpretation. There was no significant difference between sensitivities for detecting levofloxacin or moxifloxacin resistance for the two different interpretations for the other tests. In terms of specificity, there was no significant difference between any of the test mutations or the catalogue, bar the resistance detected interpretation of Genoscholar FQ + KM – TB II, which had significantly higher specificity.

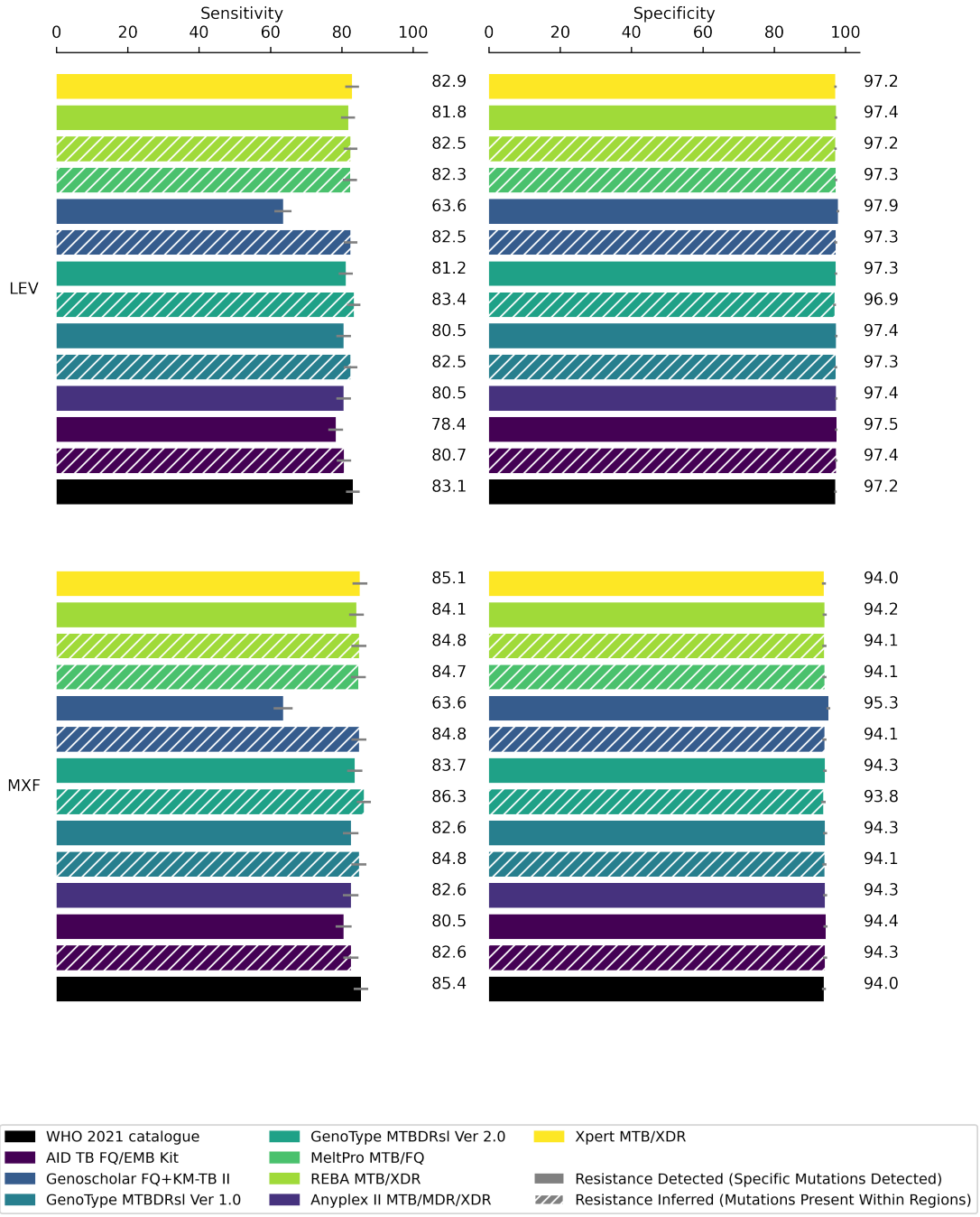

Figure 36 Sensitivity and specificity of WHO catalogue mutations and mutations detected by molecular diagnostic tests for predicting resistance to levofloxacin (LEV) and moxifloxacin (MXF). Error bars show 95% confidence intervals calculated using the method of Wilson[279].

4.3.7 Mixed alleles

I hypothesised that the resistance that could not be explained using the WHO catalogue mutations is due to the stringent filters applied during the variant calling pipeline that prevent the detection of mixed alleles in *M. tuberculosis* isolates containing a heterogeneous population (see Section 3.2.2). I defined a mixed allele, and therefore isolates taken from a heterogenous population, as those with evidence of at least two reads supported an alternative allele to the reference at one or more positions in *gyrA* or *gyrB* known to be associated with resistance (see Section 4.2.4).

A significant proportion of the isolates with an MIC measurement for a fluoroquinolone, 5.2% (641/12,354), showed evidence of mixed alleles in the DNA gyrase genes. This is likely an underestimate as there is no deep sequencing data for the CRyPTIC isolates and the sequencing depth varied between isolates and across positions, thereby leading to a dynamic and unquantifiable limit of detection. Figure 37 shows the FRS for all the alternative alleles seen and the sequencing depth at the allele position. Alternative alleles in this dataset, and at the limits of detection incurred by the sequencing data, were most frequently seen at between 0.1 and 0.2 FRS. The sequencing depths at positions with evidence of alternative alleles were most frequently between 3 and 30 reads, meaning that clinically relevant minor populations that make up a small percentage of the isolate are not identifiable. For example, with a depth of 30, the lowest possible FRS detectable for a minor allele would be 0.067 (i.e. 2 reads out of 30 supporting the alternative allele) and therefore only mixed populations with an alternative allele in >= 6.7% of the population could be identified.

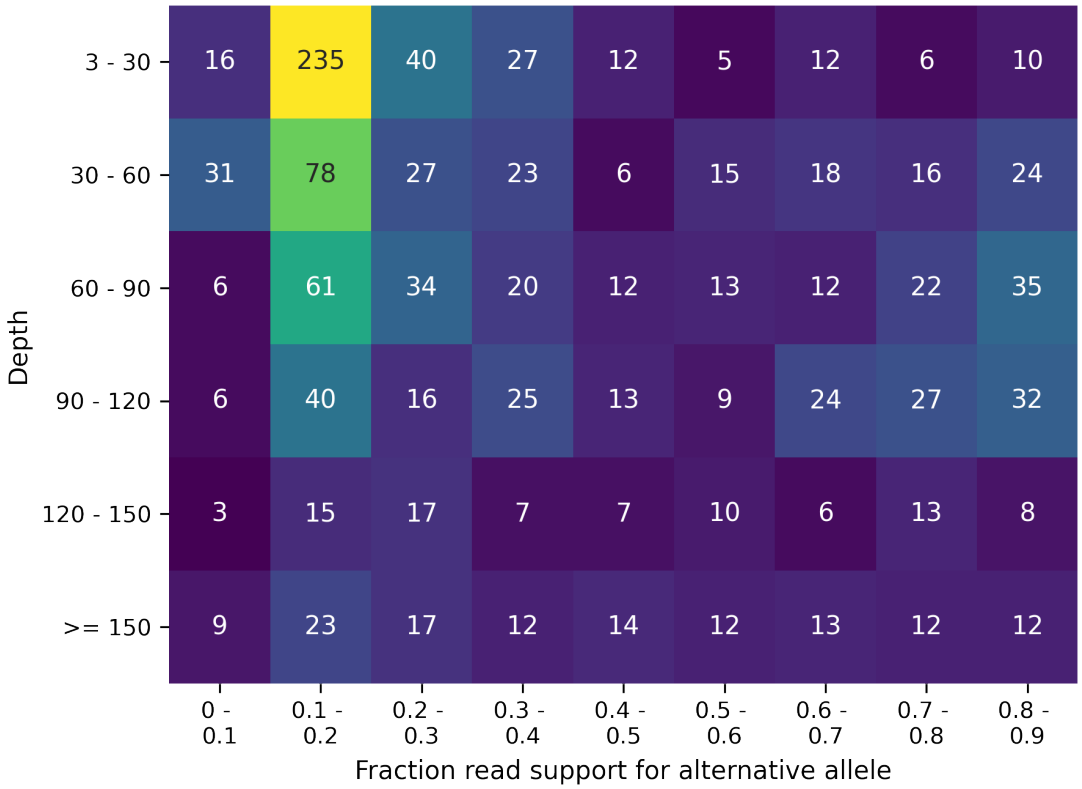

Figure 37 Heatmap of all alternative alleles seen in *gyrA* or *gyrB* in 12,354 CRyPTIC isolates at FRS > 0.9 showing the FRS for the alternative allele and the number of reads at that position.

To examine the extent of isolates that had catalogue mutations in a proportion of the population, I calculated the number of WGS isolates with at least two reads supporting an alternative allele for a catalogue resistance conferring mutation and calculated the proportion of those that had a FRS < 0.9. All of the resistance conferring mutations which are included in the catalogue were seen at FRS < 0.9 in at least one sample (Table 19). There is evidence that some mutations were more frequently seen as part of a mixed population compared to a homogeneous population than others. For instance, the *gyrA* D94G mutation was less commonly seen as part of a mixed population than in a homogenous population compared to *gyrA* D94A, S91P, D94Y or D94H and *gyrB* E501D, D461N or N499D. This is consistent with a study of fluoroquinolone resistant mixed *M. tuberculosis* infections from

Shanghai where the relative frequencies of mutations in mixed populations were inconsistent with that of homogenous populations[315].

MUTATION	N_ISOLATES	N_MIXED	%_MIXED	LB	UB
gyrA D94G	999	136	13.6	11.6	15.9
gyrA A90V	635	98	15.4	12.8	18.5
gyrA D94N	213	38	17.8	13.1	23.7
gyrA D94A	189	42	22.2	16.7	28.9
gyrA S91P	136	31	22.8	16.2	30.9
gyrA D94Y	112	35	31.3	23.0	40.7
gyrA D94H	67	22	32.8	22.2	45.4
gyrB E501D	46	20	43.5	29.3	58.7
gyrB D461N	32	9	28.1	14.4	46.5
gyrA G88C	25	1	4.0	0.2	20.0
gyrB N499D	13	6	46.2	20.4	73.7
gyrA G88A	11	3	27.3	7.3	59.0
gyrB A504V	11	2	18.2	3.2	49.6
gyrB E501V	2	1	50.0	2.7	97.3

Table 19 Isolates with catalogue resistance associated mutations and the proportion of these mutations that are found at < 0.9 FRS. 'N_ISOLATES' is the number of isolates containing the mutation, either as part of a homozygous population or a mixed population, 'N_MIXED' is the number of isolates where there is evidence that the mutation is seen as a mixed allele and '%_MIXED' is the proportion of isolates where the mutation was seen in part of a mixed population. 'LB' and 'UB' represent lower and upper bounds for confidence intervals calculated using the method of Wilson[279].

Conversely, for the five most commonly seen DNA gyrase mutations that are not associated with resistance in the WHO catalogue (and have not been associated with lineage in literature), there was no significant difference in the relative proportion of isolates with the mutation that were seen as part of a mixed population (Table 20). In general, the proportions of non-resistance associated mutations that were seen as part of a mixed population were also lower than for the resistance associated mutations (Table 19).

MUTATION	N_ISOLATES	N_MIXED	%_MIXED	LB	UB
gyrB P94L	247	9	3.64	1.79	6.92
gyrA Q613E	148	6	4.05	1.66	8.77
gyrB S661T	144	4	2.78	0.89	7.12
gyrB I271M	144	6	4.17	1.70	9.01
gyrA P472S	49	1	2.04	0.11	10.95

Table 20 Isolates with common mutations not identified as resistance conferring in the WHO catalogue and the proportion of these mutations that are found at < 0.9 FRS. 'N_ISOLATES' is the number of isolates containing the mutation, either as part of a homozygous population or a mixed population, 'N_MIXED' is the number of isolates where there is evidence that the mutation is seen as a mixed allele and '%_MIXED' is the proportion of isolates where the mutation was seen in part of a mixed population. 'LB' and 'UB' represent lower and upper bounds for confidence intervals calculated using the method of Wilson[279].

I next examined whether the FRS for resistance conferring alleles correlates with the MIC to fluoroquinolones - one might expect that if a resistance conferring allele is seen at higher FRS (i.e. more prevalent within the mixed population) it will have a greater level of resistance. I chose to test the two most frequent mutations seen in fluoroquinolone resistant isolates, *gyrA* D94G and *gyrA* A90V, which are commonly encoded by single nucleotide polymorphisms of adenine to guanine at genome index 7582 and cytosine to thymine at genome index 7570 respectively. I included both susceptible and resistant isolates containing alleles encoding the mutation at FRS < 0.9 and alleles with FRS >= 0.9 were excluded to avoid overfitting to the majority homogeneous population. Pearsons rank correlation coefficients suggest that there was no correlation between FRS for the alternative allele and MIC to either levofloxacin or moxifloxacin in isolates with either *gyrA* D94G or A90V (Figure 38a-d).

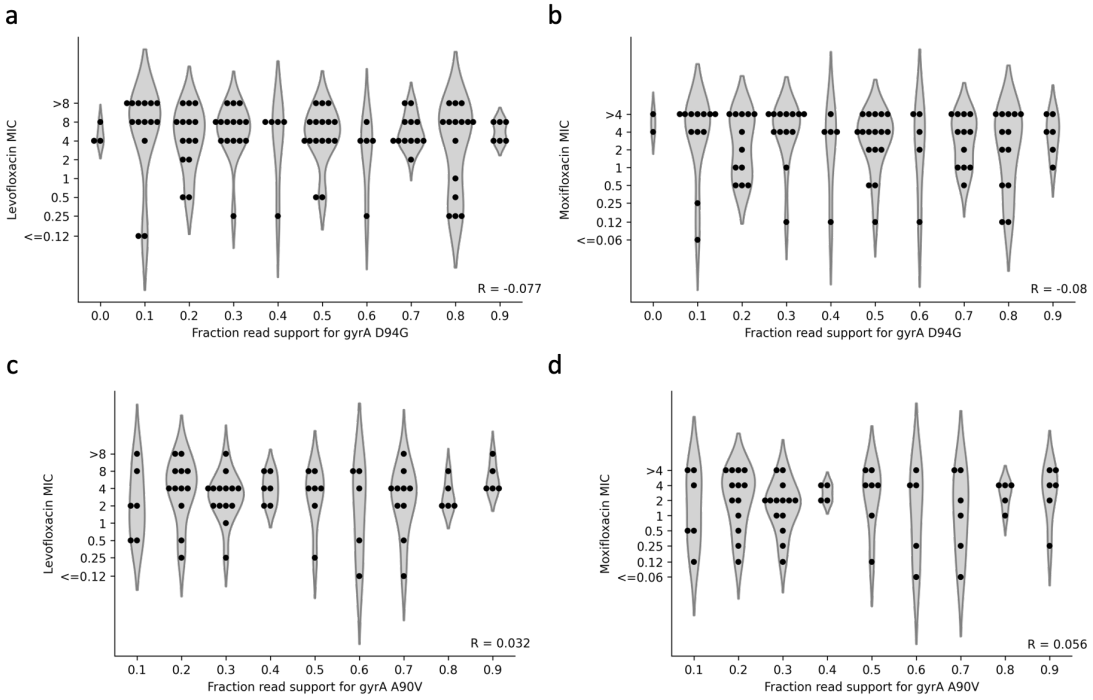

Figure 38 Distribution of FRS for *gyrA* D94G and the (a) levofloxacin and (b) moxifloxacin MIC of the *M. tuberculosis* isolate and the distribution of FRS for *gyrA* A90V and the (c) levofloxacin and (d) moxifloxacin MIC of the *M. tuberculosis* isolate. FRS was rounded to the nearest 0.1 to show the distribution of MIC values at different FRS. R shows the Pearson's rank correlation coefficient between the unrounded FRS and $\log_2$ MIC.

Including the mixed alleles increased the sensitivity of the catalogue significantly, by 9.7% when identifying levofloxacin resistance and 9.5% when identifying moxifloxacin resistance (Figure 39). With inclusion of mixed alleles, the specificity of the catalogue mutations only decreased slightly by 0.5% and 0.8% for levofloxacin and moxifloxacin respectively, and the differences were not significant at $p = 0.05$. This brings the catalogue performance for detecting fluoroquinolone resistance up to the level of first-line drugs such as rifampicin[42, 145], where the determinants of resistance are well understood.

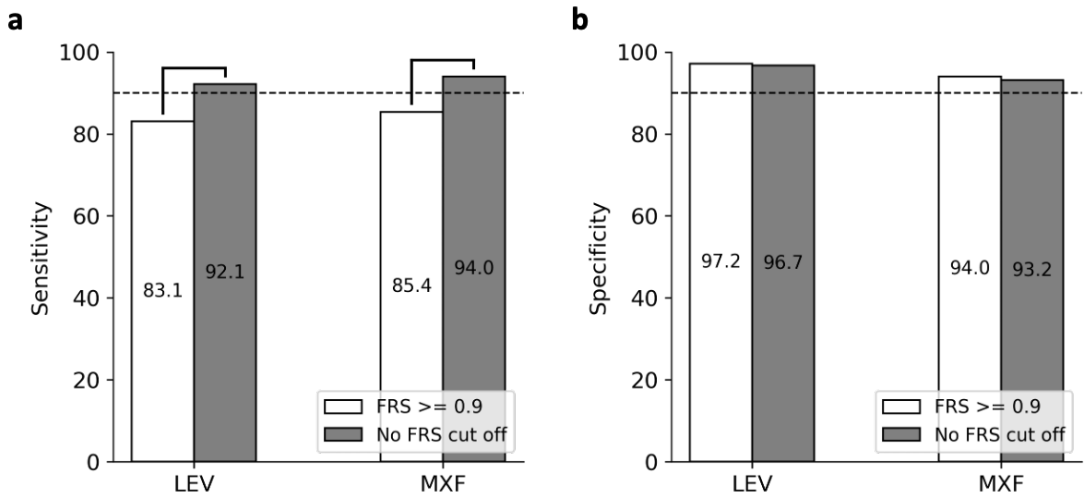

Figure 39 (a) Sensitivity and (b) specificity of fluoroquinolone resistance prediction using WHO 2021 catalogue mutations with and without mixed alleles. Brackets indicate a significant difference (z-test, $p < 0.05$).

For this dataset, 7.9% of levofloxacin resistance and 6.0% of moxifloxacin resistance is still not explained by the presence of catalogue mutations. The majority of fluoroquinolone resistant isolates with unexplained resistance contained no DNA gyrase mutations above those in the lineage background (Table 21). However, 42 of the levofloxacin resistant isolates and 28 of the moxifloxacin resistant isolates that were not identified by catalogue mutations did contain one or more non-synonymous DNA gyrase mutation (Table 21), and it is possible that these could be rare or novel resistance conferring mutations.

	Total	No DNA gyrase mutations	One DNA gyrase mutation	Two or more DNA gyrase mutations
LEV	126	84	33	9
MXF	79	51	22	6

Table 21 Number of phenotypically resistant isolates not predicted as resistant using mutations in the WHO catalogue of resistance associated mutations. 'Total' represents the total number of resistant isolates not predicted as such. 'No DNA gyrase mutations' represents the number of these isolates that did not have a mutation (at any FRS) in either *gyrA* or *gyrB*, 'One DNA gyrase mutation' is the number of isolates that have one mutation either in *gyrA* or *gyrB* and 'Two or more DNA gyrase mutations' is the number of the total that have 2 or more mutations in the *gyrA* and/or *gyrB*.

4.4 Discussion

The CRyPTIC dataset contains a diverse set of fluoroquinolone resistant *M. tuberculosis* isolates (Figure 28a-c). The majority of these isolates were from an MDR background (Figure 28c), which was unsurprising as fluoroquinolone resistance generally evolves after first line drug resistance[101]. Reassuringly, fewer isolates with rifampicin or isoniazid mono resistant backgrounds had fluoroquinolone resistance than those with an MDR background and fluoroquinolone resistance was less prevalent in an isoniazid and rifampicin susceptible background than any of the isoniazid or rifampicin resistant backgrounds (Figure 29).

The majority of fluoroquinolone resistant isolates had a DNA gyrase mutation in the *gyrA* QRDR and a small proportion only had a mutation in the *gyrB* QRDR, or elsewhere in the DNA gyrase genes (at FRS >= 0.9) (Figure 30). A range of mutations were seen in both fluoroquinolone resistant and susceptible isolates in the *gyrA* and *gyrB* QRDRs, including all 14 of the WHO catalogue resistance associated mutations (Figure 31, Figure 32, Figure 33). Consistent with previous studies, *gyrA* D94G and *gyrA* A90V were the most prevalent resistance conferring mutations seen[316].

The presence of multiple DNA gyrase mutations in *M. tuberculosis* isolates may have important implications for predicting fluoroquinolone resistance, especially if MIC is being predicted[203]. Isolates containing two resistance associated mutations at FRS >= 0.9 were rare in this dataset (1.64% of isolates contained two resistance associated mutations), which is expected as double mutations have been shown to generally have a fitness cost compared to the isolates with the individual mutations in the closely related species *M. smegmatis*[317].

gyrA A90V was more prevalent in double mutations than *gyrA* D94G in this dataset (Table 12), probably because *gyrA* A90V confers a lower level of resistance[203, 318], and this pattern has been seen in a previous study[317]. Despite any fitness cost incurred, the isolates with *gyrA* A90V + *gyrA* D94A and *gyrA* A90V + *gyrA* D94G double mutations had significantly higher MICs to levofloxacin and moxifloxacin than isolates with the single mutations in most cases (Figure 34a-d). This will be an important consideration when using higher dosage moxifloxacin treatment for the low-level resistance conferring mutations *gyrA* A90V or *gyrA* D94A[319], as further resistance could arise.

Although there was no significant difference in the prevalence of different DNA gyrase mutations seen at FRS >=0.9 in moxifloxacin resistant compared to levofloxacin resistant isolates (Table 11), there is evidence that there are differences in the genetic determinants of resistance to each drug. Depending on the particular DNA gyrase mutations an isolate possesses, moxifloxacin and levofloxacin have been shown to have different effectiveness and levels of resistance. For example *gyrB* E501D has been found to confer resistance to moxifloxacin and not levofloxacin[195], and high dose moxifloxacin, but not levofloxacin, can successfully treat *M. tuberculosis* containing a *gyrA* A90V mutation in a mouse model[210]. In this work, I found that Lineage 3 isolates with *gyrA* D94G or *gyrA* A90V were associated with lower moxifloxacin MICs compared to Lineage 2 isolates (Table 16, Table 18), but I found no significant difference in levofloxacin MIC for isolates with *gyrA* D94G or *gyrA* A90V in different lineage backgrounds (Table 15, Table 17). This suggests that moxifloxacin treatment could be more effective than levofloxacin in areas where Lineage 3 is particularly common, such as Pakistan (Figure 21). There are of course many other factors to consider when choosing an appropriate fluoroquinolone, such as the cost or safety profile of the medication and I would also argue that the difference in performance of catalogue-based

resistance prediction or resistance diagnosis using molecular tests is an important consideration. It would be preferable to use the drug for which a more reliable resistance prediction could be made, to increase the likelihood of appropriate treatment. I found that detection of moxifloxacin resistance in this dataset had generally higher sensitivity than for levofloxacin (although the difference was not significant at $p < 0.05$ for the WHO catalogue) but the specificity was generally higher for levofloxacin resistance prediction (Figure 36).

In general, the mutations used by molecular diagnostic tests identified over 80% of fluoroquinolone resistance, even in this diverse dataset, and did not perform significantly worse for resistance detection than the mutations that could be detected by WGS and catalogue-based resistance prediction (Figure 36). This is reassuring as lower income, high TB burdened countries may rely on such tests as a cheaper and more easily implementable diagnostic tool with fewer infrastructure requirements than WGS[320]. Despite differences in the number of specific mutations they detect or the specific mutations or regions probed, there was relatively little difference in the performance of molecular diagnostic tests and their different possible interpretations on the dataset as a whole (Figure 36). This is likely because the tests all identify the most prevalent resistance conferring mutations (Table 10).

Although there are concerns that the presence of other mutations in hotspot regions may affect the binding of wild-type or mutant probes[135], I found that, excluding *gyrA* S95T and other known lineage associated mutations, synonymous and non-synonymous mutations within the *gyrA* and *gyrB* QRDR were rare in fluoroquinolone susceptible isolates (Figure 31, Figure 32). This suggests that the overall performance of molecular diagnostic tests will be relatively robust to the adverse effects that these mutations may cause. Indeed the

specificity of levofloxacin and moxifloxacin resistance predictions for the 'resistance inferred' interpretation of tests simulating the absence of wild-type probe binding due to other mutations in the region queried was not significantly worse than for the resistance detected interpretations (where specific resistance conferring mutations are identified) for AID TB FQ/EMB Kit, Genotype MTBDRsl v1, Genotype MTBDRsl v2 or REBA MTB/XDR (Figure 36).

The calculated specificity of both molecular diagnostic tests and catalogue based WGS prediction on this dataset will have been impacted by the small proportion of isolates that were phenotypically susceptible to fluoroquinolones but had resistance associated mutations (Figure 31, Figure 32). As the proportion of susceptible isolates with each individual resistance conferring mutation was generally low, this could be the result of labelling errors during phenotypic testing, which has been suggested to contribute to irregularities between genotype and phenotype in similar studies[145, 272]. The presence of such samples could however be entirely legitimate – especially as all resistance conferring mutations, no matter how rare, had examples of phenotypically susceptible isolates (Figure 31, Figure 32). For isolates containing *gyrA* D94G or A90V mutations, certain lineages, countries of origin or phenotypic resistance backgrounds were negatively associated with levofloxacin and moxifloxacin MIC (Table 15, Table 16, Table 17, Table 18), and MICs of isolates with resistance conferring mutations could therefore fall below the threshold of the ECOFF used to determine phenotypic resistance. This finding suggests that one ECOFF may not be suitable for all genetic backgrounds. Continued updates to critical concentrations also highlights potential problems when using phenotypic DST as standard[115, 116]. Taken together, these factors raise the question as to whether sequencing should replace phenotypic DST as the new reference standard for susceptibility testing[212].

Although the molecular diagnostic tools performed similarly overall, the choice of molecular diagnostic based on the country of use should be considered. The country of origin has an effect on the prevalence of *gyrA* D94G and *gyrA* A90V, after controlling for confounding variables, such as lineage and phenotypic background (Table 13, Table 14). Even after applying a conservative Bonferroni correction, isolates originating from China were less likely to contain a *gyrA* D94G mutation than those from India (Table 13). Certainly, the regression models will not include all possible confounding variables as there was not sufficient data on previous antibiotic treatment, which can lead to evolution of resistance conferring mutations including *gyrA* D94G[321]. However, it is also likely that the environmental factors present in different countries could have their own selective pressures, for example the ease of access to fluoroquinolones for treating other infections, the different treatment programs used for TB or the quality of the health infrastructure may have implications for the likelihood of receiving appropriate treatment and therefore the evolution of resistance.

As lineages are geographically distinct, the effects of lineage should also be considered when choosing an appropriate diagnostic test. In this dataset, the prevalence of resistance conferring mutations was associated with lineage after controlling for country of origin and phenotypic background. Lineage 1 and Lineage 4 isolates were less likely to contain a *gyrA* D94G mutation than Lineage 2 isolates, and Lineage 4 isolates were more likely to contain a *gyrA* A90V mutation compared to Lineage 2 isolates (Table 13, Table 14). It would useful to repeat this analysis for rarer resistance conferring mutations in order to see if there are differences in their prevalence in different geographies and lineages; all the molecular diagnostic tests cover the codons of the most frequently seen resistance conferring variants

(*gyrA* 90, 91 and 94) but the tests vary in the rarer resistance associated mutations they detect (Table 10).

There were also associations seen between the prevalence of resistance conferring mutations and the level of resistance conferred in different phenotypic backgrounds. Rifampicin resistant or MDR isolates were more likely to contain a *gyrA* D94G mutation than isoniazid and rifampicin susceptible isolates and were more positively associated with levofloxacin and moxifloxacin MIC than an isoniazid and rifampicin susceptible background or an isoniazid mono-resistant background (Table 13, Table 15, Table 17). The level of resistance and selection for *gyrA* D94G could be increased in MDR isolates due to sign epistasis; possession of the *gyrA* D94G mutation has been associated with increased fitness in a range of *M. smegmatis* isolates with rifampicin resistance conferring *rpoB* mutations[322]. Increased drug efflux or decreased membrane permeability arising during development of MDR could also play a role in increasing MICs to fluoroquinolones[323]. There was no evidence of association between the different phenotypic backgrounds and the prevalence of *gyrA* A90V, but rifampicin resistance/MDR was more positively associated with the MIC of isolates containing *gyrA* A90V than an isoniazid and rifampicin susceptible background but not isoniazid mono-resistant isolates (Table 14, Table 16, Table 18). This could suggest some interplay with isoniazid resistance specifically. Ultimately *in vitro* evolution studies, MIC and relative fitness testing would be useful to investigate the hypotheses inspired by regression modelling.

The finding of no correlation between FRS and MIC for isolates with either a *gyrA* D94G or A90V mutation at FRS < 0.9 also provides evidence that these mutations are not associated

with a significant fitness cost (Figure 38a-d). This concurs with a study that showed *gyrA* D94G and *gyrA* A90V mutants did not confer a fitness cost to the closely related *Mycobacterium aurum in vitro*[324]. The proportion of isolates with the resistance conferring allele, even if this was in less than 20% of the population that was sequenced, must have grown enough to be visible on the microtiter plate after 14 days of incubation. Further the significantly lower proportion of isolates with *gyrA* D94G in a mixed population compared to several other resistance conferring mutations suggests that isolates with *gyrA* D94G may have better fitness than other resistant populations. Consistent with this hypothesis, *gyrB* mutations were more often seen as part of a mixed population with *gyrB* N499D, E501D and E501V being seen as part of a mixed population over 40% of the time, and these mutations have previously been suggested to have a fitness cost in *Mycobacterium smegmatis*[317]. Although evidence of mixed DNA gyrase alleles was only seen in 5.4% of the total isolates surveyed, mixed populations could be particularly important for the development of fluoroquinolone resistance in TB as a greater proportion of resistance conferring mutations were present in mixed alleles compared to mutations that are not associated with resistance (Table 19, Table 20).

A significant proportion of fluoroquinolone resistant isolates (16.0%) appeared to have no DNA gyrase mutations above a background of lineage associated mutations, which was surprising as the DNA gyrase is the sole antibiotic target of fluoroquinolones in *M. tuberculosis*. Through the work of this chapter, I found that the proportion of fluoroquinolone resistant isolates without DNA gyrase mutations is likely much lower than this figure, which is artificially high due to the FRS cut offs used in the CRyPTIC bioinformatic pipeline. The near 10% improvement in sensitivity without a major reduction in specificity for detecting fluoroquinolone resistance using catalogue mutations (Figure 39a,b) highlights

the importance of including mixed alleles in bioinformatic pipelines and also ensuring molecular diagnostic tests can detect low frequency resistance conferring alleles. The magnitude of improvement in sensitivity correlates with the 10% frequency of resistance conferred by alleles seen in mixed populations estimated from twelve independent studies during systematic review[108].

There are likely more minor resistance conferring alleles that could not be detected in this study due to the limitations of the sequencing technology used (Figure 37), and this could explain why a proportion of resistant isolates that were not predicted as such appeared to contain no *gyrA* or *gyrB* mutations (Table 21). The varying limits of detection in this study also makes it difficult to suggest the most appropriate FRS to use as a cut off during variant calling, especially when considering that isolates with *gyrA* D94G and A90V seen at low FRS (FRS < 0.2) had high MICs to levofloxacin and moxifloxacin (Figure 38a,b). One could instead use a conservative variant caller with high precision, for example LoFreq, which could detect minor SNPs at 3% of the population, to avoid setting a specific cutoff[325]. Although deep sequencing may be more suited to the study of minor populations in *M. tuberculosis* infection, the Illumina WGS data here provides a reasonable estimation of what could be detected in practice and can inform the sequencing depths that might be necessary for future study or diagnosis.

The CRyPTIC dataset contains a large and diverse set of fluoroquinolone resistant isolates and is therefore a valuable resource for studying fluoroquinolone resistance. Through this study, I have found that the genetic determinants of fluoroquinolone resistance are much more complex than the assumptions employed for catalogue-based resistance diagnosis or

molecular diagnostic testing. Reassuringly, these diagnostic tools can still perform well despite their limitations, but it is important to understand that they are not comprehensive and that the mechanisms of fluoroquinolone resistance are still not fully understood.

5 Chapter 5: Prediction of fluoroquinolone resistance using machine learning

5.1 Introduction

The collection of large datasets of WGS and phenotypic data to build mutation catalogues means that a wealth of information on resistance and susceptibility is available. The catalogue-based predictive approach is inherently biased by the underlying population structure of the isolates collected and may not capture rare resistance conferring variants if they do not pass the statistical criteria used for classification[42]. Another option, utilising the large datasets used to build catalogues, is to train *predictive* models to classify an infection as resistant or susceptible to a drug from their genome sequence. In tuberculosis, machine learning models have been trained to accurately predict resistance and susceptibility to a range of antitubercular drugs and predict MDR and XDR phenotypes [200, 264-266, 326], and can account for the underlying population structure of the data[326].

Machine learning models for AMR prediction often use the presence or absence of particular genes or variants within a gene as features[264-266] and so may be limited in how well they can generalise to new variants that evolve[213]. It may therefore be advantageous to include more general predictors. A promising option is to consider the changes to the structural and physiochemical properties that occur when an amino acid is mutated and the resulting effect on the interactions between the antibiotic target and the drug, rather than simply the presence or absence of a particular genetic mutation. The effect of these changes can be predicted in a quantifiable way; tools such as mCSM-lig use a combination of graph-based

structural signatures from crystal structures and physiochemical properties to train machine learning models to predict changes in protein ligand affinity associated with mutations in a range of targets[327]. Affinity predictions made by mCSM-lig correlated well with experimental data and the tool was also able to predict around 80% of the mutations associated with resistance in selected cancer and HIV drug targets[327].

In tuberculosis, protein structures have been used to help validate mutations predicted by machine learning models to be associated with resistance[326]; for example if mutations are close to the drug binding site this suggests that they are more likely to be associated with resistance. The use of protein structures to directly inform the input features for resistance prediction in tuberculosis has been somewhat limited. Protein secondary structure, amino acid composition, physicochemical properties, evolutionarily relevant properties have also been used to build feature sets for support vector machines which could predict protein sequences associated capreomycin resistance with > 80% accuracy on a small dataset of sequences[328]. Resistance and susceptibility to pyrazinamide associated with *pncA* mutations can be predicted using protein structure and physiochemical properties as features for logistic regression models, support vector machines and neural networks[329]. However, the models were trained using phenotypes for clinical isolates that had been aggregated by mutation and the best model did not have as high sensitivity scores when predicting the phenotypes of unique clinical isolates[329]. More recently, Jamal et al. used physiochemical properties to train a range of machine learners to predict resistance to rifampicin, isoniazid, pyrazinamide and fluoroquinolones in *rpoB*, *inhA/katG*, *pncA* and *gyrA/gyrB* genes respectively[330]. An artificial neural network model achieved an AUC of 1 on a non-redundant test set for several of the genes, although, the test contained <10 mutations for *inhA* and *gyrB*[330].

Thus far, structure-based machine learning models have been evaluated based on their ability to predict a binary resistant or susceptible phenotype which has been inferred by the aggregation of phenotypes associated with the presence of a particular mutation. It would be advantageous to predict resistance in individual isolates and predict the level of resistance conferred, particularly as the ECOFFs or breakpoints used to determine resistance and susceptibility in tuberculosis are continually updated[115, 116, 331]. Certainly, for fluoroquinolones, it is valuable to predict the level of resistance an isolate has because high dose moxifloxacin may be able to treat isolates with low level fluoroquinolone resistance[209]. Although structure-based features can predict a good proportion of resistance mutations, not all drug resistance in tuberculosis is target mediated. Other genetic factors, such as lineage and background, have important associations with resistance levels to many antitubercular drugs including the fluoroquinolones (see Section 4.3.5). Therefore, it is imperative that other genetic features also be considered when predicting resistance and especially when predicting the level of resistance that an isolate with a drug target mutation has.

Despite some machine learning models having high ROC AUC and accuracy scores, the use of algorithms in the clinic remains low; this is likely in part due to the 'black box' nature of approaches such as neural networks[213]. To build trust for proposed use in clinical settings, machine learning algorithms need to be transparent so that the rationale behind a prediction of resistance can be understood[213]. It is therefore encouraging that simple, interpretable models such as random forests and gradient boosted trees have performed well for predicting resistance in *M. tuberculosis*[200, 264, 265, 326], and these models have the added benefit of identifying potential resistance signatures[326]. However, a study by Chen et al. found that their neural network based approach outperformed random forest and logistic

regression models[266] and had better ROC AUC scores for 8 out of 10 drugs than another study using gradient boosted trees[265]. For protein structure and physiochemical based resistance predictions, neural network approaches have also provided better overall performance than the other machine learning models tested[329, 330], but these did not include forest-based approaches.

The aim of this chapter is to use the CRyPTIC consortium dataset to build and evaluate interpretable machine learning models to predict whether an *M. tuberculosis* isolate is resistant to levofloxacin and moxifloxacin, and the level of resistance the isolate has. To do this, I will use a combination of structure-based features associated with DNA gyrase mutations and the isolate's genetic background. I will also use the best performing models to identify important predictive features and predict the effect of all possible non-synonymous DNA gyrase mutations in a common background.

5.2 Methods

5.2.1 Dataset

Only *M. tuberculosis* isolates containing a singular *gyrA* or *gyrB* non-synonymous mutation (referred to as a "solo mutation") were considered for training machine learning models; this permits us to assume that the effect on MIC compared to wild type is solely attributable to this change. I define a solo mutation as an isolate containing a singular non-synonymous *gyrA* or *gyrB* mutation above the background of any lineage or phylogeny specific mutations[267]. I included mutations present as part of a mixed population, as the FRS was not correlated with the MIC for either levofloxacin or moxifloxacin for the two most common

resistance conferring mutations (Figure 38a-d). Known lineage-associated mutations were ignored as these are well characterised as susceptible to fluoroquinolones[267] and lineage can be included as a feature in machine learning models to account for any other effects. The final datasets used for training and testing machine learners included 2,226 *M. tuberculosis* isolates that had solo mutations and levofloxacin MIC data, and 2,088 *M. tuberculosis* that had solo mutations and moxifloxacin MIC data. The distribution of isolates MICs used for training and testing machine learning algorithms is shown in Figure 40a-b.

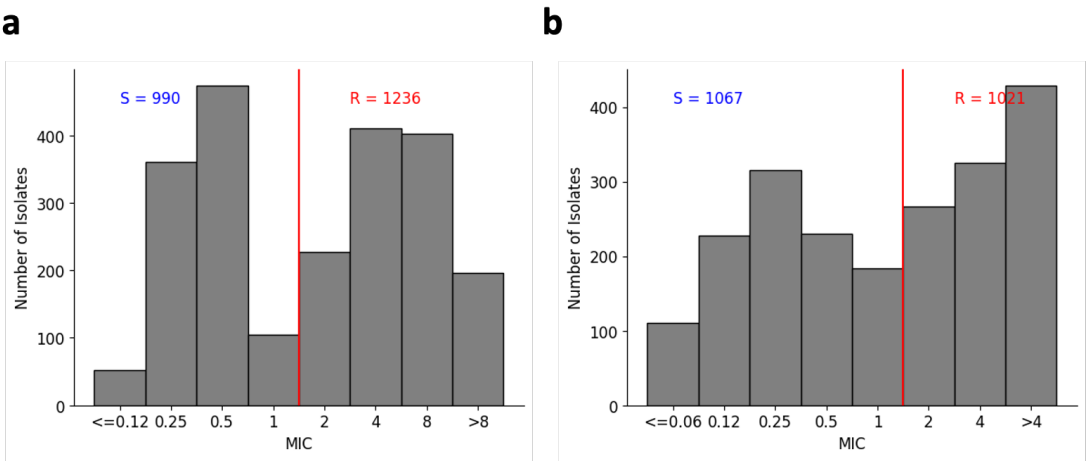

Figure 40 Distribution of (a) levofloxacin and (b) moxifloxacin Minimum Inhibitory Concentrations of *M. tuberculosis* isolates with a singular DNA gyrase mutation that were used for training and testing of machine learning models. The red line indicates the epidemiological cut off value (ECOFF) that was used to distinguish resistant and susceptible isolates.

5.2.2 Feature set for machine learning

Chemical features associated with the amino acid change of the solo mutation were calculated by: property of wild type amino acid – property of mutant amino acid. This was calculated for the number of atoms, sidechain volume, Kyte-Doolittle hydropathy score, electrical charge, number of hydrogen bond donors, number of hydrogen bond acceptors, number of rotatable bonds, number of aromatic rings and number of sulphur atoms. I also

used the SNAP2 score as a feature; this gives an estimation of the size of the effect of a mutation based on features including the evolutionarily conservation of the wild type amino acid[332].

The Protein Data Bank accessions 5BS8, of moxifloxacin bound DNA gyrase, and 5BTG, of levofloxacin bound DNA gyrase, were used to calculate all structure and distance-based features[184]. The secondary structure of each amino acid was calculated using the DSSP package[333]; the secondary structure was assumed not to change upon mutation. The relative solvent accessibility of the wild type amino acid was calculated by using solvent accessibility data from DSSP output files divided by the surface area of the wild type amino acid. Protein-ligand interaction profiler (PLIP)[334] predictions were used to assess whether there were any interactions, such as water or salt bridges, between the wild type amino acid and the moxifloxacin or levofloxacin drug. The depth of the wild type amino acid in the protein was calculated using the DEPTH server[335] and mCSM tools[336] were used to predict the effects of different mutations on protein-protein interactions between the *gyrA* and *gyrB* subunits and protein stability. The number of hydrogen bonds formed between the wild type amino acid and either other parts of the protein or the DNA, were calculated using the MDAnalysis class HydrogenBondAnalysis[337]. The distances of the wild type amino acid from the ligand, the ligand-bound magnesium ion, the catalytic magnesium ion and the DNA were calculated using MDAnalysis distance_array. Since the DNA gyrase structure contains two copies of the *gyrA* gene product and two of the *gyrB* gene product, there are two mutations to consider; the distance of both were measured and the minimum distance was taken as a consensus.

Several positions with mutations in the dataset in the *gyrA* N and C termini and in the *gyrB* N terminus are not resolved in the crystal structures. These amino acids are far from the fluroquinolone binding site and therefore the values of the various distances were set to the maximum distance seen in the rest of the dataset. For relative solvent accessibility, residue depth, DNA distance, number of hydrogen bonds formed with protein, number of hydrogen bonds formed with DNA and the predicted protein-protein affinity change, values were imputed to be the mean of all other values for that feature. All other features derived from the crystal structures for these mutations were set to zero, corresponding to 'no change'.

Whether the mutation was present as a mixed allele was also included as a feature, with 1 indicating a 'mixed population' and 0 being 'homozygous'. Lineage was included as a feature to capture any lineage-specific effects that could arise from these mutations or differences in other areas of the genome, as genetic background affect the level of fluoroquinolone resistance[304], (Table 16, Table 18). The background resistance to the first line drugs isoniazid and rifampicin was also included as a feature as fluoroquinolone resistance is associated with resistance to one or more first line drugs[338], (Table 15, Table 16, Table 17, Table 18). Country of origin was also included as I previously found this to be associated with different levels of resistance conferred by resistance associated mutations (Table 15, Table 16, Table 17, Table 18). All categorical features were one hot encoded, giving a total of 59 different features to be considered for training.

5.2.3 Machine learning pipeline

All machine learning pipelines were created using the scikit-learn python package[339]. Random forest classification and logistic regression algorithms were implemented via scikit-learn and, for these algorithms, balanced class weights were used throughout training. The MORD wrapper[223, 340] was used to implement ordinal methods including ordinal ridge, ordinal all threshold, ordinal intermediate threshold and ordinal squared error and an xgboost[216] wrapper was used to implement xgboost classification.

A range of classification models were trained to predict the binary phenotype (resistant or susceptible) to moxifloxacin or levofloxacin, using 'resistant' as the positive class. Ordinal models were also trained to predict the moxifloxacin and levofloxacin MIC. For this the MIC values were categorically encoded to a group number 0 to 7 in ascending order, e.g. for moxifloxacin an MIC of 0.06 was assigned group 0, an MIC of 0.12 was assigned to group 1 et cetera.

Isolates with singular DNA gyrase mutations were shuffled and split into stratified training (80%) and test (20%) sets. Using the training set, for each machine learner, features were subject to feature selection using cross validated (five-fold) recursive feature elimination using that method's default hyperparameters. Scoring was based on ROC AUC score (see Section 2.2.3) for methods predicting binary phenotype and on accuracy for methods predicting MIC.

Features chosen by the selection algorithm were then used in the scikit learn default models when tuning hyperparameters. Hyperparameter tuning was done by applying a random search algorithm to choose 50 different combinations of hyperparameter values for each default model and each selection of hyperparameters was evaluated using 10 times repeated 5-fold cross validation. For all methods, the number of estimators was tuned using random numbers between 100 and 1000. For random forest methods, maximum depth, minimum samples per split, and minimum samples per leaf were tuned using random integers between 2 and 6 (a minimum of 2 was used to help reduce overfitting to the training data). For xgboost methods, maximum depth was also tuned using random integers between 2 and 6 and the learning rate was tuned using uniform values between 0.05 and 0.1. For ordinal and logistic regression methods the regularisation parameter was tuned using uniform values between 0.01 and 10. For models predicting binary phenotype, ROC AUC was used to score the hyperparameters and for the models predicting MIC, accuracy was used. The final hyperparameters selected for each of the binary classification models are shown in Table 22 and the hyperparameters selected for predicting MIC are shown in Table 23.

Model class	*Levofloxacin (R/S)*	*Moxifloxacin (R/S)*
Logistic Regression	C=3.347	C=0.574
	max_iter=971	max_iter=443
Random Forest Classification	n_estimators=369	n_estimators=162
	max_depth=5	max_depth=5
	min_samples_split=3	min_samples_split=5
	min_samples_leaf=3	min_samples_leaf=5
XGBoost Classification	n_estimators=439	n_estimators=120
	max_depth=3	max_depth=2
	learning_rate=0.052	learning_rate=0.123

Table 22 Hyperparameters selected for binary classification models to predict levofloxacin or moxifloxacin resistance. Unless explicitly stated in section 5.2.3, all other parameters were default.

Model	*Levofloxacin (MIC)*	*Moxifloxacin (MIC)*
Ordinal All Threshold	alpha=0.143 max_iter=700	alpha=6.635 max_iter=101
Ordinal Intermediate Threshold	alpha=5.997 max_iter=714	alpha=0.241 max_iter=574
Ordinal Squared Error	alpha=0.065 max_iter=238	alpha=6.635 max_iter=101
Ordinal Ridge	alpha=0.241 max_iter=574	alpha=3.755 max_iter=960
Random Forest Classification	n_estimators=230 max_depth=5 min_samples_split=5 min_samples_leaf=5	n_estimators=352 max_depth=5 min_samples_split=3 min_samples_leaf=3
XGBoost Classification	n_estimators=848 max_depth=5 learning_rate=0.083	n_estimators=140 max_depth=4 learning_rate=0.057

Table 23 Hyperparameters selected for multi-class classification models to predict levofloxacin or moxifloxacin MIC. Unless explicitly stated in section 5.2.3, all other parameters were default.

The best models from hyperparameter tuning were fit on the full training set using the selected best features before making predictions on the test set.

Please see https://github.com/alicebrankin/thesis_notebooks for the codebase to reproduce the analyses and figures in this chapter.

5.3 Results

5.3.1 Binary classifiers

To assess whether machine learning could be used to predict levofloxacin and moxifloxacin resistance and susceptibility using DNA gyrase structure, chemistry and genetics-based predictors I trained and evaluated three simple machine learning algorithms; logistic regression, random forest classification and xgboost classification. A detailed introduction to these methods can be found in Section 2.2.1.

Between 20 and 53 of the 59 features were selected by recursive feature elimination for training the binary classification models to predict levofloxacin resistance; the xgboost classifier used the fewest features, and the logistic regression model the most (Appendix Table 1). At least one physiochemical, structural, genetic, geospatial and phenotypic background feature was selected in each of the three models. The models had many features in common including the change in volume of the amino acid, the change in number of aromatic rings, the distances from the amino acid to the ligand / ligand-coordinated magnesium ion / catalytic magnesium ion / DNA, whether the amino acid was present in an α-helix, whether the mutation was present as a mixed allele, whether the genetic background was Lineage 2 or Lineage 4, whether the phenotypic background was isoniazid and rifampicin susceptible or MDR and whether the isolate was from China, India, Nepal, Vietnam or South Africa.

To predict moxifloxacin resistance recursive feature elimination selected 34 features for the xgboost model and 41 features for both random forest classification and logistic regression (Appendix Table 2). As with levofloxacin resistance prediction, at least one physiochemical, structural, genetic, geospatial and phenotypic background feature was selected in each of the three models. Many of the features that the levofloxacin resistance prediction models had in common were also used for all moxifloxacin resistance prediction models with some exceptions and some additional features. The features used by all levofloxacin but not moxifloxacin resistance prediction models include the change in number of aromatic rings, the distances from the amino acid to the ligand-coordinated magnesium ion and DNA, whether the amino acid was present in an α-helix, and whether the isolate was from

Vietnam. Features used by all moxifloxacin but not levofloxacin resistance prediction models include the change in amino acid atom number, hydropathy, number of hydrogen donors, number of hydrogen acceptors, the predicted change in protein stability and protein-protein affinity, whether the isolate was in Lineage 3, whether the isoniazid or rifampicin resistance phenotype was unknown and whether the isolate was from Italy or Germany.

None of the models for either levofloxacin or moxifloxacin resistance prediction were overfit to the training set (Figure 41a,b). For levofloxacin resistance prediction, logistic regression had the highest ROC AUC score on the test set at 0.96 and for moxifloxacin resistance prediction the xgboost classifier and logistic regression models had the highest ROC AUC score on the test set at 0.93 (Figure 42a,b).

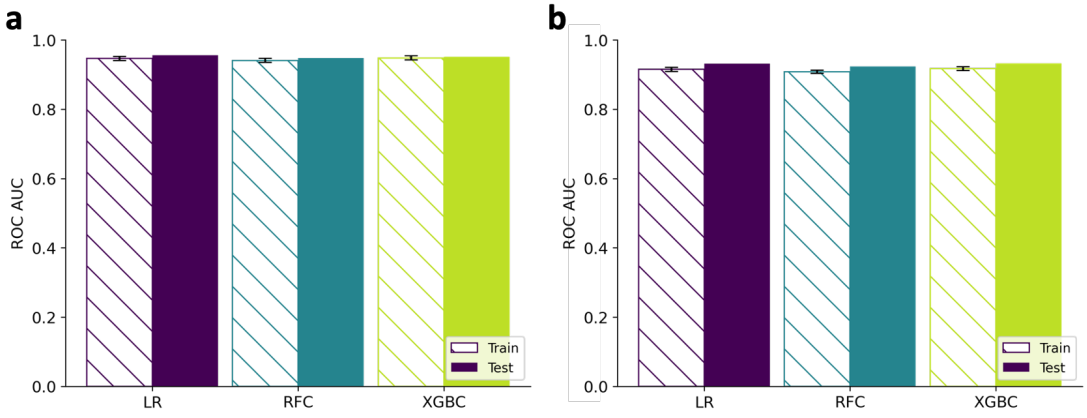

Figure 41 ROC AUC scores of models predicting resistance on training sets compared to the test set for (a) levofloxacin and (b) moxifloxacin. Error bars show the range of results from 5 validation sets. LR = logistic regression, RFC = random forest classification, XGBC = xgboost classification

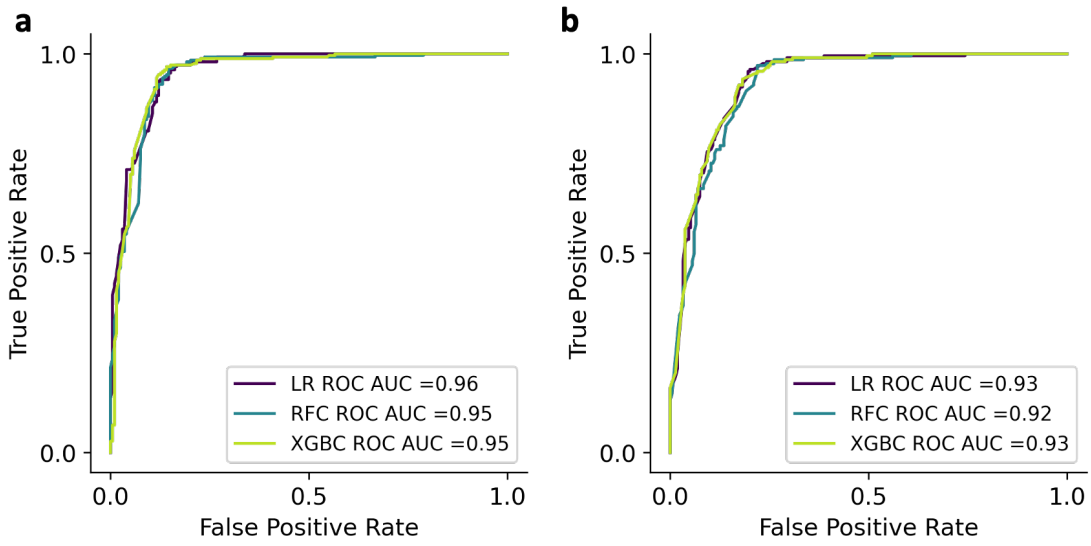

Figure 42 ROC curves of models predicting resistance on the test set for (a) levofloxacin and (b) moxifloxacin.

In terms of sensitivity, the random forest and xgboost classifier models performed best on the test set for levofloxacin resistance prediction, achieving 96.8% sensitivity (Figure 43a). The specificity scores of the models were lower, at 85.9% and 83.9% for xgboost and random forest classifiers respectively. For moxifloxacin resistance prediction, the random forest classifier had the highest sensitivity on the test set, at 97.5% but this model had significantly lower specificity than either the logistic regression or the xgboost classification models (Figure 43b). For both levofloxacin and moxifloxacin resistance prediction, the three models trained had higher sensitivity than specificity on the test set (Figure 43a,b).

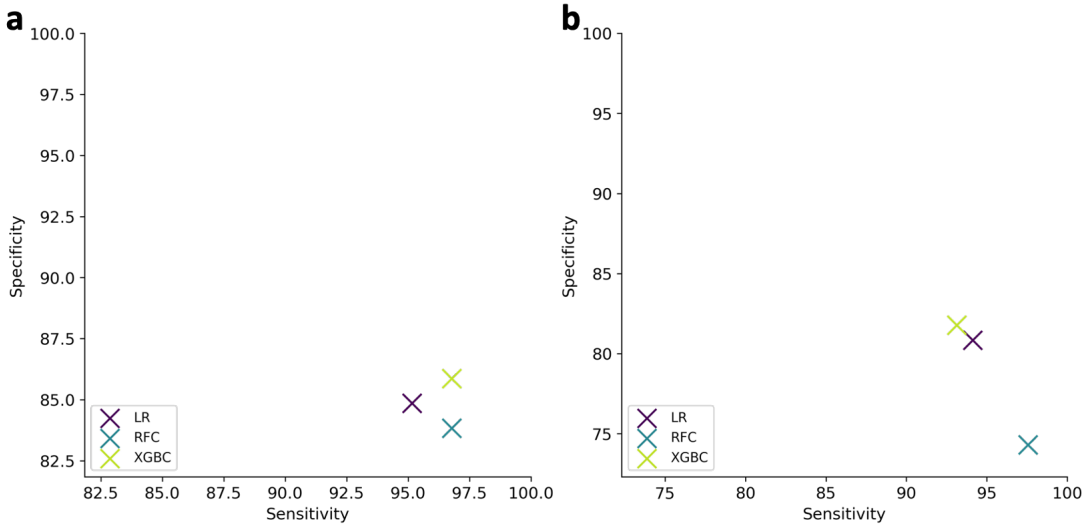

Figure 43 Sensitivity and specificity of resistant predictions on the test set for (a) levofloxacin and (b) moxifloxacin.

5.3.2 Interpretation of best models

The 'best' model was chosen as the model that scored the highest sensitivity on the test set, for levofloxacin this was the xgboost classifier and for moxifloxacin this was the random forest classifier. I investigated the relative importance of the features used to fit this model to the training set and make predictions. For both levofloxacin and moxifloxacin resistance prediction, the distance of the mutated amino acid from the drug-coordinated Magnesium ion was the most important feature, but the relative importance of the feature was much higher for levofloxacin than moxifloxacin resistance prediction (Table 24, Table 25). The other three distance-based features (distance of the amino acid from the ligand, catalytic magnesium or DNA) were more important for both levofloxacin and moxifloxacin resistance prediction than other structure based and physiochemical features. MDR phenotypic background was also more important for making predictions than most other features. The origin of the isolate and the lineage background were also generally less important for making levofloxacin and moxifloxacin resistance predictions. Interestingly, the Snap2 score

was the fifth most important predictor for moxifloxacin resistance with a feature importance of 0.105, however the feature was not used to make predictions for levofloxacin resistance at all (Table 25, Table 24).

Feature	*Importance*
LIGAND_MG_DISTANCE	0.720
LIGAND_DISTANCE	0.058
BACKGROUND_MDR	0.055
CATALYTIC_MG_DISTANCE	0.018
DNA_DISTANCE	0.018
LINEAGE_NAME_Lineage 2	0.017
IS_HET	0.013
COUNTRY_WHERE_SAMPLE_TAKEN_ZAF	0.013
COUNTRY_WHERE_SAMPLE_TAKEN_CHN	0.011
PROTEIN_HBONDS	0.010
COUNTRY_WHERE_SAMPLE_TAKEN_NPL	0.010
COUNTRY_WHERE_SAMPLE_TAKEN_PER	0.009
SECONDARY_STRUCTURE_H	0.009
AROMATIC_RING_CHANGE	0.008
BACKGROUND_INH_AND_RIF_S	0.007
COUNTRY_WHERE_SAMPLE_TAKEN_VNM	0.006
COUNTRY_WHERE_SAMPLE_TAKEN_IND	0.006
VOLUME_CHANGE	0.005
LINEAGE_NAME_Lineage 4	0.005
RSA	0.004

Table 24 Relative importance of features used by an xgboost classification model to predict levofloxacin resistance. The importance scores from all selected features sum to 1, for clarity all are rounded to 3 d.p.

Feature	*Importance*
LIGAND_MG_DISTANCE	0.155
LIGAND_DISTANCE	0.120
DNA_DISTANCE	0.108
CATALYTIC_MG_DISTANCE	0.107
SNAP2_SCORE	0.105
BACKGROUND_MDR	0.071
RESIDUE_DEPTH	0.059
RSA	0.036
SECONDARY_STRUCTURE_H	0.035
CHARGE_CHANGE	0.025
PROTEIN_HBONDS	0.018
STABILITY_CHANGE	0.018
BACKGROUND_INH_MONOR	0.014
HYDROPATHY_CHANGE	0.013
SECONDARY_STRUCTURE_NaN	0.013
LINEAGE_NAME_Lineage 3	0.013
LIGAND_INTERACTION_NaN	0.010
VOLUME_CHANGE	0.009
ATOM_CHANGE	0.009
HACCEPTOR_CHANGE	0.008
PP_AFFINITY_CHANGE	0.007
LIGAND_INTERACTION_none	0.007
BACKGROUND_INH_AND_RIF_S	0.006
DNA_HBONDS	0.006
COUNTRY_WHERE_SAMPLE_TAKEN_ZAF	0.005
ROTABLE_BOND_CHANGE	0.003
LINEAGE_NAME_Lineage 4	0.003
COUNTRY_WHERE_SAMPLE_TAKEN_PAK	0.003
IS_HET	0.002
BACKGROUND_UNKNOWN	0.002
COUNTRY_WHERE_SAMPLE_TAKEN_CHN	0.002
COUNTRY_WHERE_SAMPLE_TAKEN_IND	0.001
LINEAGE_NAME_Lineage 2	0.001
HDONOR_CHANGE	0.001
COUNTRY_WHERE_SAMPLE_TAKEN_ITA	0.001
COUNTRY_WHERE_SAMPLE_TAKEN_PER	0.001
COUNTRY_WHERE_SAMPLE_TAKEN_DEU	0.000*
COUNTRY_WHERE_SAMPLE_TAKEN_NPL	0.000*
COUNTRY_WHERE_SAMPLE_TAKEN_KGZ	0.000*
COUNTRY_WHERE_SAMPLE_TAKEN_TKM	0.000*
LINEAGE_NAME_Lineage 1	0.000*

Table 25 Relative importance of features used by a random forest classification model to predict moxifloxacin resistance. The importance scores from all selected features sum to 1, for clarity all are rounded to 3 d.p. * indicates a feature importance score of less than 0.0005.

5.3.3 Evaluation of model performance on mutations with known effects

To evaluate the likely performance of the best models for predicting levofloxacin and moxifloxacin resistance associated with specific mutations that have known effects, I used

the models to predict resistance for isolates with each of the WHO catalogue resistance associated mutations as solo. Although they scored relatively low in terms of feature importance, the genetic background features were selected as model predictors. I therefore made predictions for the mutations in isolates with a Lineage 2, MDR background from India as this represented the largest group of isolates used in training, and therefore the model is likely most optimal for this background. Although, in doing this, I note that isolates with the exact same feature set may have been seen in training, so this does not constitute a 'blind' prediction. The xgboost model for predicting levofloxacin resistance and random forest classifier model for predicting moxifloxacin resistance predicted all the catalogue resistance associated mutations in *gyrA* as resistant (Table 26). However, the models predicted some of the *gyrB* resistance associated mutations to be susceptible; *gyrB* E501D, E501V and A504V were predicted susceptible to levofloxacin and *gyrB* D461N and A504V were predicted susceptible to moxifloxacin.

Mutation	*Predicted levofloxacin phenotype*	*Predicted moxifloxacin phenotype*
gyrA G88A	R	R
gyrA G88C	R	R
gyrA A90V	R	R
gyrA S91P	R	R
gyrA D94A	R	R
gyrA D94G	R	R
gyrA D94H	R	R
gyrA D94N	R	R
gyrA D94Y	R	R
gyrB D461N	R	S
gyrB N499D	R	R
gyrB E501D	S	R
gyrB E501V	S	R
gyrB A504V	S	S

Table 26 Fluoroquinolone resistance prediction for MDR Lineage 2 *M. tuberculosis* isolates from India with WHO catalogue resistance associated mutations as solo mutations in DNA gyrase genes. A xgboost classification model was used for levofloxacin resistance prediction and a random forest classifier model for moxifloxacin resistance prediction.

I next used the models to predict resistance for *M. tuberculosis* isolates only containing lineage associated mutations that were assumed to be susceptible in the MDR, Lineage 2, Indian background. As the mutations were not considered when defining solo mutations, the models had not explicitly seen the structural and chemical features associated with these mutations in training. As expected, almost all the lineage associated mutations were predicted susceptible to both levofloxacin and moxifloxacin, including *gyrA* S95T which is proximal to the drug binding site and adjacent to the most prevalent resistance associated mutation *gyrA* D94G (Table 27). However, the susceptible *gyrA* A90G mutation was predicted to be resistant in the MDR, Lineage 2, Indian background.

Mutation	*Predicted levofloxacin phenotype*	*Predicted moxifloxacin phenotype*
gyrA E21Q	S	S
gyrA T80A	S	S
gyrA A90G	R	R
gyrA S95T	S	S
gyrA G247S	S	S
gyrA S250A	S	S
gyrA R252L	S	S
gyrA A384V	S	S
gyrA L398F	S	S
gyrA A463S	S	S
gyrA D639A	S	S
gyrA G668D	S	S
gyrA L712V	S	S
gyrA V742L	S	S
gyrB M291I	S	S
gyrB V301L	S	S
gyrB A403S	S	S

Table 27 Fluoroquinolone resistance prediction for MDR Lineage 2 *M. tuberculosis* isolates from India with lineage associated susceptible mutations as solo mutations in DNA gyrase genes. A xgboost classification model was used for levofloxacin resistance prediction and a random forest classifier model for moxifloxacin resistance prediction.

5.3.4 Prediction of novel resistance conferring mutations

The models were next used to predict the effect of all possible non-synonymous mutations in either *gyrA* or *gyrB* in isolates with a Lineage 2, MDR background from India. At least one non-synonymous mutation was predicted to be associated with levofloxacin resistance by the xgboost classifier model at 67 of the 838 amino acid positions in the *gyrA* protein (8.0%), and 38 of the 675 amino acid positions in the *gyrB* protein (5.6%), (Appendix Table 3, Appendix Table 4). Most of these mutations are located in the interior of the DNA gyrase cleavage complex and are clustered around the drug binding sites in the core (Figure 44a). Several positions far from the binding site, at the exterior of the cleavage complex, are also predicted to have resistance associated mutations. In contrast, for moxifloxacin resistance, the random forest model predicted just 21 of the positions in *gyrA* (2.5%) and 16 of the positions in *gyrB* (2.4%) to be associated with resistance (Appendix Table 5, Appendix Table 6)). The positions predicted by the random forest model to have resistance associated mutations are within the interior of the DNA gyrase cleavage complex and are clustered around the drug binding sites (Figure 44b). None of the residues forming the *gyrA* DNA exit gate or coiled coil domains were predicted to be associated with resistance to either levofloxacin or moxifloxacin. It is important to note that, due to the low specificity of both the models (Figure 43a,b), a number of these mutations will likely be false positive predictions.

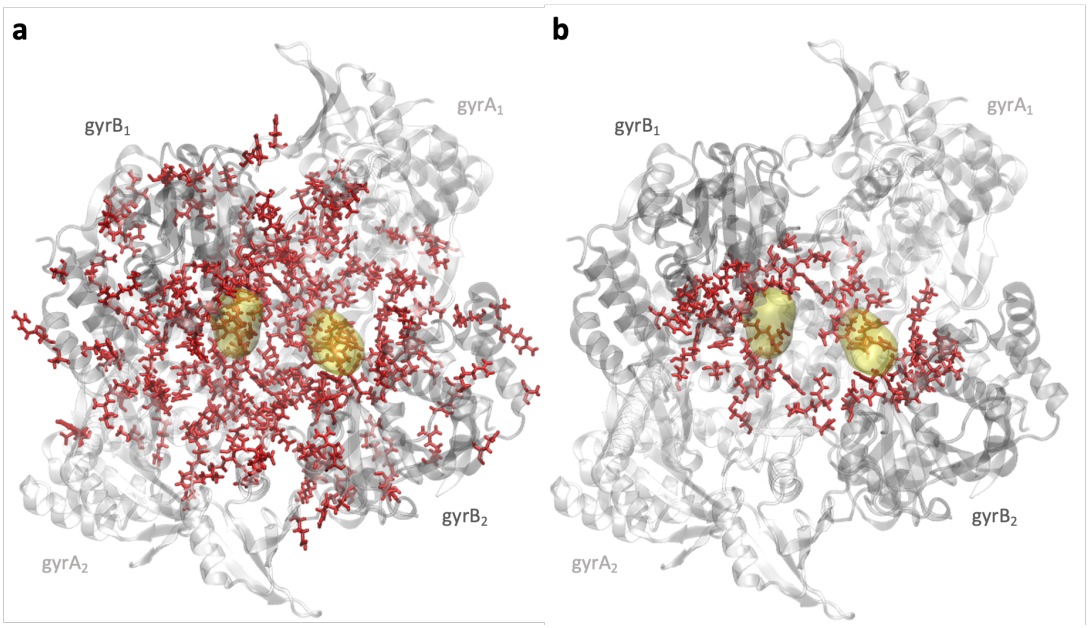

Figure 44 Positions of mutations predicted by machine learning models to be associated with resistance to levofloxacin (a) and moxifloxacin (b). The fluoroquinolone binding site is shown in yellow and predicted resistance conferring mutations in red. The full list of mutations predicted to be resistant to levofloxacin can be found in Appendix Table 3 and Appendix Table 4, and predicted resistant to moxifloxacin in Appendix Table 5 and Appendix Table 6. Image created in VMD[187] from PDB structures 5BS8 and 5BTG[184].

Any non-synonymous mutation at certain positions within *gyrA* and *gyrB* was predicted to confer resistance to fluoroquinolones. For levofloxacin, an isolate with any mutation at *gyrA* positions 52, 74, 81, 87, 88, 90, 91, 94, 98, 125, 128 and 129 or *gyrB* positions 460, 461, 482, 483 and 484 were predicted to be resistant (Appendix Table 3, Appendix Table 4), although it is unlikely positions 128 and 129 would mutate *in vivo* as these are highly conserved and essential for DNA gyrase function[341, 342]. For moxifloxacin, there were fewer examples of positions where any non-synonymous mutation resulted in a prediction of resistant, these were; *gyrA* 88, 89, 94 and 129 and *gyrB* 460, 482 and 483 (Appendix Table 5, Appendix Table 6).

5.3.5 Multi-class classifiers

To predict the levofloxacin and moxifloxacin MIC of isolates I trained a range of different machine learners including ordinal methods (ordinal all threshold, ordinal intermediate threshold, ordinal squared error and ordinal ridge) and multiclass classification models (random forest classification and xgboost classification); a full description of these methods can be found in Section 2.2.2. The training and testing sets were identical to those used for the binary classifiers.

The features, and number of features selected for levofloxacin and moxifloxacin MIC prediction were not the same as for the binary classifiers although at least one physiochemical, structural, genetic, geospatial and phenotypic background feature was selected as before (Appendix Table 7, Appendix Table 8). For levofloxacin MIC prediction, between 26 and 55 features were selected for the different machine learners with the random forest model using the least and the ordinal squared error model using the most. For moxifloxacin MIC prediction between 44 and 55 features were selected for the different models with the xgboost model using the least and the ordinal squared error and random forest models using the most. Similarly to the binary models, common features were selected for many of the models and the features selected for moxifloxacin and levofloxacin MIC prediction differed (Appendix Table 7, Appendix Table 8).

For both levofloxacin and moxifloxacin, the accuracy of correctly predicting MIC on both the training and test set was low and most algorithms were not overfit to the training set (Figure 45a,b). For both levofloxacin and moxifloxacin, the forest-based classification models

performed better than the ordinal methods on the test set; for levofloxacin and moxifloxacin MIC prediction the xgboost classification model performed best with 52.5% and 47.5% accuracy on the test set respectively.

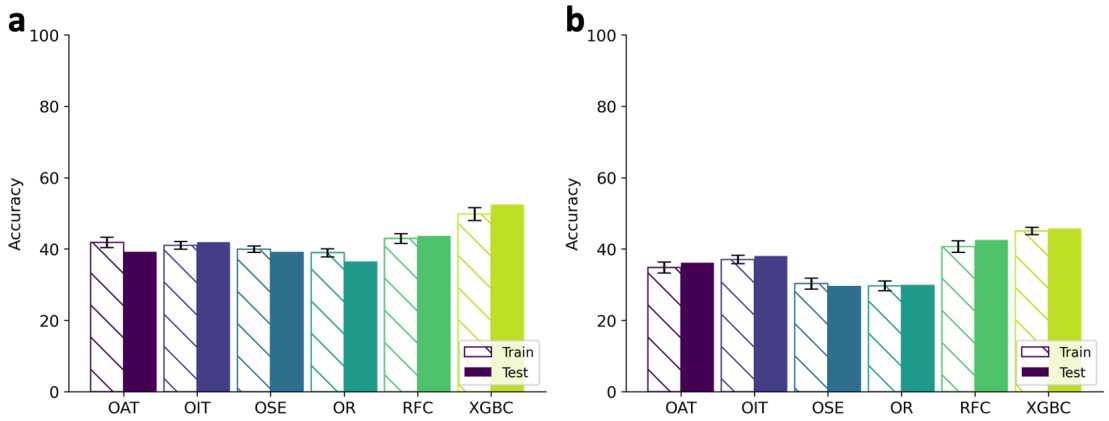

Figure 45 Accuracy of models in predicting MIC labels on training sets compared to the evaluation set for (a) levofloxacin and (b) moxifloxacin. Error bars show the range of results from 5 validation sets. OAT = ordinal all threshold, OIT = ordinal intermediate threshold, OSE = ordinal squared error, OR = ordinal ridge, RFC = random forest classification, XGBC = xgboost classification

Essential agreement is a metric commonly used to evaluate the performance of tests measuring MICs and an MIC is defined as being in essential agreement with the true MIC value if it is within one doubling dilution[343]. For prediction of levofloxacin and moxifloxacin MICs, the ordinal all threshold and xgboost classification models performed best with 87.0% and 81.1% essential agreement respectively (Figure 46a,b).

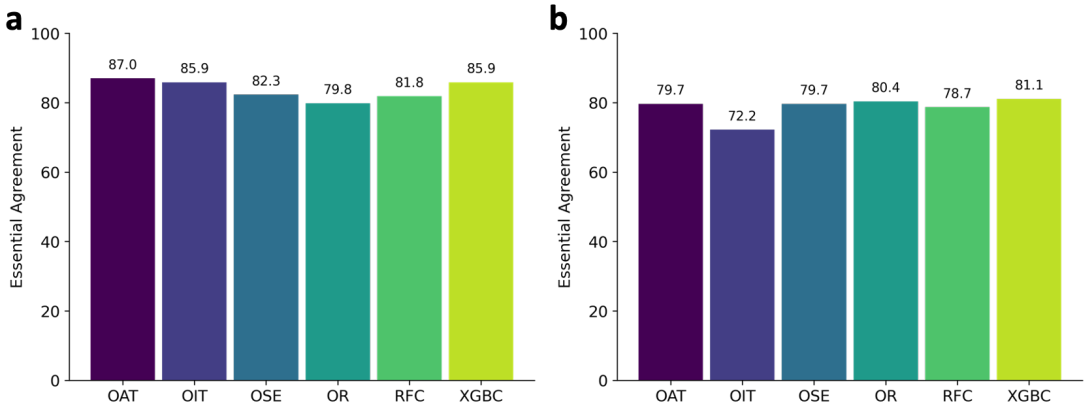

Figure 46 Essential agreement of model MIC predictions on the evaluation set for (a) levofloxacin and (b) moxifloxacin.

To see how the models performed in terms of sensitivity and specificity for predicting resistance, we binarised MIC predictions into resistant and susceptible using the ECOFFs proposed by the CRyPTIC Consortium[272]. As it is particularly important to reduce the very major error rate when making resistance predictions (a prediction of susceptible when an isolate is resistant), it is important to individually consider the sensitivity and specificity of a model. For predicting levofloxacin resistance, the ordinal ridge model had the best sensitivity score on the test set at 96.8%, although this model did score the lowest on specificity at 83.3% (Figure 47a). For moxifloxacin resistance prediction the ordinal all threshold model had the best sensitivity score on the test set at 93.6% but only achieved 79.4% specificity (Figure 47b).

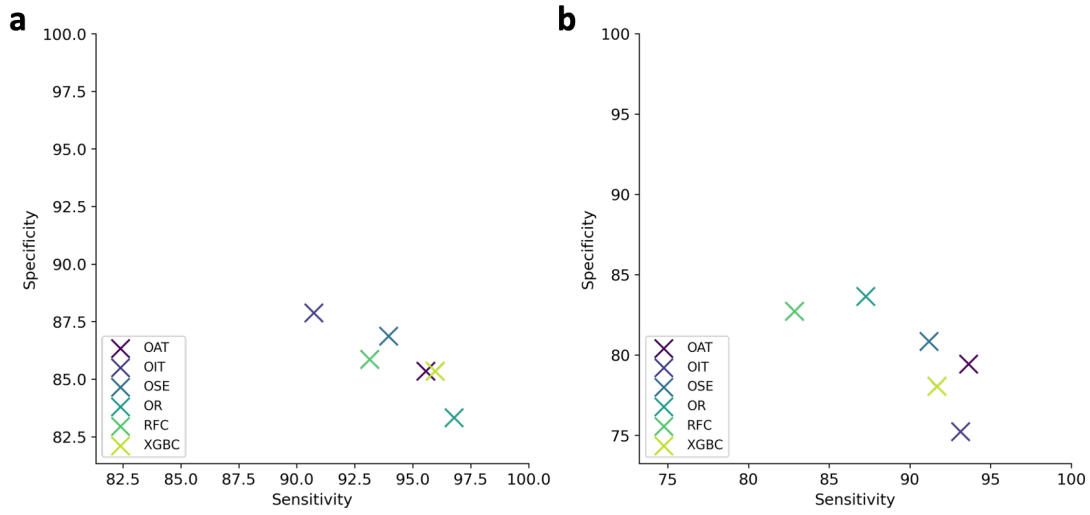

Figure 47 Binary performance of MIC prediction models. Sensitivity and specificity of resistance predictions on the test set for (a) levofloxacin and (b) moxifloxacin.

5.4 Discussion

I found that machine learning models trained using protein structure-based features, physiochemical features and features describing genetic background can be used to predict resistance and susceptibility of *M. tuberculosis* isolates to levofloxacin and moxifloxacin (Figure 41a,b), the level of resistance that is conferred (Figure 45a,b, Figure 46a,b) and predict potential novel resistance conferring mutations (Appendix Table 3, Appendix Table 4, Appendix Table 5, Appendix Table 6). Recursive feature elimination selected at least one physiochemical, structural, genetic, geospatial and phenotypic background feature for every model, suggesting that a combination of all these factors is important for fluoroquinolone resistance prediction.

Distance-based features from crystal structure were selected as features for all binary and MIC predictive models for levofloxacin and moxifloxacin (Appendix Table 1, Appendix Table 2, Appendix Table 7, Appendix Table 8), suggesting that DNA gyrase distance based features are important for predicting fluoroquinolone resistance. The distance of the mutation from the ligand co-ordinated Magnesium ion was the most important feature in the best performing binary classification models (Table 24, Table 25). This is not unexpected as the prevalent *gyrA* D94G mutation is suggested to confer resistance by disrupting a water bridge formed between *gyrA* and the drug coordinated Magnesium ion[184]. The relative importance of the distance of the mutation from the ligand co-ordinated Magnesium ion was different for levofloxacin and moxifloxacin resistance prediction models, and some different features were selected for levofloxacin compared to moxifloxacin resistance prediction models (Table 24, Table 25, Appendix Table 1, Appendix Table 2, Appendix Table 7, Appendix Table 8). This adds to evidence suggesting there are differences in the determinants of resistance to each drug[195, 203, 204].

In general, the models predicted levofloxacin resistance better than moxifloxacin resistance; binary classifiers had higher ROC AUC performance on the test set (Figure 41a,b) and MIC predictors had higher accuracy and essential agreement on the test set for levofloxacin versus moxifloxacin (Figure 45a,b, Figure 46a,b). This suggests that levofloxacin resistance can be predicted more reliably than moxifloxacin resistance, which has been previously suggested for a catalogue-based approach that was used by the CRyPTIC Consortium[1]. This could be due to the ECOFF of moxifloxacin being incorrect or because the separation between the resistant and susceptible population distributions is greater for levofloxacin (Figure 40a,b). Together with the finding that different machine learning algorithms performed best for the two different fluoroquinolones, this suggests that levofloxacin and

moxifloxacin resistance prediction should be treated as separate tasks rather than simply predicting 'fluoroquinolone' resistance, which has been done in other studies to increase the size of the training data[344].

For levofloxacin and moxifloxacin, both binary and MIC prediction models achieved over 90% sensitivity on the testing set, this is encouraging because clinically it is most important to reduce very major error (VME) rates (a prediction of susceptible when an isolate is resistant). In terms of sensitivity, the best performing binary classification models were xgboost for levofloxacin resistance prediction and random forest for moxifloxacin resistance prediction (Figure 43a,b), and these models have been particularly successful in other AMR prediction studies[345]. However, when making binary predictions from MIC predictions, the ordinal ridge and ordinal all threshold models had higher sensitivity scores than the forest based models (Figure 47a,b). The best sensitivities for levofloxacin and moxifloxacin resistance prediction were 96.8% and 97.5% respectively and the methods therefore have a VME rate of around 3% which is the maximum tolerated for the approval of a diagnostic test by the International Standards Organisation. The specificity of models was lower than the sensitivity in all cases, with no model achieving greater than 90% specificity for either drug (Figure 43a,b, Figure 47a,b). Although the major error (ME) rate is less important than the VME rate, it is important that the ME rate is low because a prediction of resistant when an isolate is susceptible could prevent a patient receiving the most effective treatment. This has implications for the clinical usage of such an algorithm; these algorithms may be best suited to rapidly predict susceptibility in cases where a catalogue cannot make a prediction; this would rule out resistance so that appropriate treatment can be started quickly. Only isolates predicted resistant by the algorithm would need to undergo standard phenotypic DST.

For predicting levofloxacin and moxifloxacin MIC, the forest-based methods had better accuracy compared to the ordinal methods, although the best accuracy scores on the test set were only 52.5% for levofloxacin resistance and 47.5% for moxifloxacin resistance (Figure 45a,b). The low accuracy could reflect the variation in MIC measurements, but could also be due to the limitations of using crystal structures to derive features; they do not show movement and are determined under non-physiological conditions. Therefore, the structures may not be truly representative of the protein *in vivo* which could have implications for the accuracy of calculated distance-based features or those based on hydrogen bonds. One should also consider that the models don't consider the whole genome, only the lineage, country and isoniazid/rifampicin resistance background. Hence, resistance signatures such as efflux, membrane permeability and transcriptional regulation are not taken into account. These mechanisms are not yet well characterised for fluoroquinolone resistance but could be added to the models as and when they are elucidated. It is noted that these kinds of resistance signatures are inherently considered and even elucidated in some genetics based machine learning models[326].

A method is considered acceptable for clinical MIC measurement if it has >=90% essential agreement with the gold standard method[343], however this threshold was not reached by any of the machine learners that I trained (Figure 46a,b). Despite having lower accuracy than the forest-based methods, the ordinal all threshold model was closest to the 90% threshold, achieving 87% essential agreement on the test for predicting levofloxacin resistance; this model could be a good candidate for further optimisation. The performance of the predictive model could be improved by feature engineering[330] or model stacking[328]. Another

option could be to retrain models using the essential agreement to score feature selection and hyperparameter tuning algorithms, rather than accuracy.

The best models for levofloxacin and moxifloxacin resistance prediction were able to predict a number of putative resistance conferring non-synonymous *gyrA* and *gyrB* mutations that were not seen in the dataset or the resistance catalogue (Appendix Table 3, Appendix Table 4, Appendix Table 5, Appendix Table 6). This adds to evidence that models based on structural and physiochemical data can generalise to predict novel mutations[329, 330]. Encouragingly, only a small proportion of amino acids within *gyrA* and *gyrB* were predicted to have resistance conferring mutations, and as the models only had moderate specificity (Figure 43a,b), it is likely that a number of these predictions will be false positives. *In vitro* evidence will be required to confirm which of the mutations identified by the models are most likely to confer resistance. Previous *in vitro* supercoiling assays provide evidence for *gyrA* A74S as a resistance conferring mutation[346] and the xgboost model predicted the mutation as resistant to levofloxacin in this study (Appendix Table 3).

The models predicted that all mutations at *gyrA* positions 88 and 94 would confer resistance to levofloxacin and moxifloxacin (Appendix Table 3, Appendix Table 5). It therefore could prove beneficial to generalise the catalogues used in bioinformatic pipelines to predict any non-synonymous mutation at these positions as resistant rather than just the commonly seen variants that reach sufficient statistical criteria. Indeed, there are already five resistance associated variants at *gyrA* D94 and two at *gyrA* G88 in the WHO catalogue[42], and there were an additional two *gyrA* D94 mutations seen in fluoroquinolone resistant CRyPTIC isolates (Figure 31).

Some mutations that are associated with resistance in the WHO catalogue were not predicted to be resistant to levofloxacin and moxifloxacin (Table 26). For example, the resistance associated mutations at position *gyrB* E501 were not predicted resistant to levofloxacin but were to moxifloxacin. This could be because the *gyrB* E501D mutation is characterised as conferring a higher level of resistance to moxifloxacin than levofloxacin[195, 204]. Additionally, in the WHO catalogue, the *gyrB* E501D mutation was upgraded to be 'associated with resistance-interim' for levofloxacin by expert rules, rather than the standardized statistical approach, due to its accepted association with moxifloxacin resistance[42, 145]. The *gyrB* D461N mutation was predicted resistant to levofloxacin but not moxifloxacin, this could be due to the ECOFFs used to determine resistance and susceptibility in this study; an MIC of 1 mg/l was the cut off above which an isolate is considered resistant to both fluoroquinolones[272]. A strain containing another mutation at the position, D461H, was shown to have an MIC of 4 mg/l to levofloxacin and 0.5 mg/l to moxifloxacin, so it is possible there is differences in the level of resistance conferred by mutations at this position to the different drugs[347]. The *gyrB* A504V mutation was not predicted resistant to either drug, however this WHO catalogue resistance associated mutation was 'associated with resistance - interim' and had uncertain significance[42, 145].

Encouragingly, almost all the lineage specific mutations were predicted to be susceptible to both levofloxacin and moxifloxacin (Table 27). It is especially encouraging that *gyrA* S95T was predicted susceptible as it is close to the moxifloxacin binding site and adjacent to the resistance conferring mutation *gyrA* D94G. A notable exception is *gyrA* A90G, which was predicted resistant to both fluoroquinolones. The mutation is well characterised as non-

resistance conferring[136, 303], has evolved in a sub group of Ugandan *M. tuberculosis* strains and has been found in Congo, Turkey and Somalia[135]. It cannot be ruled out that *gyrA* A90G could confer resistance in the Indian, MDR, Lineage 2 background for which the prediction was made, although this is unlikely as the mutation was also shown to be fluoroquinolone susceptible in *M. smegmatis*[348]. There is only one example of an isolate with *gyrA* A90G in the dataset used, and the prediction of resistant for the mutation suggests that the machine learning models may not be able to generalise to geographical areas or sub lineages that are not well represented within the CRyPTIC dataset.

There are limitations to consider when interpreting the machine learning algorithms that I have presented in this chapter. Firstly, the dataset used for training and testing has many examples of the prevalent mutations *gyrA* D94G and A90V (Figure 33). Although this reflects real world data, and many of the isolates will have different genetic backgrounds and origins, it is important to acknowledge the likelihood that there is some redundancy in the training and testing sets used. The high prevalence of some mutations may also mean that important predictive information from rarer mutations may be lost. Although, rare resistance conferring mutations were predicted resistant to both levofloxacin and moxifloxacin by the respective machine learning models (Appendix Table 3, Appendix Table 5).

Secondly, the rifampicin and isoniazid resistance phenotype were important features chosen for all of the machine learning models (Appendix Table 1, Appendix Table 2, Appendix Table 7, Appendix Table 8), but this information would not be accessible at the time that resistance predictions would need to be made in the clinic. As WGS data would be collected

and analysed by bioinformatic pipelines prior to making a fluoroquinolone resistance prediction, one could use the presence or absence of common isoniazid and rifampicin resistance conferring mutations, or indeed resistance predictions from algorithms for these drugs specifically, to infer this phenotypic information.

Finally, the trained models are limited in that they do not consider interactions that could occur where multiple DNA gyrase mutations are present, for example the restorative effect of *gyrA* A90G with some resistance conferring mutations in *gyrA*[303]. Interactions between multiple mutations should certainly be considered for future work, but due to the scarcity of multiple mutations in resistant isolates the collection of larger datasets with more examples of isolates with multiple mutations will be necessary.

In summary, machine learning models trained using structure-based features and physiochemical features, in addition to features describing genetic background, can predict phenotypic fluoroquinolone resistance in *M. tuberculosis* isolates with high sensitivity. I have also shown that it is possible to predict the level of resistance an isolate has and predict novel resistance conferring mutations. With further optimisation, and as more *M. tuberculosis* isolates are collected, machine learning models using structural based features could provide rapid resistance diagnosis to fluoroquinolones from *M. tuberculosis* WGS data. Through this work, and the work of others[327-330], I believe that structure-based and physiochemical properties associated with genetic mutations represent a significant opportunity for generalisable resistance prediction in tuberculosis and other diseases.

6 Chapter 6: Prediction of fluoroquinolone resistance and susceptibility associated with DNA gyrase mutations using free energy calculations

6.1 Introduction

As shown in the previous chapter, machine learning is a promising avenue to predict fluoroquinolone resistance associated with DNA gyrase mutations using structure-based features. However, the nature of the models means we understand little about the chemical rationale underlying the predictions made. Further, crystal structures also have important limitations. For example, they do not show movement and are determined under non-physiological conditions; therefore, they may not be truly representative of the protein *in vivo*. Molecular dynamics (MD) can simulate the dynamics of biological macromolecules in solution and allows the quantification of thermodynamic properties using theories derived from classical statistical mechanics. MD simulations can therefore be used to calculate the binding affinity of a protein to its ligand[235].

Simulating protein-ligand binding in order to compute the absolute binding free energy of the interaction is challenging, time consuming and error prone, due to the large number of atoms perturbed in the simulations. When we wish to calculate the difference in binding free energy between two related protein-ligand systems, we can instead use alchemical perturbation theory to facilitate *relative* binding free energy (RBFE) calculations. The number of atoms being perturbed is reduced, improving accuracy and error reduced compared to computing ABFE as the start and end states are more closely related. Please see Section 2.3

for a full introduction to the relevant alchemical methods and RBFE calculations used for this chapter.

Thus far RBFE methods have been largely applied to predicting how alterations to a lead compound affect its affinity to a protein target in small molecule drug design, where free energies can be predicted with roughly 1 kcal mol^{-1} error[349-351]. RBFE methods also provide a promising avenue for resistance prediction; if one assumes that a mutation causes resistance by disrupting drug binding, a protein with a resistance conferring mutation will have a lower binding affinity for the ligand than the wild-type protein, i.e. the difference in binding free energies will be positive. Predicting the effects of amino acid mutations should be practically easier than for small molecules, because forcefields are well parameterised for amino acids[352] and free software is available to automatically create the mutant amino acid topologies[353].

Several authors have now demonstrated that RBFE methods can produce accurate predictions of resistance and susceptibility for genetic mutations seen in both infectious disease and cancer[354-357]. These resistance prediction studies have applied alchemical free energy methods to simple, small monomeric targets and therefore prove the principle, but no more. However, the majority of drug targets, especially for Tuberculosis treatments, are more complex. For example, the DNA gyrase cleavage complex is tetrameric, includes bound DNA and is 199 kDa in size[184].

The calculation of free energies according to best practice requires a large number of MD simulations; firstly, apo and drug-bound systems must have their energy minimised and their dynamics equilibrated from the state in the crystal structure, then a simulation is required for each lambda window for each of three alchemical transformation steps for both the apo and drug-bound system[236, 237]. The calculations should also be repeated at least three times in order to ensure statistical reliability[236]. Therefore, the approach requires large amounts of computational resource to complete calculations in a reasonable timeframe. Despite improvements in computational architecture and software[358, 359], the simulations required may still take several days, especially if a large system is simulated[358] or if the available resource is not optimal[360].

It may be possible to use shorter simulations to reduce the time taken to make a binary prediction of resistant or susceptible to a drug on mutation of its target. The successful use of very short λ-simulations has been demonstrated for trimethoprim which targets the dihydrofolate reductase enzyme in *Staphylococcus aureus*[361]. Despite limitations in the study – the free energy estimates were not converged – the accuracy and precision of predictions were still sufficient to call mutations as resistant or susceptible to trimethoprim[361]. Clinically, as the binary labels of resistant and susceptible are used in making treatment decisions, the sensitivity and specificity of a predictive method is more important than its quantitative accuracy and precision, in contradistinction to how the performance of RBFE methods are usually assessed.

The aim of this chapter is to assess the suitability of using short RBFE calculations to predict whether individual DNA gyrase mutations confer resistance to fluoroquinolones or not. Due

to the computationally expensive nature of the method, I chose to only test the method for moxifloxacin.

6.1.1 Selection of test mutations

As described previously (see Section 1.4.3), *M. tuberculosis* DNA gyrase has a number of resistant and susceptible mutations that have been catalogued for moxifloxacin. I selected a small number of mutations to test how well alchemical free energy methods could predict fluoroquinolone resistance and susceptibility. The positions of the chosen mutations relative to the moxifloxacin binding site are shown in Figure 48 and the relative distances of the wild-type amino acids from moxifloxacin are presented in Table 28.

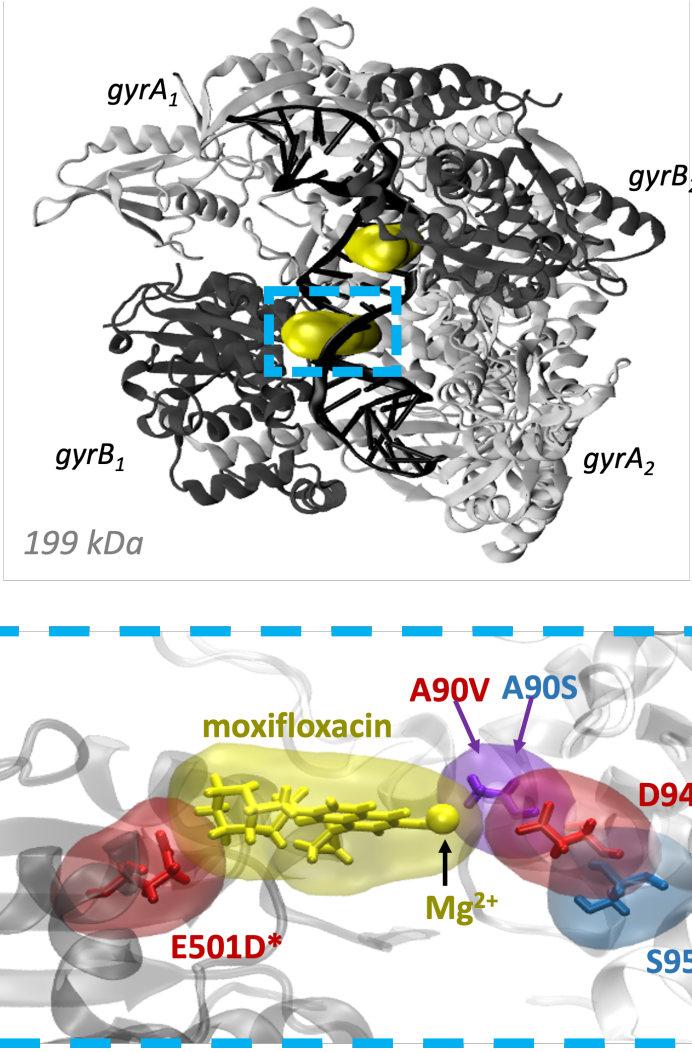

Figure 48 Structure of *M. tuberculosis* DNA gyrase cleavage complex[184], showing the selected clinical mutations associated with antibiotic resistance and susceptibility relative to the moxifloxacin binding site. For clarity nucleic acids are hidden in the close-up visualisation. Resistance-conferring mutations are drawn in red, those associated with susceptibility blue and those residues where different mutations confer different resistance phenotypes purple. An asterisk (*) indicates a mutation, all other mutations are in *gyrA.*

MUTATION	DISTANCE FROM MXF (Å)	EXPECTED PHENOTYPIC EFFECT
S95T	9.0	Susceptible
A90S	2.7	Hyper-susceptible
A90V	2.7	Resistant
E501D*	2.6	Resistant
D94G	5.1	Resistant

Table 28 Distances of DNA gyrase test mutations from moxifloxacin (MXF). Measurements are taken for the wild-type amino acid in the A or B chain (for *gyrA* and *gyrB* respectively) to the nearest bound moxifloxacin molecule. * indicates a *gyrB* mutation.

For resistant mutations, I chose to test the most common resistance conferring mutations, *gyrA* D94G and A90V[199-201], as I can have the greatest confidence in their expected effect. *gyrA* D94G poses a challenging test to the method because it involves a large change in atom number and a charge change. The mutation is also expected to cause a higher level of resistance than the other commonly seen resistance conferring mutation, *gyrA* A90V[201, 203], allowing me to investigate whether the method can distinguish between different levels of resistance conferred by different mutations. I also chose to test E501D in *gyrB*[362] which is close to the antibiotic binding site (Figure 48, Table 28), but not in the well characterised *gyrA* quinolone resistance determining region (QRDR).

For the susceptible mutations I selected *gyrA* S95T since it is very common – it is found in almost all samples except the H37Rv reference genome – and has been shown biochemically to have no effect on fluoroquinolone binding affinity[346]. This mutation is close to the drug binding site (Figure 48, Table 28) and within the *gyrA* QRDR.

A *gyrA* A90S mutation has not been seen clinically (and therefore is not present in existing catalogues), however the mutation provides an avenue to test whether RBFE methods can predict hyper susceptibility as well as resistance. Tuberculosis DNA gyrase is suggested to have some innate resistance to fluoroquinolones which is attributed to an Alanine at position 90[184]. Other bacterial gyrases such as in *E. coli*, which is more susceptible to fluoroquinolone treatment, have a Serine at position 90 which is thought to support the water network between *gyrA* D94 and the drug-coordinated Magnesium ion due to Serine's greater hydrophilicity compared to Alanine[363] due to the sidechain hydroxyl. Therefore, *gyrA* A90S is expected to increase fluoroquinolone binding affinity with TB DNA gyrase compared

to the wild-type. As this mutation is also at the same site as the well characterised resistance conferring mutation close to the binding site, *gyrA* A90V (Figure 48, Table 28), a prediction of susceptibility for *gyrA* A90S would disprove the possibility that a resistance prediction for *gyrA* A90V is simply a result of perturbing atoms near the moxifloxacin molecules – a difficult test for the method to pass.

6.2 Methods

6.2.1 System set up and equilibration

A schematic overview of the setup process from crystal structure to molecular dynamics simulation is presented in Figure 49. Firstly, the structure of moxifloxacin-bound DNA gyrase cleavage complex with DNA (PDB:5BS8)[184] was modified to create starting drug-bound and apo structures for MD simulations. One of the two DNA molecules was removed from the structure as the crystallographer could not determine directionality, and the ends of the DNA were shortened to give ending nucleotides that were compatible with GROMACS processing. Missing loops in *gyrB* were modelled in using Modloop[364] and PROPKA[365] was used to determine protonation states of amino acids. An apo structure was created by removing the ligand-coordinated Mg^{2+} ion from the crystal structure.

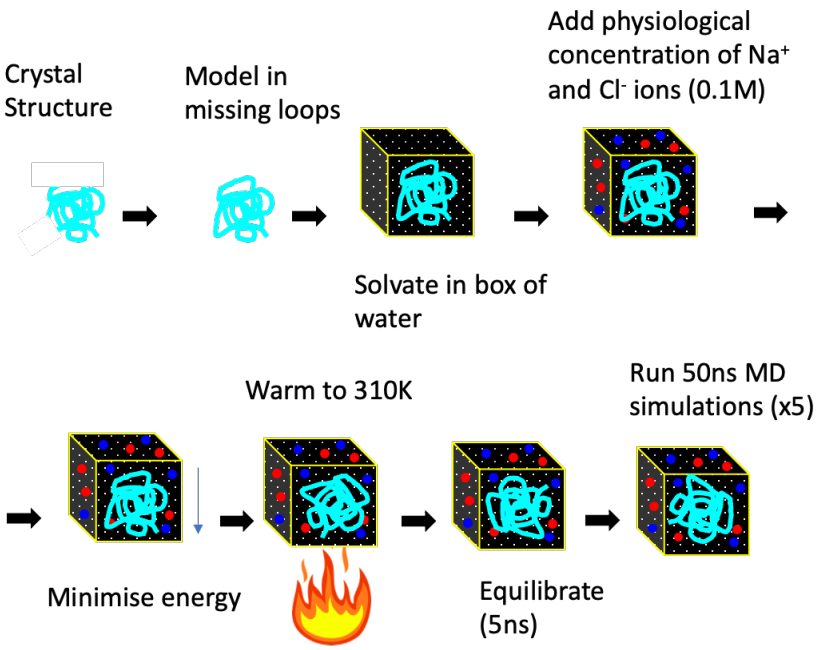

Figure 49 Schematic overview of molecular dynamics simulation set up

Parameterisation of the protein and DNA was done using the AMBER99sb-ildn forcefield and parameters for moxifloxacin were calculated using Acpype[366]. To facilitate the processing of the covalent bond between Tyrosine 129 and the phosphate backbone of DNA by GROMACS, two modified amino acids (TYX and TYY) were created. These 'hybrid residues' contained the parameters for Tyrosine, excluding the hydroxyl hydrogen, all nucleotides in the covalently bound DNA chain and the covalent bond between the Tyrosine hydroxyl oxygen and the corresponding DNA backbone phosphorus atom. The PDB file order and residue naming was adjusted to reflect the modified amino acids. Due to the size of the hybrid residues (403 atoms) it was not possible to redistribute the partial charges in a principled way using QM/MM and we instead choose to simply retain the partial charges of DNA and Tyr as found in the forcefield. Due to the 'loss' of the hydrogen atom from the tyrosine the system was left with a non-integer charge, so a solvent chloride ion was modified to provide a balancing charge. This is clearly not optimal because it is

unphysiological, however I felt it preferable to manually redistributing charge as there was no objective means to do this.

Version 2019.1 of the GROMACS software was used for all simulations and their set up. The moxifloxacin-bound structure was placed in a rhombic dodecahedron unit cell defined by three box vectors, a = (13.8 nm, 0 nm, 0 nm), b = (0 nm, 13.8 nm, 0 nm) and c = (6.9 nm, 6.9 nm, 9.8 nm), and bc, ac and ab box vector angles of 60°. This unit cell was solvated with 59,895 water molecules and 175 Na^+ and 112 Cl^- ions which provided electrical neutrality and a 0.1 M salt concentration. Likewise, for the apo system the same sized unit cell was solvated, and ions added to give a neutral charge and 0.1 M salt concentration. For each repeat (N = 5) of each apo and moxifloxacin-bound structure, following a 1,000 step energy minimisation, the system was warmed using a Langevin thermostat from 100 to 310 K for 500 ps whilst a Berendsen barostat kept the pressure at 1 bar. The system was then equilibrated for 5 ns and the pressure was maintained using a Parinello-Rahman barostat. MD simulations were then performed for 50 ns. LINCS[263] was used to constrain the length of all bonds involving a hydrogen, enabling a timestep of 2 fs to be used throughout.

During initial test simulations, the moxifloxacin-coordinated Mg^{2+} ion dissociated from the drug. A harmonic distance restraint was thus implemented to maintain the distance observed between Mg^{2+} and moxifloxacin in the crystal structure (0.209 nm) throughout all simulations and equilibration. The strength of the restraint was set at 100,000 kJ mol^{-1} nm^{-2}, lower values were tested but found to be insufficient.

6.2.2 MD Trajectory Analysis

The GROMACS trjconv tool was used to centre the protein in the box and correct any artefacts introduced by the periodic boundary conditions prior to analysis. The MDAanalysis.analysis.rms.rmsd function was used to calculate the RMSD of backbone protein atoms compared to the first frame of the trajectory after superimposition of the backbone. The VMD 'hbonds analysis' extension was used to identify and calculate the occupancy of any hydrogen bonds formed between moxifloxacin and non-solvent residues, using the default parameters (distance < 3.0 Å and the angle < 20°). Specifically, every 5 frames of each moxifloxacin-bound trajectory was loaded into VMD and hydrogen bonds for each moxifloxacin molecule were estimated separately from the last 25 ns of simulations, to allow for some equilibration. The mean occupancy and standard error of hydrogen bonds was taken from the 5 trajectories.

6.2.3 Free Energy Set Up

PMX[353] was used to mutate amino acids from frames taken at 10 ns intervals from 20 ns onwards of each of the five 50 ns MD simulations, creating the new topology including the mutated state. As DNA gyrase has two *gyrA* and two *gyrB* subunits, each mutation required the mutation of two amino acids.

GROMACS 2019.1 was used to perform equilibrium free energy calculations for each mutation for each apo and drug-bound starting structure using three steps, consistent with best practice[236]. Therefore, for each mutation three different free energies were calculated: ΔG_{qoff} – the change in free energy when the partial electrical charges on the atoms to be

perturbed are removed, ΔG_{vdW} – the change in free energy energy when the vdW parameters are perturbed and ΔG_{qon} – the change in free energy when the partial charges on the phase-in atoms are added back. Simulations were run for 500 ps using 8 equally spaced λ windows between 0 and 1. To help improve the accuracy of results obtained using thermodynamic integration, replica exchange was used; 10,000 replica exchanges were attempted between neighbouring λ-simulations every 1,000 timesteps. For mutations resulting in an increase in atom number, an Alchembed procedure[367] was introduced prior to running the 500 ps calculations; a short (1000-step) slow-growth simulation was run between the wild-type and mutated states in order to create a starting structure better able to prevent clashes arising from the growth of the new side chain. To further mitigate any clashing of sidechains or simulation crashing, for the vdW transitions for all mutations, and the qon transition of D94G, LINCS constraints were removed and the timestep appropriately reduced from 2 fs to 1 fs. To compensate for the removal of negative charge associated with the *gyrA* D94G mutation, a neutral ion approach was used, whereby the positive charge on a solvent sodium ion near the edge of the box was removed in parallel.

Mutations involving amino acids with two chiral centres were checked for correct chirality in the mutated state because on initial testing, results for the *gyrA* S95T transition converged at two different ΔG_{vdW} values (Figure 50a), which is theoretically impossible. Examination of trajectories 1-3 found that for both *gyrA* molecules, the Threonine sidechain was grown in with *S* chirality at the second chiral carbon (the first being attributable to the backbone chiral carbon common to all amino acids). In nature, this carbon of Threonine has R chirality (Figure 50b). Trajectories 4 and 5 correspond to one of the two *gyrA* Threonine 95 residues being grown in with *R* chirality and the other with *S*. Examination of the MD set up revealed that the GROMACS pdb2gmx tool, when renaming the Serine HB3 and HB2 atoms to HB1

and HB2, respectively, in the coordinate file, swapped the coordinates of the two atoms. To transform Serine into Threonine, the PMX software grows in the methyl group at the second chiral carbon and removes HB2 of Serine at this position. Therefore, PMX removed the incorrect hydrogen atom resulting in a Threonine with S chirality (Figure 50b) the majority of the time. To fix this bug, coordinates of HB1 and HB2 were manually swapped in the coordinate file before re-running the free energy set up and calculations.

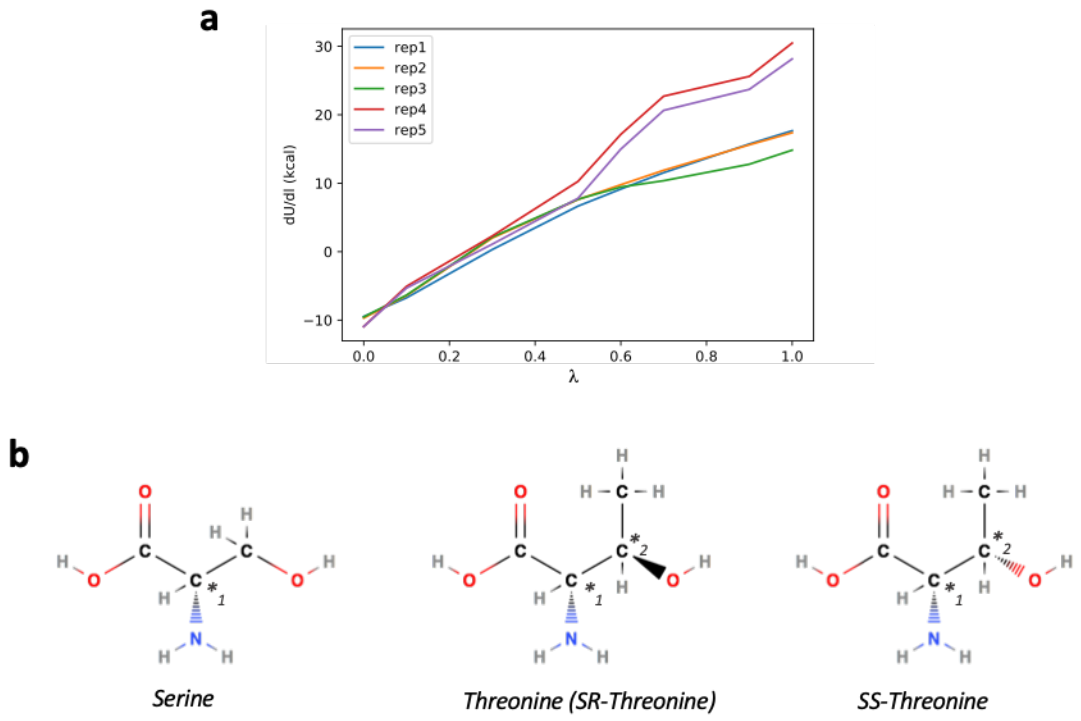

Figure 50 van der Waals transition of Serine to Threonine. a) ΔG_{vdW} for *gyrA* S95T mutation in five independent free energy calculations. b) structural formulae of Serine, naturally occurring Threonine with R chirality at the second chiral carbon, and Threonine with S chirality at the second chiral carbon.

Files and scripts necessary to reproduce the above steps, starting with the Alchembed step, for the *gyrA* A90V mutation can be found here: https://github.com/fowler-lab/tb-rbfe-setup.

6.2.4 Free Energy Calculation

The first 250 ps of each free energy simulation was discarded to minimise equilibration effects. Free energies from individual steps were then calculated using thermodynamic integration, applying the trapezium rule to calculate the area under the curve (see Section 2.3.2). To calculate the total difference in moxifloxacin binding free energy between wild type and mutated DNA gyrase ($\Delta\Delta G_{binding}$), the total free energy difference of the alchemical transition of the mutation in the apo state DNA gyrase was subtracted from that of the moxifloxacin-bound protein, as shown in the free energy cycle (Figure 51).

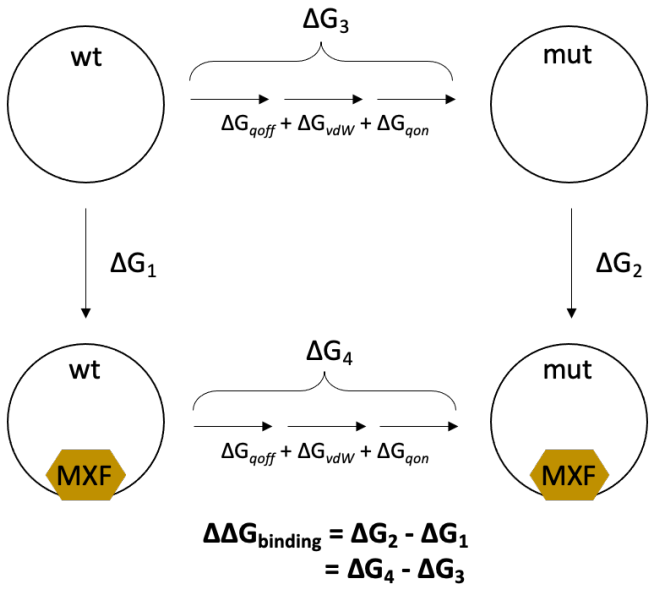

Figure 51 Free energy cycle of moxifloxacin (MXF) binding the DNA gyrase cleavage complex. The subscripts qoff, vdW and qon describe the process of first removing the electrical charge from atoms being perturbed, followed by transforming their van der Waals parameter, before finally recharging the atoms being perturbed. Free energy is a state function, and therefore the difference in binding free energy ($\Delta\Delta G_{binding}$) is a sum of the alchemical free energies (e.g. $\Delta G_4 - \Delta G_3$).

The standard error of the mean (SEM) was calculated for each step (e.g. ΔG_{vdW}), with the final error in ΔΔG estimated by adding that from all steps in quadrature. To estimate 95% confidence limits, the standard errors from each step were multiplied by the appropriate *t*-

statistic. As repeating simulations to reduce statistical error is increasingly computationally expensive, four independent free energies were initially calculated for each step as is in best practice[236]. Then, to reduce the magnitude of the estimated error in an efficient manner, additional targeted repeats were run with the aim of reducing the magnitude of the total statistical error to less than 1 kcal mol^{-1}, although this was not always possible even after running 10 or more repeat calculations for some steps (Table 29). The individual free energies for each step can be found in Appendix Table 9.

MUTATION	LEG	STEP	N	SEM
S95T	apo (ΔG3)	vdW	15	1.04
		qoff	5	0.16
		qon	5	0.16
	drug-bound (ΔG4)	vdW	15	1.06
		qoff	5	0.72
		qon	5	0.72
A90V	apo (ΔG3)	vdW	10	0.67
		qoff	5	0.07
		qon	4	0.96
	drug-bound (ΔG4)	vdW	15	0.81
		qoff	5	0.08
		qon	5	0.75
A90S	apo (ΔG3)	vdW	5	0.24
		qoff	5	0.04
		qon	5	0.79
	drug-bound (ΔG4)	vdW	5	0.73
		qoff	5	0.09
		qon	8	0.87
E501D*	apo (ΔG3)	vdW	10	1.55
		qoff	10	1.78
		qon	10	1.36
	drug-bound (ΔG4)	vdW	10	1.23
		qoff	10	0.94
		qon	9	0.88
D94G	apo (ΔG3)	vdW	4	2.73
		qoff	8	2.85
		qon	8	6.03
	drug-bound (ΔG4)	vdW	5	2.76
		qoff	10	3.06
		qon	10	3.11

Table 29 Number of repeat simulations run for alchemical transitions of DNA gyrase **mutations in apo and drug-bound systems.** * indicates a *gyrB* mutation.

As the choice to discard 250 ps from the start of each simulation was made arbitrarily, I investigated whether using different cut offs changed the results for one of the simple susceptible mutations, *gyrA* S95T, and the most complex mutation, *gyrA* D94G. On discarding 125 ps and 375 ps, compared to 250 ps, there was no difference in the final calculated ΔΔG measurement for *gyrA* S95T or *gyrA* D94G (Figure 52a,b).

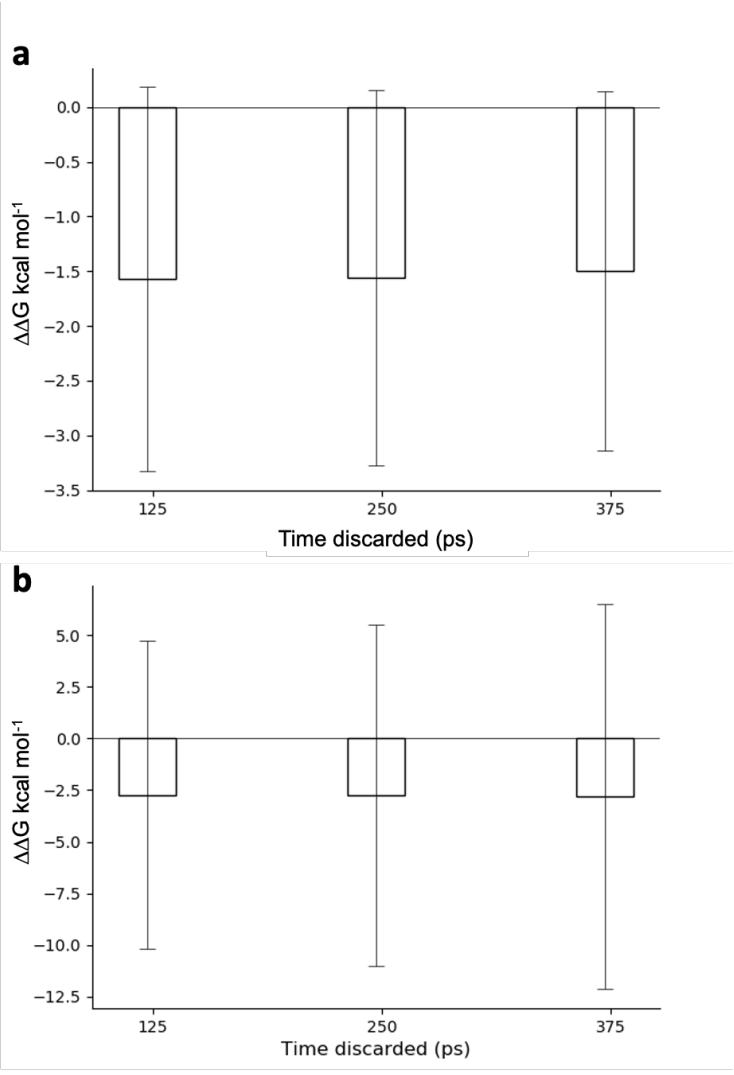

Figure 52 Effect on RFBE predictions when discarding different amounts from the start of component FE simulations for a) *gyrA* S95T and b) *gyrA* D94G

6.2.5 Calculation of the ΔΔG ECOFF equivalent and expected ΔΔG of resistance conferring mutations

To estimate the mean Minimum Inhibitory Concentration (MIC) of resistance conferring mutations, I took the geometric mean of MICs of isolates in the CRyPTIC dataset containing only that mutation above a background of any lineage associated mutations in the *gyrA* or *gyrB* genes. I am therefore assuming that any shift in MIC compared to the wild-type population is due solely to that mutation. To estimate the ΔΔG value equivalent to an MIC value (such as the epidemiological cut off value (ECOFF) used to define clinical resistance, or the mean MIC of samples containing a resistance-conferring mutation), I used the following approximation set out by Fowler *et al*.[354]:

$$\Delta\Delta G = \ln(v / w) \times kT$$

where v is the MIC value to convert (e.g. ECOFF, geometric mean), w is the geometric mean MIC of a genotypically wild type *M. tuberculosis* population (calculated using moxifloxacin MICs of isolates that the CRyPTIC Consortium had identified as genotypically wild-type), k is Boltzmann's constant and T is 310 K.

6.3 Results

6.3.1 Molecular dynamics simulations

The root mean squared deviation (RMSD) of the protein backbone remains at between 2 and 4 Å during both the apo and moxifloxacin bound structure simulations and the plots begin to level off as the simulation time increases (Figure 53a,b). This low RMSD indicates that the gyrase structure is relatively stable, especially when considering that the complex is ~200 kDa. The plots suggest that several different confirmations have been sampled as there are higher

and lower RMSD events during the timeframe, but that complete convergence is unlikely as the RMSD is not plateaued in all cases; there are likely events occurring over a longer timeframe than that which I have simulated.

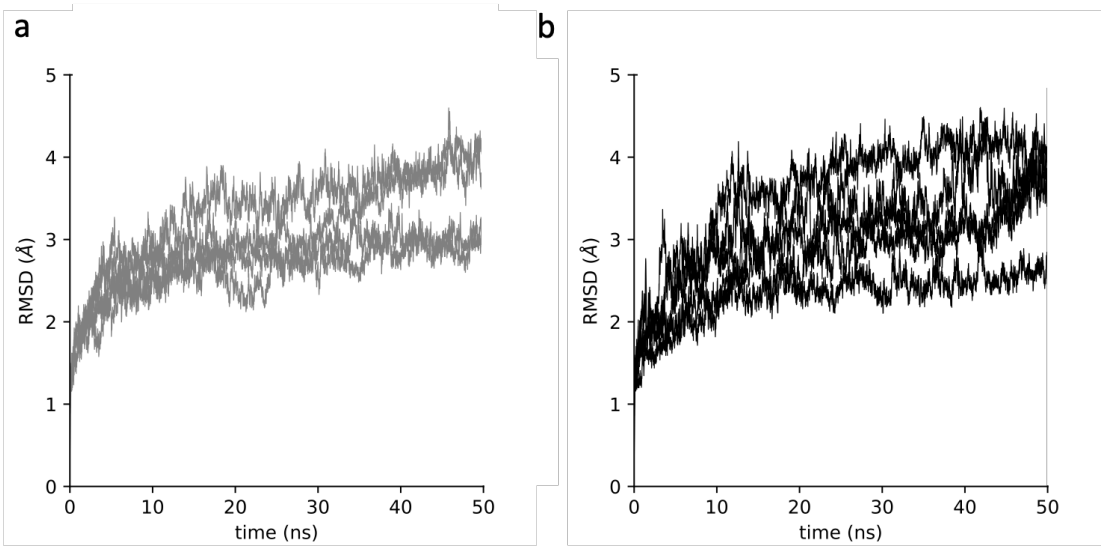

Figure 53 RMSD of protein backbone atoms of (a) apo and (b) moxifloxacin-bound DNA gyrase cleavage complex protein backbone during 5x 50ns MD simulations

The only hydrogen bonding between the fluoroquinolones and the DNA gyrase in the crystal structure is mediated by the water bridge co-ordinated by *gyrA* D94G. Since MD allows a dynamic study of the system it could suggest the presence of intermittent direct interactions such as hydrogen bonds. Analysis of inferred hydrogen bonds between the fluoroquinolones and the gyrase from the second half of the equilibration simulations suggest that a hydrogen bond forms occasionally between *gyrA* Arginine 128 and a carboxylic oxygen of moxifloxacin (Table 30, Figure 54).

HBOND (DONOR-ACCEPTOR)	MEAN OCCUPANCY (%)
ARG128-SIDE-MFX677-SIDE	21.6 ± 6.2
MFX677-SIDE-TYY129-SIDE	0.5 ± 0.3

Table 30 Hydrogen bonds formed between moxifloxacin (residue 677 in the crystal structure) and DNA gyrase cleavage complex residues and their mean occupancy from 5 independent MD simulations. Mean occupancy is calculated from occupancy over the last 25ns of MD simulation and standard error of the mean is shown. SIDE indicates atoms from the side chain of residues are involved in the interaction. The side chain of TYY129 includes all nucleotides in the covalently bound DNA strand.

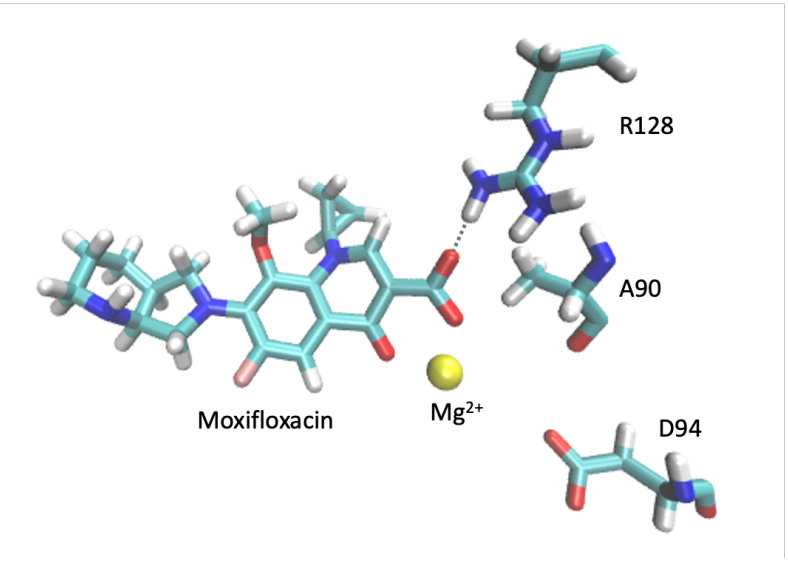

Figure 54 Hydrogen bond formed between *gyrA* Arginine 128 and moxifloxacin (residue number 677 in PDB:5BS8[184]) during molecular dynamics simulation

There is a difference between the two fluoroquinolone binding positions whereby the *gyrA* Arginine 128 residue adjacent to one moxifloxacin molecule forms these interactions but the other does not (Table 31). A few other hydrogen bonds are inferred, but these are at lower than 10% occupancy and typically only seen in one of the five simulations.

HBOND (DONOR-ACCEPTOR)	MEAN OCCUPANCY (%)
ARG482-SIDE-MFX676-SIDE	5.8 ± 4.7
TYY129-SIDE-MFX676-SIDE	0.1 ± 0.1
ARG128-SIDE-MFX676-SIDE	0.1 ± 0.1

Table 31 Hydrogen bonds formed between moxifloxacin (residue 676 in the crystal structure) and DNA gyrase cleavage complex residues and their mean occupancy from 5 independent MD simulations. Mean occupancy is calculated from occupancy over the last 25ns of each MD simulation and standard error of the mean is shown. SIDE indicates atoms from the side chain of residues are involved in the interaction. The side chain of TYY129 includes all nucleotides in the covalently bound DNA strand.

6.3.2 Free energy calculations

A positive value of the change in binding free energy of the antibiotic (ΔΔG > 0) indicates that the antibiotic binds less well to the target following the mutation and therefore would be predicted to confer resistance to that drug. Clinically, however, a sample is categorized as 'resistant' if its minimum inhibitory concentration (MIC) is greater than a critical concentration, often the ECOFF, which is defined as the MIC of the 99th percentile of a collection of phenotypically-wildtype samples. The threshold for moxifloxacin resistance (1.2 kcal mol^{-1}) was derived using published ECOFF values[272] as described in Section 6.2.5.

Four independent values of ΔΔG were first calculated. Each value of ΔΔG required the calculation of 6 alchemical free energies (Figure 51). Repeats of the alchemical free energy components exhibiting the greatest variation were then run to efficiently reduce the confidence limits of the prediction as described in Section 6.2.4. First let us consider the overall values of ΔΔG and whether successful predictions can be made.

Both negative controls (S95T and A90S) were correctly predicted to not affect the binding of moxifloxacin to the DNA gyrase (Figure 55, Table 32). Although hyper-susceptibility is expected for A90S, the magnitude of the confidence limits are too large to confirm or refute this. No definite prediction could be made for any of the three mutations associated with moxifloxacin resistance since the confidence limits of all three mutations straddled the clinical threshold. However, the A90V mutation does significantly decrease the binding affinity for moxifloxacin, as the ΔΔG is positive. The mutations that involved charged residues (E501D and D94G) had the largest estimated errors, and the magnitude of these errors prevent any conclusions being drawn.

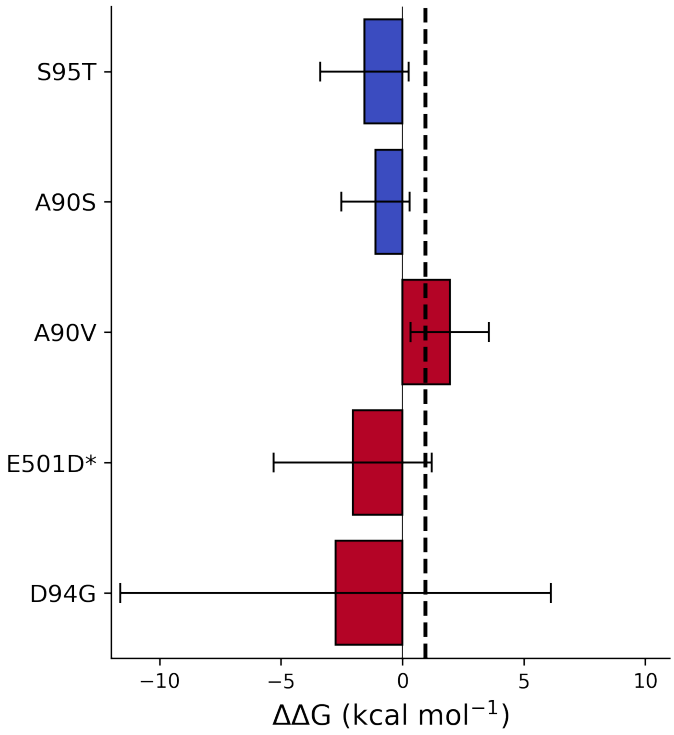

Figure 55 The calculated effect of the listed mutations on the binding free energy of moxifloxacin to DNA gyrase. Dotted lines represent the value of ΔΔG equivalent to the epidemiological cutoff value for moxifloxacin; above this value an *M. tuberculosis* isolate would be considered clinically resistant. Bars represent the mean ΔΔG for each susceptible (blue) and resistant (red) mutation compared to the wild-type protein and 95% confidence limits are shown, calculated using the appropriate t-statistic. An asterisk (*) indicates a mutation in gyrB.

MUTATION	EXPECTATION	ΔΔG (KCAL MOL^{-1})	N	PREDICTION
S95T	Susceptible	-1.6 ± 1.8	50	Susceptible
A90S	Hyper-susceptible	-1.1 ± 1.4	33	Susceptible
A90V	Resistant	2.0 ± 1.6	44	Uncertain
E501D*	Resistant	-2.0 ± 3.2	59	Uncertain
D94G	Resistant	-2.8 ± 8.9	45	Uncertain

Table 32 Summary of free energy calculations for DNAG mutations. N is the total number of free energy calculations used to calculate the ΔΔG. * indicates a mutation in *gyrB*.

To see how the ΔΔG values compared with clinical resistance measurements, the 'expected ΔΔG' corresponding to the geometric mean of MICs associated with each of the resistance conferring mutations, was calculated using previously described methods[354]. However, the errors in both the 'expected ΔΔG' and the ΔΔG values calculated by RBFE were too large to draw any conclusions about how well the values compare with one another (Figure 56).

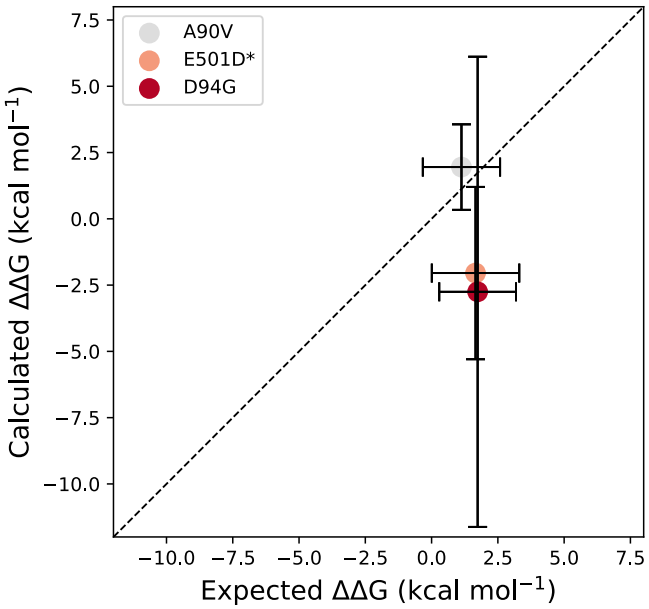

Figure 56 RBFE calculated mean ΔΔG measurements of moxifloxacin resistance conferring mutations compared to the expected ΔΔG measurement. Expected ΔΔG measurements for each mutation were calculated from the geometric mean minimum inhibitory concentration (MIC) of a population of isolates containing each resistance conferring mutation and no other *gyrA*/*gyrB* mutation in an otherwise genetically wild-type background, using previously described methods[354]. Error bars represent 95% confidence interval.

6.3.3 Investigation of sources of error

To further examine what is driving the magnitudes and confidence limits of the individual $\Delta\Delta G$ values in Figure 55, I analysed the alchemical free energy components from the de-charging (ΔG_{qoff}), van der Waals (ΔG_{vdW}) and re-charging (ΔG_{qon}) transitions (Figure 51) for both apo and drug-bound legs of the free energy calculations (Figure 57). As expected, there were no significant differences for the susceptible mutations between the mean apo and drug-bound values of ΔG_{qoff}, ΔG_{vdW} and ΔG_{qon} and the estimated error is generally low.

Differences in ΔG_{vdW} between the apo- and complexed DNA gyrase appear mainly responsible for the positive value of $\Delta\Delta G$ for A90V. Since the mutation involves the introduction of a larger sidechain that is oriented towards moxifloxacin, this is consistent with decreased binding affinity arising primarily through steric hindrance of the moxifloxacin binding site.

For E501D and D94G, all three transitions contribute significant error (Figure 57), which since they add in quadrature, leads to a large overall error in $\Delta\Delta G$. This is not surprising since both mutations involve turning off (and on) electrical charge and D94G involves a net charge change that must be compensated for elsewhere in the system. For D94G, the error arising from the ΔG_{qoff}, and ΔG_{qon} is larger than that of the ΔG_{vdW} despite efforts to minimize the overall error by running more repeats for those transitions (Table 29) to reduce their individual estimated errors.

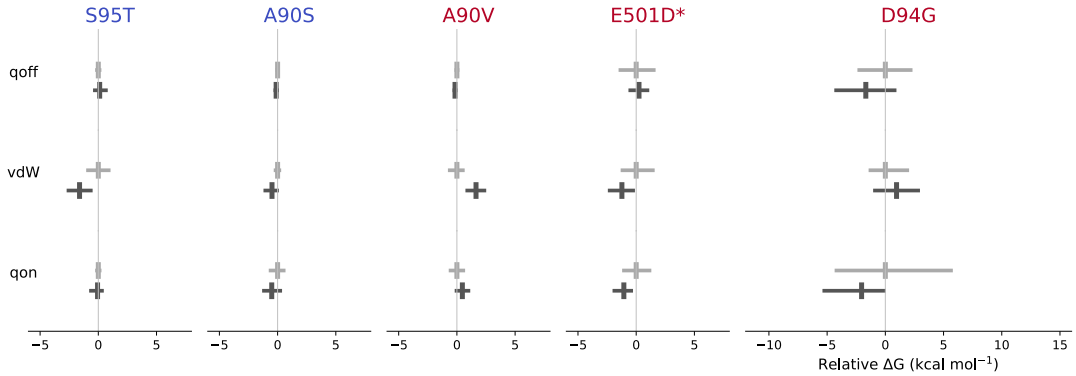

Figure 57 Apo (light grey) and drug-bound (dark grey) free energy calculations for DNA gyrase mutations for de-charging (qoff), van der Waals (vdW), and re-charging (qon) transitions. All results are normalized to the mean of the calculations for the apo leg for each transition for each mutation. Mean values are denoted by a cross and the error bars describe the 95% confidence limits, calculated from the SEM using the appropriate t-statistic. An asterisk (*) indicates a mutation.

By starting each simulation from a different structural seed and discarding the first half of the alchemical free energy simulations and then applying statistics to the resulting values of ΔG, it is assumed that they are independent. If true, then one would also expect the values to be normally distributed; this would appear to be the case for most sets of ΔG values (Figure 58). Applying the Shapiro-Wilks test of normality to the data confirms that, despite the small numbers of samples in some cases, the majority of ΔG values are indeed normally distributed with the exceptions of the ΔG_{vdW} for the apo leg of S95T and ΔG_{qon} for both the apo and drug bound leg of D94G.

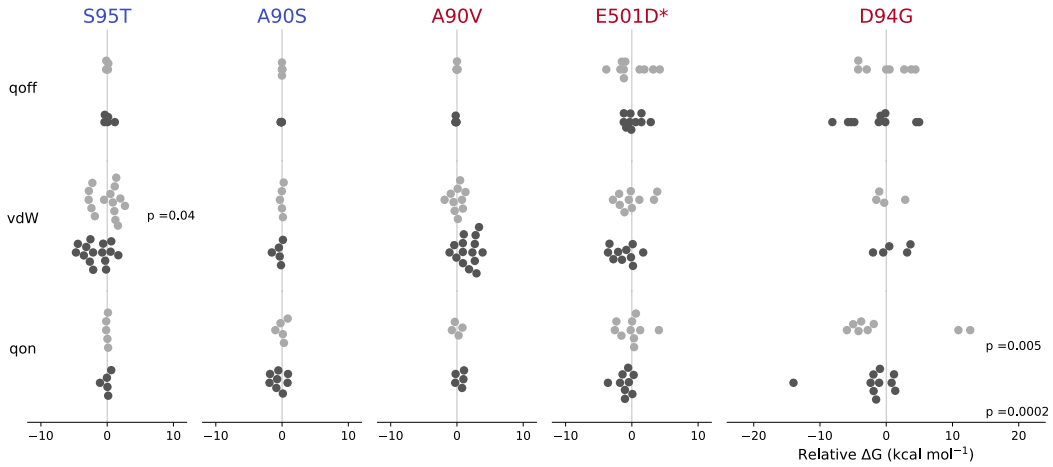

Figure 58 Swarm plots of individual results from apo (light grey) and drug bound (dark grey) alchemical free energy calculations for mutations in the DNA gyrase. All results are normalized to the mean of the calculations for the apo leg for each qoff, vdW or qon transition for each mutation. p-values from Shapiro Wilks test are displayed for each transition showing evidence of non-normality in the repeated calculations, transitions where no p-value is shown indicates there was no evidence of non-normality in the data ($p > 0.05$). An asterisk (*) indicates a mutation.

To test how far the simulations are from normality, four apo and five drug-bound simulations underlying the most variable component of the most complex mutation, D94G (Figure 57), were extended by an order of magnitude (from 0.5 ns to 5 ns). As assessed by the Shapiro-Wilks test, the resulting distributions of apo- and drug-bound free energies no longer showed evidence of non-normality after 5 ns of simulation ($p = 0.92$ and $p = 0.16$) but the distribution of results for the repeated calculations, and therefore the error, remains large (Figure 59).

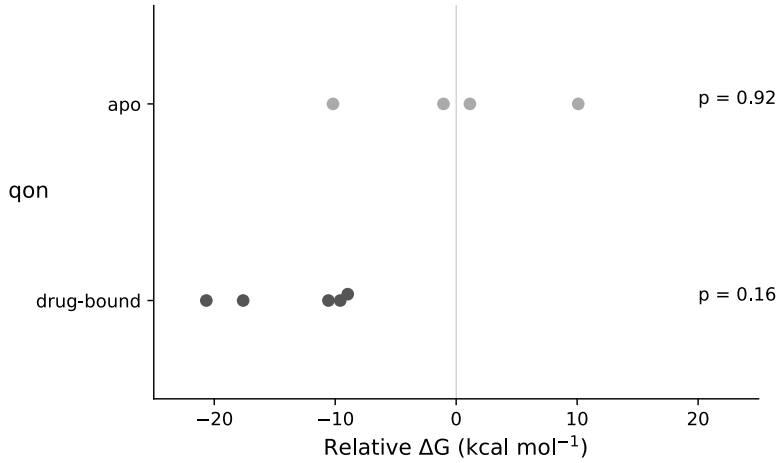

Figure 59 Swarm plots of individual results from apo and drug-bound 5 ns alchemical free energy calculations for the qon transition of DNA gyrase *gyrA* D94G mutation. Results are normalised to the mean of the calculations for the apo leg. p values from Shapiro Wilks test are displayed.

As I am using thermodynamic integration with the trapezium rule to calculate free energies, the accuracy, bias and error introduced is dependent in part on the degree of curvature in the component free energy calculations[236, 368]. More complex transitions, i.e. those with a greater difference between starting and end states, are more likely to have increased curvature especially at λ values close to the extremes (0 and 1). I therefore investigated the curvature in individual qoff, vdW and qon free energy components for a simple susceptible mutation, *gyrA* S95T, which involves the growing in of a methyl-group and no change in charge, and the most complex mutation, *gyrA* D94G, which involves a significant reduction in atom number and removal of a full negative charge (Figure 60a-b). As expected, for *gyrA* S95T, there was no curvature observed for either qon and qoff transitions as there is no major charge change. The vdW transition for both apo and drug-bound systems, where the key change occurs in growing in a methyl group, do show some gentle curvature through mid-range values of λ (Figure 60a), and the error for this transition was larger than for qoff or qon transitions (Figure 57). In contrast, for both apo and drug-bound legs of the D94G mutation, there is a high degree of curvature towards λ=1 in the vdW transition where the

large side chain of Aspartate is removed (Figure 60b). Unexpectedly there was little curvature observed in the qoff transition, where the main negative charge of the Aspartate side chain is removed, and the transition contributed a large amount of error (Figure 57). In general, the individual <dU/dλ> values are much larger in magnitude for the *gyrA* D94G mutation than S95T, which is not unexpected due to the magnitude of change, but may explain why there is larger error associated with the *gyrA* D94G prediction (Figure 55).

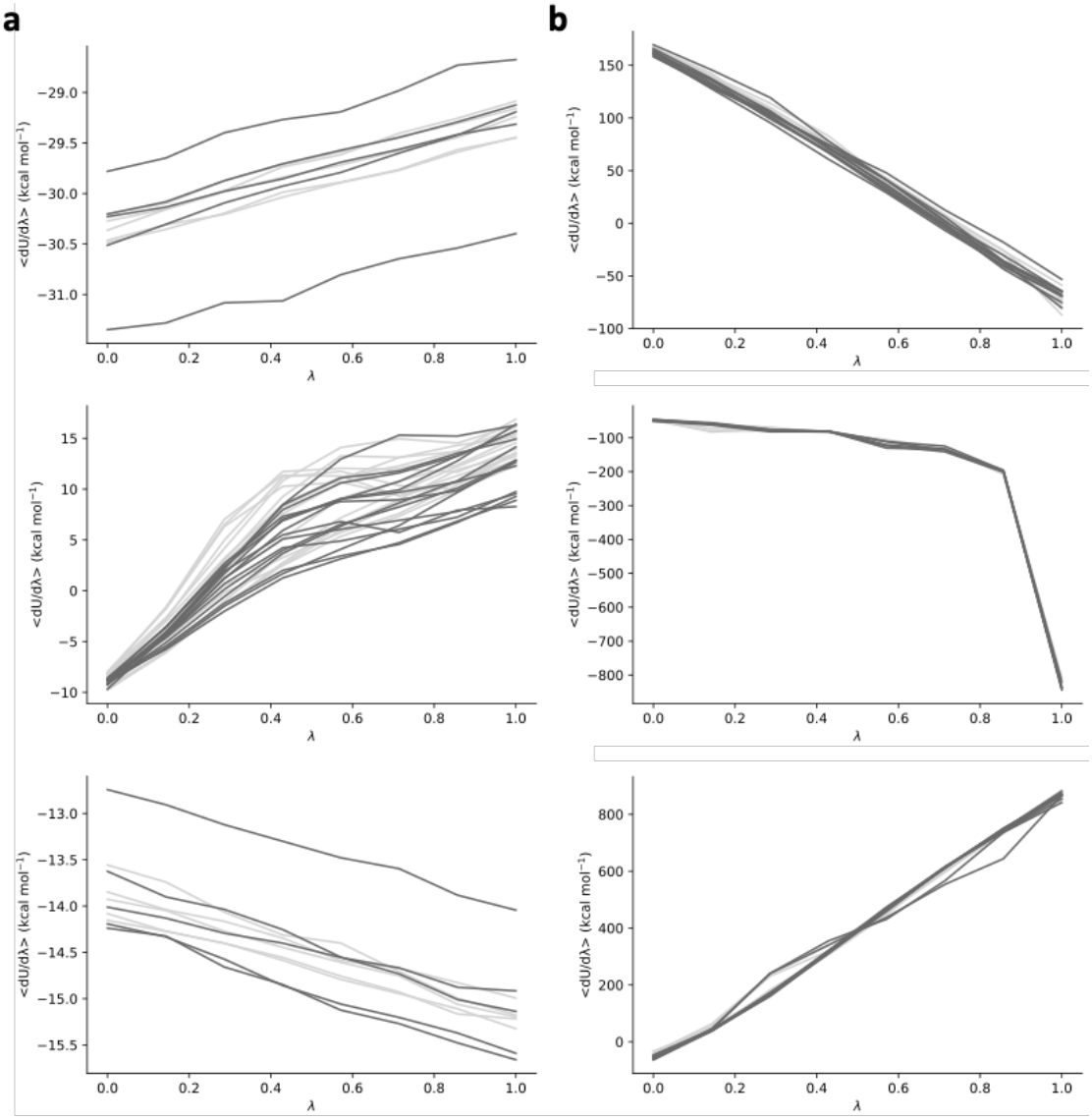

Figure 60 Curvature in qoff, vdW and qon $\lambda_{0\to1}$ free energy calculations for (a) *gyrA* S95T and (b) *gyrA* D94G. Apo results are shown in light grey and drug-bound results in dark grey.

6.4 Discussion

As exemplified by previous studies, short RBFE calculations can make correct resistant and susceptible predictions[354, 361]. However, on the large, complex DNA gyrase system, the combination of small fold increases in MIC and significant changes in properties associated with resistance conferring mutations meant that the estimated error of $\Delta\Delta G$ was too large for definite predictions to be made in several instances (Figure 55). It is not wholly unexpected that we saw larger errors than in other RBFE studies[349-351] as the DNA gyrase system is very large and complex, and in addition, each genetic mutation results in two mutations in the protein complex, which both contribute error.

There was more error associated with the mutations that involved charged residues, *gyrA* D94G and *gyrB* E501D than those that did not, and substantial error arose from the qoff and qon transitions (Figure 55, Figure 57). In order to reduce the estimated statistical error from these transitions to half the current value, at least 4x the number of simulations would need to be run (assuming the simulations are independent). The introduction of more, targeted, λ windows to reduce bias and error from regions of transitions with high curvature in calculations for complex mutations, such as *gyrA* D94G, could also help reduce error, however the transitions with the greatest curvature were not necessarily the most variable (Figure 58, Figure 60b).

To calculate statistical significance, it is assumed that conformations used to seed each calculation are independent of one another or that the λ simulations are long enough to allow the initial state to be 'forgotten'. Given the use of very short λ simulations the latter is

almost certainly not true and the observed non-normality of the ΔG_{qon} free energy for D94G indicates that these values are not always independent (Figure 58). To solve the non-normality problem, either the λ simulations would have to be extended by at least an order of magnitude or the equilibration simulations would need to be more numerous as well as longer. It is not unreasonable to suggest that the equilibration simulations should be longer as the structures contain DNA, for which equilibration in the order of μs has been suggested[369].

If the magnitude of the error was sufficiently reduced, it is still probable that predictions will be influenced by inaccuracies elsewhere in the system: there remains limitations in that forcefield parameters for ions and small molecules are not always accurate[370] and indeed in this instance they are unlikely to be so as harmonic distance restraints were required between moxifloxacin and its coordinated Magnesium ion. Further, accurate predictions for charge changing mutations such as *gyrA* D94G may still prove challenging, as unwanted electrostatic artifacts may prevail in the system despite keeping the total system net neutral[371]. There are however several mathematical correction options to account for these effects which could be considered[371-373]. Finally, by using the RBFE method to predict resistance, one assumes that the mutation causes resistance by disrupting antibiotic binding and this may not always be the case – for example a *gyrB* A642P mutation is associated with increased resistance to fluoroquinolones[203], yet the mutation is located at the exterior of the protein, far from the drug-binding site. In future, isothermal titration calorimetry data could help confirm whether the mutations cause a decrease in binding affinity for the drug and would provide a better quantitative benchmark for RBFE results than MIC data.

As discussed elsewhere in this thesis (see Section 5.4), it may be more useful to be able to predict susceptibility accurately and reproducibly than resistance since this can allow a patient to immediately start treatment on an appropriate regimen. Although both susceptible mutations were correctly predicted in this instance, the scale of estimated errors suggest it may be more difficult to predict susceptibility than resistance using RBFE. Let us assume that most susceptible mutations will not affect the binding affinity for the antibiotic, and hence would have a $\Delta\Delta G$ of zero. The predicted $\Delta\Delta G$ of such mutations would therefore require the estimated error to be at least less than the value of the clinical cut off used (for DNA gyrase 0.9 kcal mol^{-1}) to make a confident susceptible prediction. The magnitude of error for the mutations in this study was greater than the ECOFF in all cases (Table 32). It is likely that the error could be reduced by running a greater number of repeats, however some mutations can result in a small increase in MIC but not enough to confer resistance (as susceptibility can be defined as any MIC up to the ECOFF), and in such cases even a lower level of error (± 0.5 kcal mol^{-1}) may still prove insufficient for prediction.

Despite the limitations in making resistance and susceptibility predictions using RBFE, the MD simulations generated are useful for probing system dynamics which can direct hypotheses to explain *in vitro* behaviours or improve drug efficacy. Analysis of static crystal structure suggests no gyrase amino acids make specific contact with fluoroquinolones, aside from the water bridge network coordinated by *gyrA* $D94^{184}$, but my study suggests a hydrogen bond could form between *gyrA* R128 and a carboxylic oxygen of moxifloxacin (Table 30, Figure 54). Interestingly, this interaction was only observed at one fluoroquinolone binding site. The only difference between the two binding sites is the DNA sequence at each position, suggesting that the different nucleotides may cause the differential hydrogen bond networks seen at the two sites. Indeed, this could rationalise why

fluoroquinolone-mediated cleavage has some DNA sequence preference in other bacterial species[374, 375]. If the hydrogen bonding interaction is genuine, this, together with information about DNA sequence preference, could provide direction for the design of more effective fluoroquinolone antibiotics, particularly as *gyrA* R128 is essential for catalysis and therefore well conserved[341, 342].

I conclude that the RBFE approach is not yet suitable for making a rapid fluoroquinolone resistant or susceptible prediction for DNA gyrase mutations in Tuberculosis. This is principally due to the large errors and inability to make predictions with sufficient confidence. Implementation of the solutions suggested to reduce error and ensure independence in this discussion using the software and compute that were employed for this study (see Section 6.2) would be prohibitively computationally expensive. However, yearly software updates with increased performance and use of modern Graphics Processing Units (GPUs) and cloud computing could offer improved speed and efficiency to facilitate further study[359, 376]. Hopefully, continued improvements software, compute, forcefields and correction schemes could make RBFE calculation a viable option for studying large and complex biochemical systems such as this in the future.

7 Summary and conclusions

The aim of this thesis was to build upon our knowledge of fluoroquinolone resistance and test new methods to predict fluoroquinolone resistance from WGS data to ultimately help improve the performance of sequencing-based methods for fluoroquinolone DST. In this chapter I will summarise the main findings of this thesis, their significance, and implications for the future of fluoroquinolone DST and further research. I will then emphasise the overarching strengths and limitations of my work, before concluding.

7.1 Main findings and their implications

7.1.1 There is a high level of fluoroquinolone resistance

Analysis of the CRyPTIC dataset revealed that 38.8% of RR/MDR isolates collected were pre-XDR (Figure 24) and of the true MDR isolates over 40% had fluoroquinolone resistance (Figure 29). This is far higher than the WHO estimate of 20%. It is acknowledged that the proportion of fluoroquinolone resistant MDR isolates in this dataset is artificially high due to the large number of samples contributed by India (Figure 25), but the removal of this site, and others associated with a high prevalence of fluoroquinolone resistance still resulted in a figure of 33.0% of RR/MDR isolates having resistance to a fluoroquinolone (see Section 3.3.3). This level of resistance is particularly concerning because fluoroquinolones are recommended for most MDR treatment regimens and are one of the safest and most effective drug choices (Table 4, Table 2).

The potential shortfall in the WHO estimate of fluoroquinolone resistance prevalence further highlights the acknowledged lack of fluoroquinolone resistance surveillance and DST[110]. This, combined with the likely increased use of fluoroquinolones for TB treatment as the use of second-line injectables is phased out and new drug-susceptible regimens are adopted, means that it is imperative to increase DST for fluoroquinolones. The high proportion of MDR TB samples with fluoroquinolone resistance, indicates that fluoroquinolone DST should be performed at the same time as testing for MDR.

Additionally, if levels of fluoroquinolone resistance increase further, as is probable, we will see higher levels of pre-XDR TB, for which the BPaL regimen is recommended, but may have limited efficacy[109]. We may also see an increase in XDR cases, which are difficult to treat, as the NRDs bedaquiline and linezolid become more widely used without an effective fluoroquinolone (Table 4). As there are currently no PCR based diagnostic tests for the NRDs, and we have limited knowledge of the genetic determinants of resistance thus far, time-consuming culture-based DST will be required to diagnose these cases.

7.1.2 Fluoroquinolone resistance can arise prior to first line treatment

Concerningly, the dataset shows that fluoroquinolone resistance can emerge *prior* to resistance to first line treatments and therefore, one assumes, before the TB was treated (Figure 29). Fluoroquinolone drugs are used to treat a range of other infections and prior fluoroquinolone treatment is associated with a three times greater likelihood of having fluoroquinolone resistant TB[190]. This suggests acquisition of resistance can occur during latent infection, and if true, this could be more likely for fluoroquinolone resistance than for

other drug resistance because the most common resistance conferring mutations may not have a fitness cost[324] (Figure 38a-d) and therefore would not be outcompeted by any remaining wild type population once fluoroquinolone treatment stops.

It is imperative that the availability and prescription of fluoroquinolones for any disease is more carefully considered, especially when one considers the estimated prevalence of latent TB is around 25% of the global population[9]. Fluoroquinolones are among the safest and most effective drugs for TB treatment (Table 2) and their stewardship is therefore vital for good treatment outcomes. This finding also has implications for DST, in that if resistance emerges prior to MDR, fluoroquinolone DST should be considered at the same time as rifampicin DST.

7.1.3 Moxifloxacin and levofloxacin are different

Throughout this thesis we have found several differences between the fluoroquinolones, levofloxacin and moxifloxacin. Firstly, there were differences in the level of resistance to each of the drugs:

- Isolates in the CRyPTIC dataset were more likely to have resistance to levofloxacin than moxifloxacin in all backgrounds tested (Figure 27a-d).
- There are several examples of isolates that were resistant to moxifloxacin and not levofloxacin and vice versa (Figure 23).

Although there was no significant difference seen in the prevalence of different non-synonymous SNPs in levofloxacin and moxifloxacin resistant isolates in the largest background group (MDR, Lineage 2 and originating from India) (Table 11), there were

different associations between the level of moxifloxacin and levofloxacin resistance conferred by the most common resistance conferring mutations:

- Lineage 3 isolates with *gyrA* D94G or *gyrA* A90V were associated with lower moxifloxacin MICs compared to Lineage 2 isolates (Table 16, Table 18), but there was no significant difference in levofloxacin MIC for isolates with *gyrA* D94G or *gyrA* A90V in different lineage backgrounds (Table 15, Table 17).

Finally, the performance of resistance prediction methods was different for levofloxacin and moxifloxacin:

- The detection of moxifloxacin resistance in the CRyPTIC dataset using the WHO catalogue and the mutations detected by molecular diagnostic tests was more sensitive for moxifloxacin than for levofloxacin (Figure 36) but the specificity was generally higher for levofloxacin resistance prediction (Figure 36).
- For binary predictions made using machine learning models there was relatively little difference in sensitivity for the best performing models (96.8% vs 97.5%) but the specificity was again higher when predicting levofloxacin resistance (Figure 43a,b).
- For machine learning models predicting MIC, the accuracy and essential agreement for levofloxacin MIC prediction was higher than for moxifloxacin (Figure 45a,b, Figure 46a,b).

These observed differences suggest that the drugs should be treated separately for resistance prediction, yet the fluoroquinolones are often grouped together when making catalogue-based or machine learning predcitions[344]. A prediction of simply 'fluoroquinolone

resistant' may result in the proportion of patients that would be susceptible to one of the two fluoroquinolones (Figure 23) being denied treatment with the most safe and effective choice (Table 2). The difference in the performance of the DST or predictive tool used for each of the drugs should be considered when choosing a treatment option; it would be preferable to use the drug for which a more reliable resistance prediction could be made, to increase the likelihood of appropriate treatment and to achieve more reflective resistance surveillance. However, I acknowledge that other considerations such as cost and local availability, and safety and effectiveness (see Section 1.4.1), will likely remain the most important factors when selecting the most appropriate fluoroquinolone for TB treatment.

7.1.4 There are complex associations with fluoroquinolone resistance

I found that some mutations are associated with lineage, country of origin and phenotypic resistance backgrounds and that the level of resistance to levofloxacin and moxifloxacin is also associated with several different factors. Specifically:

- Fluoroquinolone resistant isolates of Lineage 1 and Lineage 4 were less likely to have a *gyrA* D94G mutation than Lineage 2 isolates and Lineage 4 isolates were more likely to have a *gyrA* A90V mutation than Lineage 2 isolates (Table 13, Table 14).
- Isolates from China were less likely to contain the most common resistance conferring mutation *gyrA* D94G than isolates from India (Table 13), but the levofloxacin and moxifloxacin MIC of isolates with a *gyrA* D94G mutation was higher in China than in India (Table 15, Table 16).
- Lineage 3 isolates with either a *gyrA* D94G or *gyrA* A90V mutation were associated with lower moxifloxacin MICs compared to Lineage 2 isolates (Table 16, Table 18)

- The moxifloxacin and levofloxacin MIC of isolates containing *gyrA* D94G were negatively associated with an isoniazid and rifampicin susceptible background and an isoniazid mono-resistant background compared to a rifampicin resistant/MDR background (Table 15, Table 16).
- The moxifloxacin and levofloxacin MIC of isolates containing *gyrA* A90V were negatively associated with an isoniazid and rifampicin susceptible background compared to a rifampicin resistant/MDR background (Table 17, Table 18).
- Double resistance conferring mutations, although rare, confer a higher level of resistance than either mutation individually (Table 12, Figure 34).

These findings suggest that a 'one size fits all' approach to predicting resistance such as molecular diagnostics or WGS with catalogue-based prediction will have limited performance, and indeed no single approach detected all the fluoroquinolone resistance present in the CRyPTIC dataset (Figure 36a,b). It is therefore important that new predictive options that consider a range of background factors, such as the machine learning models in chapter 5, are considered. To determine whether the associations are causative, *in vitro* evolution studies and MIC and relative fitness testing will be necessary to confirm or refute the hypotheses inspired by the regression models.

7.1.5 Including mixed alleles significantly improves catalogue performance

The prevalence of samples with mixed alleles in the DNA gyrase genes was at least 5.2% in the CRyPTIC dataset and all 14 catalogue resistance associated mutations were seen as part of a mixed population in at least one sample (Table 19). When predicting levofloxacin and moxifloxacin resistance from WGS data using the catalogue approach, there was a near 10% increase in the sensitivity without a significant decrease in specificity when including mixed

alleles (Figure 39a,b), bringing the fluoroquinolones into line with the first-line drugs isoniazid, rifampicin and ethambutol[42]. I also found that there was no correlation between a proxy for the relative size of the minority population (FRS) and MIC which suggests that even populations where a resistance conferring allele is only present in a small fraction can rapidly grow and outcompete a majority wild-type population in the presence of a fluoroquinolone antibiotic (Figure 38a-d).

The prevalence of mixed populations and the huge improvements their inclusion provides for catalogue-based resistance prediction means it is vital that all fluoroquinolone DST methods detect resistance present in a minor population. Together the findings also have strong implications for WGS pipelines; adequate sequencing depth across the DNA gyrase genes is important to detect very small minor populations as these are highly likely to be clinically relevant, and all mixed alleles at resistance associated genome indexes should be reported.

An important follow on to this work is to create a standardised bioinformatic pipeline for WGS variant calling in TB and specifically determine the optimal filters, including the optimal FRS cut-off, for how levofloxacin or moxifloxacin resistant variants should be called. A deep sequencing study could be considered to define the optimal cut offs; the CRyPTIC dataset provides relatively limited insight into very minor populations due to the low and variable sequencing depth (Figure 37). It would also be worth investigating whether any thresholds should be specific for each gene or resistance associated mutation.

7.1.6 Machine learning models can predict fluoroquinolone resistance

The machine learning models trained to predict resistance and susceptibility to levofloxacin and moxifloxacin had high sensitivities (Figure 43a,b) and the best performing models correctly predicted the phenotypes of most of the known mutations (Table 26, Table 27). Little explored until now in this discipline, protein structured-based machine learning models were also able to predict the MIC of isolates with good essential agreement (Figure 46a,b). As more data are collected to inform these and other machine learning models, their performance will continue to improve.

These results are especially encouraging because once machine learning models are trained, a prediction for a new clinical sample can be made almost instantaneously, and one can even predict the effect of all possible amino acid mutations (Appendix Table 3, Appendix Table 4, Appendix Table 5, Appendix Table 6) which would be highly impractical for more computationally intensive approaches such as RBFE calculations. The particularly high sensitivity of models suggest they might be best used to predict susceptibility, and in doing so, rule out resistance, to enable patients to start treatment programmes rapidly.

The high sensitivity of structure-based machine learning models for fluoroquinolone resistance prediction (Figure 43a,b), and the relative importance of protein structural features in the models (Table 24, Table 25), have implications for the wider field of infectious disease. A structure-based approach could be successful for predicting resistance in a range of other bacteria with DNA gyrase enzymes that are treated with fluoroquinolones: there are DNA gyrase crystal structures resolved from several infectious

species including *Staphylococcus aureus*[377], *Streptococcus pneumoniae*[378] and *Escherichia Coli*[379]. DNA gyrase enzymes are also highly conserved[342] and therefore structure-based predictions for one species could also be useful for making initial predictions in others where a crystal structure is not yet elucidated, and this could be explored in further work. Although the focus of this thesis is the fluoroquinolones, in theory, structure-based machine learning models could be useful for resistance prediction for any drug where resistance is primarily target mediated and a crystal structure has been elucidated.

7.1.7 The RBFE calculations could not make confident resistance predictions

Unfortunately, RBFE calculations could not make sufficiently confident predictions for two of the three resistance conferring mutations tested (Figure 55) and there was evidence of inaccuracy and non-normality in the calculations (Figure 58, Figure 60). The RBFE calculations were also difficult to set up; adaptations were required for modelling the protein-DNA covalent bonds, the moxifloxacin-Mg^{2+} interactions and mutations involving chiral centres (see Sections 6.2.1, 6.2.3). Together, these findings suggest RBFE may not yet be appropriate from making rapid predictions for large and complex protein targets.

I consider the main barrier to using the RBFE approach to predict fluoroquinolone resistance is the computational requirement that would be necessary to make a confident prediction. However, I think it important to remember that other tools, including machine learning, were originally shunned due to their computational requirements. I am therefore optimistic that continued improvement to MD codes and computational architecture could make RBFE a viable option in future. Further, the set-up of calculations is becoming more user friendly

as toolkits and packages for setting up and running simulations are continually developed by the community through the addition of new features[380].

Although the use of RBFE is not yet a practical primary predictive tool, there are important advantages to consider for a MD-based predictive approach; MD is theoretically exact and the chemical rationale for predictions can be understood. A combination of molecular dynamics and machine learning could therefore be considered, and I think there are two options that would be viable to explore:

- One could use features derived from molecular dynamics simulations to inform the machine learning models as these may be more representative than equivalent features measured from the static crystal structure.
- One could consider a funnelled approach, as we have previously proposed[381]. The effect of a mutation could be first predicted by a machine learning model where susceptible isolates could be quickly and accurately identified, as machine learning models predicting resistance had high sensitivity (Figure 43a,b). Any resistant predictions with low confidence could then be tested using an RFBE approach as further evidence.

7.2 Strengths and limitations

Although the strengths and limitations of individual analyses are discussed in the relevant chapter discussions, there are several overarching strengths and limitations of this work that should be highlighted and acknowledged.

7.2.1 Strengths

The major strength of this work is in the size of the dataset used to support my findings, which is unparalleled. This work shows how clinical data can be used to increase our understanding of patterns associated with resistance and build models to predict resistance. Further, the availability of the dataset and the codes used for analysis means analyses are reproducible and translatable to facilitate and inspire further research.

7.2.2 Limitations

There are several limitations to individual analyses which are discussed in the relevant chapter discussions, but there are some limitations that apply to the entirety of this thesis. Firstly, many findings are dependent on the ECOFFs used to distinguish resistance and susceptibility, and these could be wrong. Secondly the information collected for each CRyPTIC sample is limited, for example there is little or no data on prior treatment and the origin of a significant proportion of samples was not recorded, which could cause residual confounding. The genetic information on each sample is also constrained by the moderate depth to which each sample was sequenced. This is insufficient to probe just how small a minority population can be yet still give rise to resistance at the level of the sample in the presence of a fluoroquinolone, leading to possible inaccuracies in statistical models, training sets for machine learning and the average MICs calculated to compare with predictions from free energy calculations. Finally, the dataset is biased, although this is by design to ensure a diverse range of resistant isolates, and it means the data are not truly representative of global resistance patterns and makes comparison between different geographies and lineages difficult. Further, it is important to consider that the data contain relatively few

Lineage 6 isolates and no Lineage 5 isolates, so my conclusions may not hold for all TB infections.

7.3 Conclusion

Fluoroquinolone resistance poses a major threat to the treatment of TB and it is imperative that resistance in clinical samples is detected and action taken. WGS offers an attractive alternative to time consuming phenotypic DST and could become more accessible as cheaper sequencing technologies, such as Oxford Nanopore, are adopted. Catalogue-based resistance prediction from WGS data has important limitations due to the complexity of the genetic determinants of fluoroquinolone resistance, including minor alleles, lineage, country of origin and additional resistance, and the approach will never be exhaustive. Predictive methods that consider the effect of mutations on protein structure and chemistry could help address the exhaustivity problem: machine learning offers an attractive solution to improve performance of WGS for DST, whilst RBFE calculations could have application in future. There is still much to learn about the genetic determinants of fluoroquinolone resistance and this work highlights the impact of large, high-quality, matched phenotypic and WGS data to build both associative and predictive models.

8 References

1. The, CRyPTIC Consortium, A data compendium associating the genomes of 12,289 Mycobacterium tuberculosis isolates with quantitative resistance phenotypes to 13 antibiotics. *PLOS Biology* **2022,** *20* (8), e3001721.

2. Brankin, A. E.; Fowler, P. W., Predicting antibiotic resistance in complex protein targets using alchemical free energy methods. *Journal of Computational Chemistry* **2022**, *43* (26), 1771.

3. Emadi, M.; Delavari, S.; Bayati, M., Global socioeconomic inequality in the burden of communicable and non-communicable diseases and injuries: an analysis on global burden of disease study 2019. *BMC Public Health* **2021,** *21* (1), 1771.

4. Coates, M. M.; Ezzati, M.; Robles Aguilar, G.; Kwan, G. F.; Vigo, D.; Mocumbi, A. O.; Becker, A. E.; Makani, J.; Hyder, A. A.; Jain, Y.; Stefan, D. C.; Gupta, N.; Marx, A.; Bukhman, G., Burden of disease among the world's poorest billion people: An expert-informed secondary analysis of Global Burden of Disease estimates. *PloS one* **2021,** *16* (8), e0253073.

5. Wang, H.; Paulson, K. R.; Pease, S. A.; Watson, S.; Comfort, H.; Zheng, P.; Aravkin, A. Y.; Bisignano, C.; Barber, R. M.; Alam, T.; Fuller, J. E.; May, E. A.; Jones, D. P.; Frisch, M. E.; Abbafati, C.; Adolph, C.; Allorant, A.; Amlag, J. O.; Bang-Jensen, B.; Bertolacci, G. J.; Bloom, S. S.; Carter, A.; Castro, E.; Chakrabarti, S.; Chattopadhyay, J.; Cogen, R. M.; Collins, J. K.; Cooperrider, K.; Dai, X.; Dangel, W. J.; Daoud, F.; Dapper, C.; Deen, A.; Duncan, B. B.; Erickson, M.; Ewald, S. B.; Fedosseeva, T.; Ferrari, A. J.; Frostad, J. J.; Fullman, N.; Gallagher, J.; Gamkrelidze, A.; Guo, G.; He, J.; Helak, M.; Henry, N. J.; Hulland, E. N.; Huntley, B. M.; Kereselidze, M.; Lazzar-Atwood, A.; LeGrand, K. E.; Lindstrom, A.; Linebarger, E.; Lotufo, P. A.; Lozano, R.; Magistro, B.; Malta, D. C.; Månsson, J.; Mantilla Herrera, A. M.; Marinho, F.; Mirkuzie, A. H.; Misganaw, A. T.; Monasta, L.; Naik, P.; Nomura, S.; O'Brien, E. G.; O'Halloran, J. K.; Olana, L. T.; Ostroff, S. M.; Penberthy, L.; Reiner Jr, R. C.; Reinke, G.; Ribeiro, A. L. P.; Santomauro, D. F.; Schmidt, M. I.; Shaw, D. H.; Sheena, B. S.; Sholokhov, A.; Skhvitaridze, N.; Sorensen, R. J. D.; Spurlock, E. E.; Syailendrawati, R.; Topor-Madry, R.; Troeger, C. E.; Walcott, R.; Walker, A.; Wiysonge, C. S.; Worku, N. A.; Zigler, B.; Pigott, D. M.; Naghavi, M.; Mokdad, A. H.; Lim, S. S.; Hay, S. I.; Gakidou, E.; Murray, C. J. L., Estimating excess mortality due to the COVID-19 pandemic: a systematic analysis of COVID-19-related mortality. *The Lancet* **2022,** *399* (10334), 1513-1536.

6. Dheda, K.; Perumal, T.; Moultrie, H.; Perumal, R.; Esmail, A.; Scott, A. J.; Udwadia, Z.; Chang, K. C.; Peter, J.; Pooran, A.; von Delft, A.; von Delft, D.; Martinson, N.; Loveday, M.; Charalambous, S.; Kachingwe, E.; Jassat, W.; Cohen, C.; Tempia, S.; Fennelly, K.; Pai, M., The intersecting pandemics of tuberculosis and COVID-19: population-level and patient-level impact, clinical presentation, and corrective interventions. *The Lancet Respiratory Medicine* **2022,** *10* (6), 603-622.

7. Cave, A.; Demonstrator, A., The evidence for the incidence of tuberculosis in ancient Egypt. *British Journal of Tuberculosis* **1939,** *33* (3), 142-152.

8. Hershkovitz, I.; Donoghue, H. D.; Minnikin, D. E.; Besra, G. S.; Lee, O. Y.; Gernaey, A. M.; Galili, E.; Eshed, V.; Greenblatt, C. L.; Lemma, E.; Bar-Gal, G. K.; Spigelman, M., Detection and molecular characterization of 9,000-year-old Mycobacterium tuberculosis from a Neolithic settlement in the Eastern Mediterranean. *PloS one* **2008,** *3* (10), e3426.

9. Houben, R. M.; Dodd, P. J., The Global Burden of Latent Tuberculosis Infection: A Re-estimation Using Mathematical Modelling. *PLoS Med* **2016,** *13* (10), e1002152.

10. World Health Organization (WHO) *Tuberculosis Fact Sheet*; 2021.

11. Getahun, H.; Matteelli, A.; Chaisson, R. E.; Raviglione, M., Latent Mycobacterium tuberculosis infection. *N Engl J Med* **2015,** *372* (22), 2127-35.

12. World Health Organization (WHO), Global Tuberculosis Report 2021. **2021**.

13. Urdahl, K. B.; Shafiani, S.; Ernst, J. D., Initiation and regulation of T-cell responses in tuberculosis. *Mucosal Immunol* **2011,** *4* (3), 288-93.

14. Wayne, L. G.; Hayes, L. G., An in vitro model for sequential study of shiftdown of Mycobacterium tuberculosis through two stages of nonreplicating persistence. *Infect Immun* **1996,** *64* (6), 2062-9.

15. Gideon, H. P.; Flynn, J. L., Latent tuberculosis: what the host "sees"? *Immunol Res* **2011,** *50* (2-3), 202-12.

16. Hernández-Pando, R.; Jeyanathan, M.; Mengistu, G.; Aguilar, D.; Orozco, H.; Harboe, M.; Rook, G. A.; Bjune, G., Persistence of DNA from Mycobacterium tuberculosis in superficially normal lung tissue during latent infection. *Lancet* **2000,** *356* (9248), 2133-8.

17. Bishai, W. R., Rekindling old controversy on elusive lair of latent tuberculosis. *Lancet* **2000,** *356* (9248), 2113-4.

18. Neyrolles, O.; Hernández-Pando, R.; Pietri-Rouxel, F.; Fornès, P.; Tailleux, L.; Barrios Payán, J. A.; Pivert, E.; Bordat, Y.; Aguilar, D.; Prévost, M. C.; Petit, C.; Gicquel, B., Is adipose tissue a place for Mycobacterium tuberculosis persistence? *PloS one* **2006,** *1* (1), e43.

19. Chao, M. C.; Rubin, E. J., Letting sleeping dos lie: does dormancy play a role in tuberculosis? *Annu Rev Microbiol* **2010,** *64*, 293-311.

20. Gengenbacher, M.; Kaufmann, S. H., Mycobacterium tuberculosis: success through dormancy. *FEMS Microbiol Rev* **2012,** *36* (3), 514-32.

21. van Soolingen, D.; Hoogenboezem, T.; de Haas, P. E.; Hermans, P. W.; Koedam, M. A.; Teppema, K. S.; Brennan, P. J.; Besra, G. S.; Portaels, F.; Top, J.; Schouls, L. M.; van Embden, J. D., A novel pathogenic taxon of the Mycobacterium tuberculosis complex, Canetti: characterization of an exceptional isolate from Africa. *Int J Syst Bacteriol* **1997,** *47* (4), 1236-45.

22. Thompson, P. J.; Cousins, D. V.; Gow, B. L.; Collins, D. M.; Williamson, B. H.; Dagnia, H. T., Seals, seal trainers, and mycobacterial infection. *The American review of respiratory disease* **1993,** *147* (1), 164-7.

23. Smith, N. H.; Crawshaw, T.; Parry, J.; Birtles, R. J., Mycobacterium microti: More Diverse than Previously Thought. *Journal of Clinical Microbiology* **2009,** *47* (8), 2551-2559.

24. Prodinger, W. M.; Indra, A.; Koksalan, O. K.; Kilicaslan, Z.; Richter, E., Mycobacterium caprae infection in humans. *Expert Rev Anti Infect Ther* **2014,** *12* (12), 1501-13.

25. Palmer, M. V.; Thacker, T. C.; Waters, W. R.; Gortázar, C.; Corner, L. A. L., Mycobacterium bovis: A Model Pathogen at the Interface of Livestock, Wildlife, and Humans. *Veterinary Medicine International* **2012,** *2012*, 236205.

26. van Ingen, J.; Rahim, Z.; Mulder, A.; Boeree, M. J.; Simeone, R.; Brosch, R.; van Soolingen, D., Characterization of Mycobacterium orygis as *M. tuberculosis* complex subspecies. *Emerg Infect Dis* **2012,** *18* (4), 653-5.

27. Koch, R., Die Ätiologie der Tuberkulose (1882). In *Robert Koch : Zentrale Texte*, Springer Berlin Heidelberg: Berlin, Heidelberg, 2018; pp 113-131.

28. Cole, S. T.; Brosch, R.; Parkhill, J.; Garnier, T.; Churcher, C.; Harris, D.; Gordon, S. V.; Eiglmeier, K.; Gas, S.; Barry, C. E., 3rd; Tekaia, F.; Badcock, K.; Basham, D.; Brown, D.; Chillingworth, T.; Connor, R.; Davies, R.; Devlin, K.; Feltwell, T.; Gentles, S.; Hamlin, N.; Holroyd, S.; Hornsby, T.; Jagels, K.; Krogh, A.; McLean, J.; Moule, S.; Murphy, L.; Oliver, K.; Osborne, J.; Quail, M. A.; Rajandream, M. A.; Rogers, J.; Rutter, S.; Seeger, K.; Skelton, J.; Squares, R.; Squares, S.; Sulston, J. E.; Taylor, K.; Whitehead, S.; Barrell, B. G., Deciphering the biology of Mycobacterium tuberculosis from the complete genome sequence. *Nature* **1998,** *393* (6685), 537-44.

29. Heath, R. J.; White, S. W.; Rock, C. O., Inhibitors of fatty acid synthesis as antimicrobial chemotherapeutics. *Appl Microbiol Biotechnol* **2002,** *58* (6), 695-703.

30. Borisova, O. F.; Shchyolkina, A. K.; Chernov, B. K.; Tchurikov, N. A., Relative stability of AT and GC pairs in parallel DNA duplex formed by a natural sequence. *FEBS Lett* **1993,** *322* (3), 304-6.

31. Hu, E.-Z.; Lan, X.-R.; Liu, Z.-L.; Gao, J.; Niu, D.-K., A positive correlation between GC content and growth temperature in prokaryotes. *BMC Genomics* **2022,** *23* (1), 110.

32. Comas, I.; Coscolla, M.; Luo, T.; Borrell, S.; Holt, K. E.; Kato-Maeda, M.; Parkhill, J.; Malla, B.; Berg, S.; Thwaites, G.; Yeboah-Manu, D.; Bothamley, G.; Mei, J.; Wei, L.; Bentley, S.; Harris, S. R.; Niemann, S.; Diel, R.; Aseffa, A.; Gao, Q.; Young, D.; Gagneux, S., Out-of-Africa migration and Neolithic coexpansion of Mycobacterium tuberculosis with modern humans. *Nat Genet* **2013,** *45* (10), 1176-82.

33. Gagneux, S.; DeRiemer, K.; Van, T.; Kato-Maeda, M.; Jong, B. C. d.; Narayanan, S.; Nicol, M.; Niemann, S.; Kremer, K.; Gutierrez, M. C.; Hilty, M.; Hopewell, P. C.; Small, P. M., Variable host pathogen compatibility in Mycobacterium tuberculosis. *Proceedings of the National Academy of Sciences* **2006,** *103* (8), 2869-2873.

34. Gagneux, S., Ecology and evolution of Mycobacterium tuberculosis. *Nat Rev Microbiol* **2018,** *16* (4), 202-213.

35. Menardo, F.; Rutaihwa, L. K.; Zwyer, M.; Borrell, S.; Comas, I.; Conceição, E. C.; Coscolla, M.; Cox, H.; Joloba, M.; Dou, H. Y.; Feldmann, J.; Fenner, L.; Fyfe, J.; Gao, Q.; García de Viedma, D.; Garcia-Basteiro, A. L.; Gygli, S. M.; Hella, J.; Hiza, H.; Jugheli, L.; Kamwela, L.; Kato-Maeda, M.; Liu, Q.; Ley, S. D.; Loiseau, C.; Mahasirimongkol, S.; Malla, B.; Palittapongarnpim, P.; Rakotosamimanana, N.; Rasolofo, V.; Reinhard, M.; Reither, K.; Sasamalo, M.; Silva Duarte, R.; Sola, C.; Suffys, P.; Batista Lima, K. V.; Yeboah-Manu, D.; Beisel, C.; Brites, D.; Gagneux, S., Local adaptation in populations of Mycobacterium tuberculosis endemic to the Indian Ocean Rim. *F1000Res* **2021,** *10*, 60.

36. Freschi, L.; Vargas, R., Jr.; Husain, A.; Kamal, S. M. M.; Skrahina, A.; Tahseen, S.; Ismail, N.; Barbova, A.; Niemann, S.; Cirillo, D. M.; Dean, A. S.; Zignol, M.; Farhat, M. R., Population structure, biogeography and transmissibility of Mycobacterium tuberculosis. *Nat Commun* **2021,** *12* (1), 6099.

37. Brosch, R.; Gordon, S. V.; Marmiesse, M.; Brodin, P.; Buchrieser, C.; Eiglmeier, K.; Garnier, T.; Gutierrez, C.; Hewinson, G.; Kremer, K.; Parsons, L. M.; Pym, A. S.; Samper, S.; van Soolingen, D.; Cole, S. T., A new evolutionary scenario for the Mycobacterium tuberculosis complex. *Proc Natl Acad Sci U S A* **2002,** *99* (6), 3684-9.

38. Bottai, D.; Frigui, W.; Sayes, F.; Di Luca, M.; Spadoni, D.; Pawlik, A.; Zoppo, M.; Orgeur, M.; Khanna, V.; Hardy, D.; Mangenot, S.; Barbe, V.; Medigue, C.; Ma, L.; Bouchier, C.; Tavanti, A.; Larrouy-Maumus, G.; Brosch, R., TbD1 deletion as a driver of the evolutionary success of modern epidemic *Mycobacterium tuberculosis* lineages. *Nature communications* **2020,** *11* (1), 684.

39. Blouin, Y.; Hauck, Y.; Soler, C.; Fabre, M.; Vong, R.; Dehan, C.; Cazajous, G.; Massoure, P. L.; Kraemer, P.; Jenkins, A.; Garnotel, E.; Pourcel, C.; Vergnaud, G., Significance of the identification in the Horn of Africa of an exceptionally deep branching Mycobacterium tuberculosis clade. *PloS one* **2012,** *7* (12), e52841.

40. Gehre, F.; Kumar, S.; Kendall, L.; Ejo, M.; Secka, O.; Ofori-Anyinam, B.; Abatih, E.; Antonio, M.; Berkvens, D.; de Jong, B. C., A Mycobacterial Perspective on Tuberculosis in West Africa: Significant Geographical Variation of M. africanum and Other *M. tuberculosis* Complex Lineages. *PLoS neglected tropical diseases* **2016,** *10* (3), e0004408.

41. Netikul, T.; Palittapongarnpim, P.; Thawornwattana, Y.; Plitphonganphim, S., Estimation of the global burden of Mycobacterium tuberculosis lineage 1. *Infection, Genetics and Evolution* **2021,** *91*, 104802.

42. Walker, T. M.; Miotto, P.; Köser, C. U.; Fowler, P. W.; Knaggs, J.; Iqbal, Z.; Hunt, M.; Chindelevitch, L.; Farhat, M. R.; Cirillo, D. M.; Comas, I.; Posey, J.; Omar, S. V.; Peto, T. E. A.; Suresh, A.; Uplekar, S.; Laurent, S.; Colman, R. E.; Nathanson, C.-M.; Zignol, M.; Walker, A. S.; Crook, D. W.; Ismail, N.; Rodwell, T. C.; Walker, A. S.; Steyn, A. J. C.; Lalvani, A.; Baulard, A.; Christoffels, A.; Mendoza-Ticona, A.; Trovato, A.; Skrahina, A.; Lachapelle, A. S.; Brankin, A.; Piatek, A.; Gibertoni Cruz, A.; Koch, A.; Cabibbe, A. M.; Spitaleri, A.; Brandao, A. P.; Chaiprasert, A.; Suresh, A.; Barbova, A.; Van Rie, A.; Ghodousi, A.; Bainomugisa, A.; Mandal, A.; Roohi, A.; Javid, B.; Zhu, B.; Letcher, B.; Rodrigues, C.; Nimmo, C.; Nathanson, C.-M.; Duncan, C.; Coulter, C.; Utpatel, C.; Liu, C.; Grazian, C.;

Kong, C.; Köser, C. U.; Wilson, D. J.; Cirillo, D. M.; Matias, D.; Jorgensen, D.; Zimenkov, D.; Chetty, D.; Moore, D. A. J.; Clifton, D. A.; Crook, D. W.; van Soolingen, D.; Liu, D.; Kohlerschmidt, D.; Barreira, D.; Ngcamu, D.; Santos Lazaro, E. D.; Kelly, E.; Borroni, E.; Roycroft, E.; Andre, E.; Böttger, E. C.; Robinson, E.; Menardo, F.; Mendes, F. F.; Jamieson, F. B.; Coll, F.; Gao, G. F.; Kasule, G. W.; Rossolini, G. M.; Rodger, G.; Smith, E. G.; Meintjes, G.; Thwaites, G.; Hoffmann, H.; Albert, H.; Cox, H.; Laurenson, I. F.; Comas, I.; Arandjelovic, I.; Barilar, I.; Robledo, J.; Millard, J.; Johnston, J.; Posey, J.; Andrews, J. R.; Knaggs, J.; Gardy, J.; Guthrie, J.; Taylor, J.; Werngren, J.; Metcalfe, J.; Coronel, J.; Shea, J.; Carter, J.; Pinhata, J. M. W.; Kus, J. V.; Todt, K.; Holt, K.; Nilgiriwala, K. S.; Ghisi, K. T.; Malone, K. M.; Faksri, K.; Musser, K. A.; Joseph, L.; Rigouts, L.; Chindelevitch, L.; Jarrett, L.; Grandjean, L.; Ferrazoli, L.; Rodrigues, M.; Farhat, M.; Schito, M.; Fitzgibbon, M. M.; Loembé, M. M.; Wijkander, M.; Ballif, M.; Rabodoarivelo, M.-S.; Mihalic, M.; Wilcox, M.; Hunt, M.; Zignol, M.; Merker, M.; Egger, M.; O'Donnell, M.; Caws, M.; Wu, M.-H.; Whitfield, M. G.; Inouye, M.; Mansjö, M.; Dang Thi, M. H.; Joloba, M.; Kamal, S. M. M.; Okozi, N.; Ismail, N.; Mistry, N.; Hoang, N. N.; Rakotosamimanana, N.; Paton, N. I.; Rancoita, P. M. V.; Miotto, P.; Lapierre, P.; Hall, P. J.; Tang, P.; Claxton, P.; Wintringer, P.; Keller, P. M.; Thai, P. V. K.; Fowler, P. W.; Supply, P.; Srilohasin, P.; Suriyaphol, P.; Rathod, P.; Kambli, P.; Groenheit, R.; Colman, R. E.; Ong, R. T.-H.; Warren, R. M.; Wilkinson, R. J.; Diel, R.; Oliveira, R. S.; Khot, R.; Jou, R.; Tahseen, S.; Laurent, S.; Gharbia, S.; Kouchaki, S.; Shah, S.; Plesnik, S.; Earle, S. G.; Dunstan, S.; Hoosdally, S. J.; Mitarai, S.; Gagneux, S.; Omar, S. V.; Yao, S.-Y.; Grandjean Lapierre, S.; Battaglia, S.; Niemann, S.; Pandey, S.; Uplekar, S.; Halse, T. A.; Cohen, T.; Cortes, T.; Prammananan, T.; Kohl, T. A.; Thuong, N. T. T.; Teo, T. Y.; Peto, T. E. A.; Rodwell, T. C.; William, T.; Walker, T. M.; Rogers, T. R.; Surve, U.; Mathys, V.; Furió, V.; Cook, V.; Vijay, S.; Escuyer, V.; Dreyer, V.; Sintchenko, V.; Saphonn, V.; Solano, W.; Lin, W.-H.; van Gemert, W.; He, W.; Yang, Y.; Zhao, Y.; Qin, Y.; Xiao, Y.-X.; Hasan, Z.; Iqbal, Z.; Puyen, Z. M., The 2021 WHO catalogue of Mycobacterium tuberculosis complex mutations associated with drug resistance: a genotypic analysis. *The Lancet Microbe* **2022,** *3* (4), e265.

43. Nathavitharana, R. R.; Jijon, D. F.; Pal, P.; Rane, S., Diagnosing active tuberculosis in primary care. *BMJ* **2021,** *374*, n1590.

44. Pfyffer, G. E.; Wittwer, F., Incubation time of mycobacterial cultures: how long is long enough to issue a final negative report to the clinician? *J Clin Microbiol* **2012,** *50* (12), 4188-9.

45. Chihota, V. N.; Grant, A. D.; Fielding, K.; Ndibongo, B.; van Zyl, A.; Muirhead, D.; Churchyard, G. J., Liquid vs. solid culture for tuberculosis: performance and cost in a resource-constrained setting. *The international journal of tuberculosis and lung disease : the official journal of the International Union against Tuberculosis and Lung Disease* **2010,** *14* (8), 1024-1031.

46. Kik, S. V.; Denkinger, C. M.; Chedore, P.; Pai, M., Replacing smear microscopy for the diagnosis of tuberculosis: what is the market potential? *European Respiratory Journal* **2014,** *43* (6), 1793-1796.

47. Cattamanchi, A.; Davis, J. L.; Pai, M.; Huang, L.; Hopewell, P. C.; Steingart, K. R., Does bleach processing increase the accuracy of sputum smear microscopy for diagnosing pulmonary tuberculosis? *J Clin Microbiol* **2010,** *48* (7), 2433-9.

48. Theron, G.; Peter, J.; Dowdy, D.; Langley, I.; Squire, S. B.; Dheda, K., Do high rates of empirical treatment undermine the potential effect of new diagnostic tests for tuberculosis in high-burden settings? *Lancet Infect Dis* **2014,** *14* (6), 527-32.

49. World Health Organization (WHO), WHO consolidated guidelines on tuberculosis. Module 3: Diagnosis - Rapid diagnostics for tuberculosis detection 2021 update. **2021**.

50. Adelman, M. W.; Kurbatova, E.; Wang, Y. F.; Leonard, M. K.; White, N.; McFarland, D. A.; Blumberg, H. M., Cost analysis of a nucleic acid amplification test in the diagnosis of pulmonary tuberculosis at an urban hospital with a high prevalence of TB/HIV. *PloS one* **2014,** *9* (7), e100649.

51. Atherton, R. R.; Cresswell, F. V.; Ellis, J.; Kitaka, S. B.; Boulware, D. R., Xpert MTB/RIF Ultra for Tuberculosis Testing in Children: A Mini-Review and Commentary. *Front Pediatr* **2019,** *7*, 34.

52. Pfuetze, K. H.; Pyle, M. M.; Hinshaw, H. C.; Feldman, W. H., The First Clinical Trial of Streptomycin in Human Tuberculosis. *American Review of Tuberculosis and Pulmonary Diseases* **1955,** *71* (5), 752-754.

53. Iseman, M. D., Tuberculosis therapy: past, present and future. *European Respiratory Journal* **2002,** *20* (36 suppl), 87S-94s.

54. VARIOUS combinations of isoniazid with streptomycin or with P.A.S. in the treatment of pulmonary tuberculosis; seventh report to the Medical Research Council by their Tuberculosis Chemotherapy Trials Committee. *Br Med J* **1955,** *1* (4911), 435-45.

55. Doster, B.; Murray, F. J.; Newman, R.; Woolpert, S. F., Ethambutol in the initial treatment of pulmonary tuberculosis. U.S. Public Health Service tuberculosis therapy trials. *The American review of respiratory disease* **1973,** *107* (2), 177-90.

56. Tsukamura, M.; Nakamura, E.; Yoshii, S.; Amano, H., Therapeutic effect of a new antibacterial substance ofloxacin (DL8280) on pulmonary tuberculosis. *The American review of respiratory disease* **1985,** *131* (3), 352-356.

57. Ryan, N. J.; Lo, J. H., Delamanid: first global approval. *Drugs* **2014,** *74* (9), 1041-5.

58. Cox, E.; Laessig, K., FDA Approval of Bedaquiline — The Benefit–Risk Balance for Drug-Resistant Tuberculosis. *New England Journal of Medicine* **2014,** *371* (8), 689-691.

59. Lee, M.; Lee, J.; Carroll, M. W.; Choi, H.; Min, S.; Song, T.; Via, L. E.; Goldfeder, L. C.; Kang, E.; Jin, B.; Park, H.; Kwak, H.; Kim, H.; Jeon, H. S.; Jeong, I.; Joh, J. S.; Chen, R. Y.; Olivier, K. N.; Shaw, P. A.; Follmann, D.; Song, S. D.; Lee, J. K.; Lee, D.; Kim, C. T.; Dartois, V.; Park, S. K.; Cho, S. N.; Barry, C. E., 3rd, Linezolid for treatment of chronic extensively drug-resistant tuberculosis. *N Engl J Med* **2012,** *367* (16), 1508-18.

60. Tang, S.; Yao, L.; Hao, X.; Liu, Y.; Zeng, L.; Liu, G.; Li, M.; Li, F.; Wu, M.; Zhu, Y.; Sun, H.; Gu, J.; Wang, X.; Zhang, Z., Clofazimine for the Treatment of Multidrug-Resistant Tuberculosis: Prospective, Multicenter, Randomized Controlled Study in China. *Clinical Infectious Diseases* **2015,** *60* (9), 1361-1367.

61. Pawar, A.; Jha, P.; Konwar, C.; Chaudhry, U.; Chopra, M.; Saluja, D., Ethambutol targets the glutamate racemase of Mycobacterium tuberculosis-an enzyme involved in peptidoglycan biosynthesis. *Appl Microbiol Biotechnol* **2019,** *103* (2), 843-851.

62. Gopal, P.; Sarathy, J. P.; Yee, M.; Ragunathan, P.; Shin, J.; Bhushan, S.; Zhu, J.; Akopian, T.; Kandror, O.; Lim, T. K.; Gengenbacher, M.; Lin, Q.; Rubin, E. J.; Grüber, G.; Dick, T., Pyrazinamide triggers degradation of its target aspartate decarboxylase. *Nat Commun* **2020,** *11* (1), 1661.

63. Koul, A.; Dendouga, N.; Vergauwen, K.; Molenberghs, B.; Vranckx, L.; Willebrords, R.; Ristic, Z.; Lill, H.; Dorange, I.; Guillemont, J.; Bald, D.; Andries, K., Diarylquinolines target subunit c of mycobacterial ATP synthase. *Nat Chem Biol* **2007,** *3* (6), 323-4.

64. Cholo, M. C.; Mothiba, M. T.; Fourie, B.; Anderson, R., Mechanisms of action and therapeutic efficacies of the lipophilic antimycobacterial agents clofazimine and bedaquiline. *Journal of Antimicrobial Chemotherapy* **2016,** *72* (2), 338-353.

65. Manjunatha, U.; Boshoff, H. I.; Barry, C. E., The mechanism of action of PA-824: Novel insights from transcriptional profiling. *Commun Integr Biol* **2009,** *2* (3), 215-8.

66. Van den Bossche, A.; Varet, H.; Sury, A.; Sismeiro, O.; Legendre, R.; Coppee, J. Y.; Mathys, V.; Ceyssens, P. J., Transcriptional profiling of a laboratory and clinical Mycobacterium tuberculosis strain suggests respiratory poisoning upon exposure to delamanid. *Tuberculosis (Edinb)* **2019,** *117*, 18-23.

67. Zheng, J.; Rubin, E. J.; Bifani, P.; Mathys, V.; Lim, V.; Au, M.; Jang, J.; Nam, J.; Dick, T.; Walker, J. R.; Pethe, K.; Camacho, L. R., para-Aminosalicylic acid is a prodrug targeting dihydrofolate reductase in Mycobacterium tuberculosis. *J Biol Chem* **2013,** *288* (32), 23447-56.

68. World Health Organization (WHO), Rapid Communication: Key changes to treatment of multidrug- and rifampicin-resistant tuberculosis (MDR/RR-TB) **2018**.

69. World Health Organization (WHO), WHO consolidated guidelines on tuberculosis. Module 4: treatment - drug-susceptible tuberculosis treatment. **2020**.

70. World Health Organization (WHO), WHO consolidated guidelines on tuberculosis. Module 4: treatment - drug-resistant tuberculosis treatment. **2020,** *Licence: CC BY-NC-SA 3.0 IGO.*

71. World Health Organization (WHO), WHO consolidated guidelines on tuberculosis. Module 1: Prevention. **2020,** *Licence: CC BY-NC-SA 3.0 IGO.*

72. Badje, A.; Moh, R.; Gabillard, D.; Guéhi, C.; Kabran, M.; Ntakpé, J. B.; Carrou, J. L.; Kouame, G. M.; Ouattara, E.; Messou, E.; Anzian, A.; Minga, A.; Gnokoro, J.; Gouesse, P.; Emieme, A.; Toni, T. D.; Rabe, C.; Sidibé, B.; Nzunetu, G.; Dohoun, L.; Yao, A.; Kamagate, S.; Amon, S.; Kouame, A. B.; Koua, A.; Kouamé, E.; Daligou, M.; Hawerlander, D.; Ackoundzé, S.; Koule, S.; Séri, J.; Ani, A.; Dembélé, F.; Koné, F.; Oyebi, M.; Mbakop, N.; Makaila, O.; Babatunde, C.; Babatunde, N.; Bleoué, G.; Tchoutedjem, M.; Kouadio, A. C.; Sena, G.; Yededji, S. Y.; Karcher, S.; Rouzioux, C.; Kouame, A.; Assi, R.; Bakayoko, A.; Domoua, S. K.; Deschamps, N.; Aka, K.; N'Dri-Yoman, T.; Salamon, R.; Journot, V.; Ahibo,

H.; Ouassa, T.; Menan, H.; Inwoley, A.; Danel, C.; Eholié, S. P.; Anglaret, X., Effect of isoniazid preventive therapy on risk of death in west African, HIV-infected adults with high CD4 cell counts: long-term follow-up of the Temprano ANRS 12136 trial. *Lancet Glob Health* **2017,** *5* (11), e1080-e1089.

73. Abossie, A.; Yohanes, T., Assessment of isoniazid preventive therapy in the reduction of tuberculosis among ART patients in Arba Minch Hospital, Ethiopia. *Ther Clin Risk Manag* **2017,** *13*, 361-366.

74. Russom, M.; Woldu, H. G.; Berhane, A.; Jeannetot, D. Y. B.; Stricker, B. H.; Verhamme, K., Effectiveness of a 6-Month Isoniazid on Prevention of Incident Tuberculosis Among People Living with HIV in Eritrea: A Retrospective Cohort Study. *Infectious Diseases and Therapy* **2022,** *11* (1), 559-579.

75. Mangtani, P.; Abubakar, I.; Ariti, C.; Beynon, R.; Pimpin, L.; Fine, P. E.; Rodrigues, L. C.; Smith, P. G.; Lipman, M.; Whiting, P. F.; Sterne, J. A., Protection by BCG vaccine against tuberculosis: a systematic review of randomized controlled trials. *Clin Infect Dis* **2014,** *58* (4), 470-80.

76. Abubakar, I.; Pimpin, L.; Ariti, C.; Beynon, R.; Mangtani, P.; Sterne, J. A.; Fine, P. E.; Smith, P. G.; Lipman, M.; Elliman, D.; Watson, J. M.; Drumright, L. N.; Whiting, P. F.; Vynnycky, E.; Rodrigues, L. C., Systematic review and meta-analysis of the current evidence on the duration of protection by bacillus Calmette-Guérin vaccination against tuberculosis. *Health Technol Assess* **2013,** *17* (37), 1-372, v-vi.

77. Van Der Meeren, O.; Hatherill, M.; Nduba, V.; Wilkinson, R. J.; Muyoyeta, M.; Van Brakel, E.; Ayles, H. M.; Henostroza, G.; Thienemann, F.; Scriba, T. J.; Diacon, A.; Blatner, G. L.; Demoitié, M.-A.; Tameris, M.; Malahleha, M.; Innes, J. C.; Hellström, E.; Martinson, N.; Singh, T.; Akite, E. J.; Khatoon Azam, A.; Bollaerts, A.; Ginsberg, A. M.; Evans, T. G.; Gillard, P.; Tait, D. R., Phase 2b Controlled Trial of M72/AS01E Vaccine to Prevent Tuberculosis. *New England Journal of Medicine* **2018,** *379* (17), 1621-1634.

78. Oxlade, O.; Murray, M., Tuberculosis and poverty: why are the poor at greater risk in India? *PloS one* **2012,** *7* (11), e47533.

79. Pyle, M. M., Relative numbers of resistant tubercle bacilli in sputa of patients before and during treatment with streptomycin. *Proc Staff Meet Mayo Clin* **1947,** *22* (21), 465-73.

80. Skrahina, A.; Hurevich, H.; Zalutskaya, A.; Sahalchyk, E.; Astrauko, A.; van Gemert, W.; Hoffner, S.; Rusovich, V.; Zignol, M., Alarming levels of drug-resistant tuberculosis in Belarus: results of a survey in Minsk. *Eur Respir J* **2012,** *39* (6), 1425-31.

81. Glasauer, S.; Altmann, D.; Hauer, B.; Brodhun, B.; Haas, W.; Perumal, N., First-line tuberculosis drug resistance patterns and associated risk factors in Germany, 2008-2017. *PloS one* **2019,** *14* (6), e0217597.

82. Chung-Delgado, K.; Guillen-Bravo, S.; Revilla-Montag, A.; Bernabe-Ortiz, A., Mortality among MDR-TB cases: comparison with drug-susceptible tuberculosis and associated factors. *PloS one* **2015,** *10* (3), e0119332.

83. Zürcher, K.; Reichmuth, M. L.; Ballif, M.; Loiseau, C.; Borrell, S.; Reinhard, M.; Skrivankova, V.; Hömke, R.; Sander, P.; Avihingsanon, A.; Abimiku, A. l. G.; Marcy, O.; Collantes, J.; Carter, E. J.; Wilkinson, R. J.; Cox, H.; Yotebieng, M.; Huebner, R.; Fenner, L.; Böttger, E. C.; Gagneux, S.; Egger, M., Mortality from drug-resistant tuberculosis in high-burden countries comparing routine drug susceptibility testing with whole-genome sequencing: a multicentre cohort study. *The Lancet Microbe* **2021,** *2* (7), e320-e330.

84. World Health Organization (WHO) *Meeting report of the WHO expert consultation on the definition of extensively drug-resistant tuberculosis*; 2020.

85. Coll, F.; Phelan, J.; Hill-Cawthorne, G. A.; Nair, M. B.; Mallard, K.; Ali, S.; Abdallah, A. M.; Alghamdi, S.; Alsomali, M.; Ahmed, A. O.; Portelli, S.; Oppong, Y.; Alves, A.; Bessa, T. B.; Campino, S.; Caws, M.; Chatterjee, A.; Crampin, A. C.; Dheda, K.; Furnham, N.; Glynn, J. R.; Grandjean, L.; Minh Ha, D.; Hasan, R.; Hasan, Z.; Hibberd, M. L.; Joloba, M.; Jones-Lopez, E. C.; Matsumoto, T.; Miranda, A.; Moore, D. J.; Mocillo, N.; Panaiotov, S.; Parkhill, J.; Penha, C.; Perdigao, J.; Portugal, I.; Rchiad, Z.; Robledo, J.; Sheen, P.; Shesha, N. T.; Sirgel, F. A.; Sola, C.; Oliveira Sousa, E.; Streicher, E. M.; Helden, P. V.; Viveiros, M.; Warren, R. M.; McNerney, R.; Pain, A.; Clark, T. G., Genome-wide analysis of multi- and extensively drug-resistant Mycobacterium tuberculosis. *Nat Genet* **2018,** *50* (2), 307-316.

86. Ramaswamy, S.; Musser, J. M., Molecular genetic basis of antimicrobial agent resistance in Mycobacterium tuberculosis: 1998 update. *Tuber Lung Dis* **1998,** *79* (1), 3-29.

87. Zhang, Y.; Heym, B.; Allen, B.; Young, D.; Cole, S., The catalase-peroxidase gene and isoniazid resistance of Mycobacterium tuberculosis. *Nature* **1992,** *358* (6387), 591-3.

88. Guo, H.; Seet, Q.; Denkin, S.; Parsons, L.; Zhang, Y., Molecular characterization of isoniazid-resistant clinical isolates of Mycobacterium tuberculosis from the USA. *Journal of medical microbiology* **2006,** *55* (Pt 11), 1527-1531.

89. Remm, S.; Earp, J. C.; Dick, T.; Dartois, V.; Seeger, M. A., Critical discussion on drug efflux in Mycobacterium tuberculosis. *FEMS Microbiology Reviews* **2021,** *46* (1).

90. Jarlier, V.; Nikaido, H., Mycobacterial cell wall: structure and role in natural resistance to antibiotics. *FEMS Microbiol Lett* **1994,** *123* (1-2), 11-8.

91. Sarathy, J. P.; Dartois, V.; Lee, E. J., The role of transport mechanisms in mycobacterium tuberculosis drug resistance and tolerance. *Pharmaceuticals (Basel)* **2012,** *5* (11), 1210-35.

92. Manson, A. L.; Cohen, K. A.; Abeel, T.; Desjardins, C. A.; Armstrong, D. T.; Barry, C. E., 3rd; Brand, J.; Chapman, S. B.; Cho, S. N.; Gabrielian, A.; Gomez, J.; Jodals, A. M.; Joloba, M.; Jureen, P.; Lee, J. S.; Malinga, L.; Maiga, M.; Nordenberg, D.; Noroc, E.; Romancenco, E.; Salazar, A.; Ssengooba, W.; Velayati, A. A.; Winglee, K.; Zalutskaya, A.; Via, L. E.; Cassell, G. H.; Dorman, S. E.; Ellner, J.; Farnia, P.; Galagan, J. E.; Rosenthal, A.; Crudu, V.; Homorodean, D.; Hsueh, P. R.; Narayanan, S.; Pym, A. S.; Skrahina, A.; Swaminathan, S.; Van der Walt, M.; Alland, D.; Bishai, W. R.; Cohen, T.; Hoffner, S.; Birren, B. W.; Earl, A. M., Genomic analysis of globally diverse Mycobacterium tuberculosis strains provides insights into the emergence and spread of multidrug resistance. *Nat Genet* **2017,** *49* (3), 395-402.

93. Pym, A. S.; Saint-Joanis, B.; Cole, S. T., Effect of katG mutations on the virulence of Mycobacterium tuberculosis and the implication for transmission in humans. *Infect Immun* **2002,** *70* (9), 4955-60.

94. Sulis, G.; Pai, M., Isoniazid-resistant tuberculosis: A problem we can no longer ignore. *PLOS Medicine* **2020,** *17* (1), e1003023.

95. Cohen, K. A.; Abeel, T.; Manson McGuire, A.; Desjardins, C. A.; Munsamy, V.; Shea, T. P.; Walker, B. J.; Bantubani, N.; Almeida, D. V.; Alvarado, L.; Chapman, S. B.; Mvelase, N. R.; Duffy, E. Y.; Fitzgerald, M. G.; Govender, P.; Gujja, S.; Hamilton, S.; Howarth, C.; Larimer, J. D.; Maharaj, K.; Pearson, M. D.; Priest, M. E.; Zeng, Q.; Padayatchi, N.; Grosset, J.; Young, S. K.; Wortman, J.; Mlisana, K. P.; O'Donnell, M. R.; Birren, B. W.; Bishai, W. R.; Pym, A. S.; Earl, A. M., Evolution of Extensively Drug-Resistant Tuberculosis over Four Decades: Whole Genome Sequencing and Dating Analysis of Mycobacterium tuberculosis Isolates from KwaZulu-Natal. *PLoS Med* **2015,** *12* (9), e1001880.

96. Cox, H. S.; Sibilia, K.; Feuerriegel, S.; Kalon, S.; Polonsky, J.; Khamraev, A. K.; Rüsch-Gerdes, S.; Mills, C.; Niemann, S., Emergence of extensive drug resistance during treatment for multidrug-resistant tuberculosis. *N Engl J Med* **2008,** *359* (22), 2398-400.

97. Bloemberg, G. V.; Keller, P. M.; Stucki, D.; Trauner, A.; Borrell, S.; Latshang, T.; Coscolla, M.; Rothe, T.; Hömke, R.; Ritter, C.; Feldmann, J.; Schulthess, B.; Gagneux, S.; Böttger, E. C., Acquired Resistance to Bedaquiline and Delamanid in Therapy for Tuberculosis. *N Engl J Med* **2015,** *373* (20), 1986-8.

98. Yoshiyama, T.; Takaki, A.; Aono, A.; Mitarai, S.; Okumura, M.; Ohta, K.; Kato, S., Multidrug Resistant Tuberculosis With Simultaneously Acquired Drug Resistance to Bedaquiline and Delamanid. *Clin Infect Dis* **2021,** *73* (12), 2329-2331.

99. Zhou, Y.; van den Hof, S.; Wang, S.; Pang, Y.; Zhao, B.; Xia, H.; Anthony, R.; Ou, X.; Li, Q.; Zheng, Y.; Song, Y.; Zhao, Y.; van Soolingen, D., Association between genotype and drug resistance profiles of Mycobacterium tuberculosis strains circulating in China in a national drug resistance survey. *PloS one* **2017,** *12* (3), e0174197.

100. Niemann, S.; Diel, R.; Khechinashvili, G.; Gegia, M.; Mdivani, N.; Tang, Y. W., Mycobacterium tuberculosis Beijing lineage favors the spread of multidrug-resistant tuberculosis in the Republic of Georgia. *J Clin Microbiol* **2010,** *48* (10), 3544-50.

101. Ektefaie, Y.; Dixit, A.; Freschi, L.; Farhat, M. R., Globally diverse Mycobacterium tuberculosis resistance acquisition: a retrospective geographical and temporal analysis of whole genome sequences. *The Lancet. Microbe* **2021,** *2* (3), e96-e104.

102. Seung, K. J.; Gelmanova, I. E.; Peremitin, G. G.; Golubchikova, V. T.; Pavlova, V. E.; Sirotkina, O. B.; Yanova, G. V.; Strelis, A. K., The effect of initial drug resistance on treatment response and acquired drug resistance during standardized short-course chemotherapy for tuberculosis. *Clin Infect Dis* **2004,** *39* (9), 1321-8.

103. Seung, K. J.; Keshavjee, S.; Rich, M. L., Multidrug-Resistant Tuberculosis and Extensively Drug-Resistant Tuberculosis. *Cold Spring Harb Perspect Med* **2015,** *5* (9), a017863.

104. Ford, C.; Yusim, K.; Ioerger, T.; Feng, S.; Chase, M.; Greene, M.; Korber, B.; Fortune, S., Mycobacterium tuberculosis--heterogeneity revealed through whole genome sequencing. *Tuberculosis (Edinb)* **2012,** *92* (3), 194-201.

105. Sun, G.; Luo, T.; Yang, C.; Dong, X.; Li, J.; Zhu, Y.; Zheng, H.; Tian, W.; Wang, S.; Barry, C. E., 3rd; Mei, J.; Gao, Q., Dynamic population changes in Mycobacterium tuberculosis during acquisition and fixation of drug resistance in patients. *J Infect Dis* **2012,** *206* (11), 1724-33.

106. de Vos, M.; Ley, S. D.; Wiggins, K. B.; Derendinger, B.; Dippenaar, A.; Grobbelaar, M.; Reuter, A.; Dolby, T.; Burns, S.; Schito, M.; Engelthaler, D. M.; Metcalfe, J.; Theron, G.; van Rie, A.; Posey, J.; Warren, R.; Cox, H., Bedaquiline Microheteroresistance after Cessation of Tuberculosis Treatment. *N Engl J Med* **2019,** *380* (22), 2178-2180.

107. Dheda, K.; Lenders, L.; Magombedze, G.; Srivastava, S.; Raj, P.; Arning, E.; Ashcraft, P.; Bottiglieri, T.; Wainwright, H.; Pennel, T.; Linegar, A.; Moodley, L.; Pooran, A.; Pasipanodya, J. G.; Sirgel, F. A.; van Helden, P. D.; Wakeland, E.; Warren, R. M.; Gumbo, T., Drug-Penetration Gradients Associated with Acquired Drug Resistance in Patients with Tuberculosis. *Am J Respir Crit Care Med* **2018,** *198* (9), 1208-1219.

108. Ye, M.; Yuan, W.; Molaeipour, L.; Azizian, K.; Ahmadi, A.; Kouhsari, E., Antibiotic heteroresistance in Mycobacterium tuberculosis isolates: a systematic review and meta-analysis. *Ann Clin Microbiol Antimicrob* **2021,** *20* (1), 73.

109. World Health Organization (WHO) *Rapid communication: Key changes to the treatment of drug-resistant tuberculosis*; 2022.

110. World Health Organization (WHO), Global Tuberculosis Report 2020. **2020,** *Licence: CC BY-NC-SA 3.0 IGO.*

111. World Health Organization (WHO). The End TB Strategy: global strategy and targets for tuberculosis prevention, care and control after 2015 2015. http://www.who.int/tb /strategy/End_TB_Strategy.pdf?ua=1.

112. World Health Organization (WHO). Implementing the End TB Strategy: the essentials 2015. https://www.who.int/tb/publications/2015/end_tb_essential.pdf.

113. Organization, W. H. *Technical manual for drug susceptibility testing of medicines used in the treatment of tuberculosis*; 9789241514842; World Health Organization: Geneva, 2018, 2018.

114. Canetti, G.; Froman, S.; Grosset, J.; Hauduroy, P.; Langerova, M.; Mahler, H. T.; Meissner, G.; Mitchison, D. A.; Sula, L., MYCOBACTERIA: LABORATORY METHODS FOR TESTING DRUG SENSITIVITY AND RESISTANCE. *Bull World Health Organ* **1963,** *29* (5), 565-78.

115. World Health Organization (WHO), Technical Report on critical concentrations for drug susceptibility testing of medicines used in the treatment of drug-resistant tuberculosis. **2018**.

116. World Health Organization (WHO), Technical report on critical concentrations for drug susceptibility testing of isoniazid and the rifamycins (rifampicin, rifabutin and rifapentine). **2021**.

117. Kontos, F.; Maniati, M.; Costopoulos, C.; Gitti, Z.; Nicolaou, S.; Petinaki, E.; Anagnostou, S.; Tselentis, I.; Maniatis, A. N., Evaluation of the fully automated Bactec MGIT 960 system for the susceptibility testing of Mycobacterium tuberculosis to first-line drugs: a multicenter study. *Journal of microbiological methods* **2004,** *56* (2), 291-294.

118. Scarparo, C.; Ricordi, P.; Ruggiero, G.; Piccoli, P., Evaluation of the fully automated BACTEC MGIT 960 system for testing susceptibility of Mycobacterium tuberculosis to pyrazinamide, streptomycin, isoniazid, rifampin, and ethambutol and comparison with the radiometric BACTEC 460TB method. *J Clin Microbiol* **2004,** *42* (3), 1109-14.

119. Tortoli, E.; Cichero, P.; Piersimoni, C.; Simonetti, M. T.; Gesu, G.; Nista, D., Use of BACTEC MGIT 960 for recovery of mycobacteria from clinical specimens: multicenter study. *J Clin Microbiol* **1999,** *37* (11), 3578-82.

120. Lawson, L.; Emenyonu, N.; Abdurrahman, S. T.; Lawson, J. O.; Uzoewulu, G. N.; Sogaolu, O. M.; Ebisike, J. N.; Parry, C. M.; Yassin, M. A.; Cuevas, L. E., Comparison of Mycobacterium tuberculosis drug susceptibility using solid and liquid culture in Nigeria. *BMC Res Notes* **2013,** *6*, 215.

121. Ardito, F.; Posteraro, B.; Sanguinetti, M.; Zanetti, S.; Fadda, G., Evaluation of BACTEC Mycobacteria Growth Indicator Tube (MGIT 960) automated system for drug susceptibility testing of Mycobacterium tuberculosis. *J Clin Microbiol* **2001,** *39* (12), 4440-4.

122. Schön, T.; Werngren, J.; Machado, D.; Borroni, E.; Wijkander, M.; Lina, G.; Mouton, J.; Matuschek, E.; Kahlmeter, G.; Giske, C.; Santin, M.; Cirillo, D. M.; Viveiros, M.; Cambau, E., Multicentre testing of the EUCAST broth microdilution reference method for MIC determination on Mycobacterium tuberculosis. *Clin Microbiol Infect* **2021,** *27* (2), 288.e1-288.e4.

123. Schön, T.; Werngren, J.; Machado, D.; Borroni, E.; Wijkander, M.; Lina, G.; Mouton, J.; Matuschek, E.; Kahlmeter, G.; Giske, C.; Santin, M.; Cirillo, D. M.; Viveiros, M.; Cambau, E., Antimicrobial susceptibility testing of Mycobacterium tuberculosis complex isolates - the EUCAST broth microdilution reference method for MIC determination. *Clin Microbiol Infect* **2020,** *26* (11), 1488-1492.

124. Kunte, S.; Karmarkar, A.; Dharmashale, S.; Hatolkar, S., Resazurin microtitre assay (REMA) plate - A simple, rapid and inexpensive method for detection of drug resistance in Mycobacterium tuberculosis. *European Respiratory Journal* **2012,** *40* (Suppl 56), P1412.

125. Lee, J.; Armstrong, D. T.; Ssengooba, W.; Park, J. A.; Yu, Y.; Mumbowa, F.; Namaganda, C.; Mboowa, G.; Nakayita, G.; Armakovitch, S.; Chien, G.; Cho, S. N.; Via, L. E.; Barry, C. E., 3rd; Ellner, J. J.; Alland, D.; Dorman, S. E.; Joloba, M. L., Sensititre MYCOTB MIC plate for testing Mycobacterium tuberculosis susceptibility to first- and second-line drugs. *Antimicrob Agents Chemother* **2014,** *58* (1), 11-8.

126. Rancoita, P. M. V.; Cugnata, F.; Gibertoni Cruz, A. L.; Borroni, E.; Hoosdally, S. J.; Walker, T. M.; Grazian, C.; Davies, T. J.; Peto, T. E. A.; Crook, D. W.; Fowler, P. W.; Cirillo,

D. M.; The CRyPTIC Consortium, Validating a 14-Drug Microtiter Plate Containing Bedaquiline and Delamanid for Large-Scale Research Susceptibility Testing of Mycobacterium tuberculosis. *Antimicrob Agents Chemother* **2018,** *62* (9).

127. Somoskovi, A.; Clobridge, A.; Larsen, S. C.; Sinyavskiy, O.; Surucuoglu, S.; Parsons, L. M.; Salfinger, M., Does the MGIT 960 system improve the turnaround times for growth detection and susceptibility testing of the Mycobacterium tuberculosis complex? *J Clin Microbiol* **2006,** *44* (6), 2314-5.

128. Steingart, K. R.; Schiller, I.; Horne, D. J.; Pai, M.; Boehme, C. C.; Dendukuri, N., Xpert® MTB/RIF assay for pulmonary tuberculosis and rifampicin resistance in adults. *Cochrane Database Syst Rev* **2014,** *2014* (1), Cd009593.

129. Sachdeva, K. S.; Raizada, N.; Sreenivas, A.; Van't Hoog, A. H.; van den Hof, S.; Dewan, P. K.; Thakur, R.; Gupta, R. S.; Kulsange, S.; Vadera, B.; Babre, A.; Gray, C.; Parmar, M.; Ghedia, M.; Ramachandran, R.; Alavadi, U.; Arinaminpathy, N.; Denkinger, C.; Boehme, C.; Paramasivan, C. N., Use of Xpert MTB/RIF in Decentralized Public Health Settings and Its Effect on Pulmonary TB and DR-TB Case Finding in India. *PloS one* **2015,** *10* (5), e0126065.

130. Lawn, S. D.; Nicol, M. P., Xpert® MTB/RIF assay: development, evaluation and implementation of a new rapid molecular diagnostic for tuberculosis and rifampicin resistance. *Future Microbiol* **2011,** *6* (9), 1067-82.

131. Cao, Y.; Parmar, H.; Gaur, R. L.; Lieu, D.; Raghunath, S.; Via, N.; Battaglia, S.; Cirillo, D. M.; Denkinger, C.; Georghiou, S.; Kwiatkowski, R.; Persing, D.; Alland, D.; Chakravorty, S., Xpert MTB/XDR: a 10-Color Reflex Assay Suitable for Point-of-Care Settings To Detect Isoniazid, Fluoroquinolone, and Second-Line-Injectable-Drug Resistance Directly from Mycobacterium tuberculosis-Positive Sputum. *J Clin Microbiol* **2021,** *59* (3).

132. Sanchez-Padilla, E.; Merker, M.; Beckert, P.; Jochims, F.; Dlamini, T.; Kahn, P.; Bonnet, M.; Niemann, S., Detection of drug-resistant tuberculosis by Xpert MTB/RIF in Swaziland. *N Engl J Med* **2015,** *372* (12), 1181-2.

133. Makhado, N. A.; Matabane, E.; Faccin, M.; Pincon, C.; Jouet, A.; Boutachkourt, F.; Goeminne, L.; Gaudin, C.; Maphalala, G.; Beckert, P.; Niemann, S.; Delvenne, J. C.; Delmee, M.; Razwiedani, L.; Nchabeleng, M.; Supply, P.; de Jong, B. C.; Andre, E., Outbreak of multidrug-resistant tuberculosis in South Africa undetected by WHO-endorsed commercial tests: an observational study. *Lancet Infect Dis* **2018,** *18* (12), 1350-1359.

134. Wasserman, S.; Furin, J., Clarity with INHindsight: High-Dose Isoniazid for Drug-Resistant Tuberculosis with inhA Mutations. *Am J Respir Crit Care Med* **2020,** *201* (11), 1331-1333.

135. Ajileye, A.; Alvarez, N.; Merker, M.; Walker, T. M.; Akter, S.; Brown, K.; Moradigaravand, D.; Schon, T.; Andres, S.; Schleusener, V.; Omar, S. V.; Coll, F.; Huang, H.; Diel, R.; Ismail, N.; Parkhill, J.; de Jong, B. C.; Peto, T. E.; Crook, D. W.; Niemann, S.; Robledo, J.; Smith, E. G.; Peacock, S. J.; Koser, C. U., Some Synonymous and Nonsynonymous gyrA Mutations in Mycobacterium tuberculosis Lead to Systematic False-Positive Fluoroquinolone Resistance Results with the Hain GenoType MTBDRsl Assays. *Antimicrob Agents Chemother* **2017,** *61* (4).

136. Aubry, A.; Sougakoff, W.; Bodzongo, P.; Delcroix, G.; Armand, S.; Millot, G.; Jarlier, V.; Courcol, R.; Lemaître, N., First evaluation of drug-resistant Mycobacterium tuberculosis clinical isolates from Congo revealed misdetection of fluoroquinolone resistance by line probe assay due to a double substitution T80A-A90G in GyrA. *PloS one* **2014,** *9* (4), e95083.

137. Gillespie, S. H., Evolution of drug resistance in Mycobacterium tuberculosis: clinical and molecular perspective. *Antimicrob Agents Chemother* **2002,** *46* (2), 267-74.

138. Meehan, C. J.; Goig, G. A.; Kohl, T. A.; Verboven, L.; Dippenaar, A.; Ezewudo, M.; Farhat, M. R.; Guthrie, J. L.; Laukens, K.; Miotto, P.; Ofori-Anyinam, B.; Dreyer, V.; Supply, P.; Suresh, A.; Utpatel, C.; van Soolingen, D.; Zhou, Y.; Ashton, P. M.; Brites, D.; Cabibbe, A. M.; de Jong, B. C.; de Vos, M.; Menardo, F.; Gagneux, S.; Gao, Q.; Heupink, T. H.; Liu, Q.; Loiseau, C.; Rigouts, L.; Rodwell, T. C.; Tagliani, E.; Walker, T. M.; Warren, R. M.; Zhao, Y.; Zignol, M.; Schito, M.; Gardy, J.; Cirillo, D. M.; Niemann, S.; Comas, I.; Van Rie, A., Whole genome sequencing of Mycobacterium tuberculosis: current standards and open issues. *Nat Rev Microbiol* **2019,** *17* (9), 533-545.

139. Marin, M.; Vargas, R.; Harris, M.; Jeffrey, B.; Epperson, L. E.; Durbin, D.; Strong, M.; Salfinger, M.; Iqbal, Z.; Akhundova, I.; Vashakidze, S.; Crudu, V.; Rosenthal, A.; Farhat, M. R., Benchmarking the empirical accuracy of short-read sequencing across the *M. tuberculosis* genome. *Bioinformatics* **2022,** *38* (7), 1781-7.

140. World Health Organization (WHO), The use of next-generation sequencing technologies for the detection of mutations associated with drug resistance in Mycobacterium tuberculosis complex: technical guide. **2018,** *WHO/CDS/TB/2018.19*.

141. Mitchell, K.; Brito, J. J.; Mandric, I.; Wu, Q.; Knyazev, S.; Chang, S.; Martin, L. S.; Karlsberg, A.; Gerasimov, E.; Littman, R.; Hill, B. L.; Wu, N. C.; Yang, H. T.; Hsieh, K.; Chen, L.; Littman, E.; Shabani, T.; Enik, G.; Yao, D.; Sun, R.; Schroeder, J.; Eskin, E.; Zelikovsky, A.; Skums, P.; Pop, M.; Mangul, S., Benchmarking of computational error-correction methods for next-generation sequencing data. *Genome Biology* **2020,** *21* (1), 71.

142. Kubica, G. P.; Kim, T. H.; Dunbar, F. P., Designation of Strain H37Rv as the Neotype of Mycobacterium tuberculosis. *International Journal of Systematic and Evolutionary Microbiology* **1972,** *22* (2), 99-106.

143. Miotto, P.; Tessema, B.; Tagliani, E.; Chindelevitch, L.; Starks, A. M.; Emerson, C.; Hanna, D.; Kim, P. S.; Liwski, R.; Zignol, M.; Gilpin, C.; Niemann, S.; Denkinger, C. M.; Fleming, J.; Warren, R. M.; Crook, D.; Posey, J.; Gagneux, S.; Hoffner, S.; Rodrigues, C.; Comas, I.; Engelthaler, D. M.; Murray, M.; Alland, D.; Rigouts, L.; Lange, C.; Dheda, K.; Hasan, R.; Ranganathan, U. D. K.; McNerney, R.; Ezewudo, M.; Cirillo, D. M.; Schito, M.; Koser, C. U.; Rodwell, T. C., A standardised method for interpreting the association between mutations and phenotypic drug resistance in Mycobacterium tuberculosis. *Eur Respir J* **2017,** *50* (6).

144. The CRyPTIC Consortium; The Genomes Project; Allix-Beguec, C.; Arandjelovic, I.; Bi, L.; Beckert, P.; Bonnet, M.; Bradley, P.; Cabibbe, A. M.; Cancino-Munoz, I.; Caulfield, M. J.; Chaiprasert, A.; Cirillo, D. M.; Clifton, D. A.; Comas, I.; Crook, D. W.; De Filippo, M. R.; de Neeling, H.; Diel, R.; Drobniewski, F. A.; Faksri, K.; Farhat, M. R.; Fleming, J.; Fowler, P.; Fowler, T. A.; Gao, Q.; Gardy, J.; Gascoyne-Binzi, D.; Gibertoni-Cruz, A. L.; Gil-

Brusola, A.; Golubchik, T.; Gonzalo, X.; Grandjean, L.; He, G.; Guthrie, J. L.; Hoosdally, S.; Hunt, M.; Iqbal, Z.; Ismail, N.; Johnston, J.; Khanzada, F. M.; Khor, C. C.; Kohl, T. A.; Kong, C.; Lipworth, S.; Liu, Q.; Maphalala, G.; Martinez, E.; Mathys, V.; Merker, M.; Miotto, P.; Mistry, N.; Moore, D. A. J.; Murray, M.; Niemann, S.; Omar, S. V.; Ong, R. T.; Peto, T. E. A.; Posey, J. E.; Prammananan, T.; Pym, A.; Rodrigues, C.; Rodrigues, M.; Rodwell, T.; Rossolini, G. M.; Sanchez Padilla, E.; Schito, M.; Shen, X.; Shendure, J.; Sintchenko, V.; Sloutsky, A.; Smith, E. G.; Snyder, M.; Soetaert, K.; Starks, A. M.; Supply, P.; Suriyapol, P.; Tahseen, S.; Tang, P.; Teo, Y. Y.; Thuong, T. N. T.; Thwaites, G.; Tortoli, E.; van Soolingen, D.; Walker, A. S.; Walker, T. M.; Wilcox, M.; Wilson, D. J.; Wyllie, D.; Yang, Y.; Zhang, H.; Zhao, Y.; Zhu, B., Prediction of Susceptibility to First-Line Tuberculosis Drugs by DNA Sequencing. *N Engl J Med* **2018,** *379* (15), 1403-1415.

145. World Health Organization (WHO), Catalogue of mutations in Mycobacterium tuberculosis complex and their association with drug resistance. **2021,** *License: CC BY-NC-SA 3.0 IGO.*

146. Cohen, K. A.; Manson, A. L.; Desjardins, C. A.; Abeel, T.; Earl, A. M., Deciphering drug resistance in Mycobacterium tuberculosis using whole-genome sequencing: progress, promise, and challenges. *Genome Medicine* **2019,** *11* (1), 45.

147. Olaru, I. D.; Patel, H.; Kranzer, K.; Perera, N., Turnaround time of whole genome sequencing for mycobacterial identification and drug susceptibility testing in routine practice. *Clinical Microbiology and Infection* **2018,** *24* (6), 659.e5-659.e7.

148. Walker, T. M.; Cruz, A. L. G.; Peto, T. E.; Smith, E. G.; Esmail, H.; Crook, D. W., Tuberculosis is changing. *Lancet Infect Dis* **2017,** *17* (4), 359-361.

149. Sanchini, A.; Jandrasits, C.; Tembrockhaus, J.; Kohl, T. A.; Utpatel, C.; Maurer, F. P.; Niemann, S.; Haas, W.; Renard, B. Y.; Kröger, S., Improving tuberculosis surveillance by detecting international transmission using publicly available whole genome sequencing data. *Eurosurveillance* **2021,** *26* (2), 1900677.

150. Cohen, K. A.; Manson, A. L.; Abeel, T.; Desjardins, C. A.; Chapman, S. B.; Hoffner, S.; Birren, B. W.; Earl, A. M., Extensive global movement of multidrug-resistant *M. tuberculosis* strains revealed by whole-genome analysis. *Thorax* **2019,** *74* (9), 882-889.

151. Dookie, N.; Khan, A.; Padayatchi, N.; Naidoo, K., Application of Next Generation Sequencing for Diagnosis and Clinical Management of Drug-Resistant Tuberculosis: Updates on Recent Developments in the Field. *Frontiers in Microbiology* **2022,** *13*.

152. McNerney, R.; Clark, T. G.; Campino, S.; Rodrigues, C.; Dolinger, D.; Smith, L.; Cabibbe, A. M.; Dheda, K.; Schito, M., Removing the bottleneck in whole genome sequencing of Mycobacterium tuberculosis for rapid drug resistance analysis: a call to action. *Int J Infect Dis* **2017,** *56*, 130-135.

153. Doyle, R. M.; Burgess, C.; Williams, R.; Gorton, R.; Booth, H.; Brown, J.; Bryant, J. M.; Chan, J.; Creer, D.; Holdstock, J.; Kunst, H.; Lozewicz, S.; Platt, G.; Romero, E. Y.; Speight, G.; Tiberi, S.; Abubakar, I.; Lipman, M.; McHugh, T. D.; Breuer, J.; Mellmann, A., Direct Whole-Genome Sequencing of Sputum Accurately Identifies Drug-Resistant Mycobacterium tuberculosis Faster than MGIT Culture Sequencing. *Journal of Clinical Microbiology* **2018,** *56* (8), e00666-18.

154. Colman, R. E.; Anderson, J.; Lemmer, D.; Lehmkuhl, E.; Georghiou, S. B.; Heaton, H.; Wiggins, K.; Gillece, J. D.; Schupp, J. M.; Catanzaro, D. G.; Crudu, V.; Cohen, T.; Rodwell, T. C.; Engelthaler, D. M., Rapid Drug Susceptibility Testing of Drug-Resistant Mycobacterium tuberculosis Isolates Directly from Clinical Samples by Use of Amplicon Sequencing: a Proof-of-Concept Study. *J Clin Microbiol* **2016,** *54* (8), 2058-67.

155. Kambli, P.; Ajbani, K.; Kazi, M.; Sadani, M.; Naik, S.; Shetty, A.; Tornheim, J. A.; Singh, H.; Rodrigues, C., Targeted next generation sequencing directly from sputum for comprehensive genetic information on drug resistant Mycobacterium tuberculosis. *Tuberculosis (Edinb)* **2021,** *127*, 102051.

156. Cabibbe, A. M.; Spitaleri, A.; Battaglia, S.; Colman, R. E.; Suresh, A.; Uplekar, S.; Rodwell, T. C.; Cirillo, D. M., Application of targeted Next Generation Sequencing assay on a portable sequencing platform for culture-free detection of drug resistant tuberculosis from clinical samples. *Journal of Clinical Microbiology* **2020**, JCM.00632-20.

157. Colman, R. E.; Mace, A.; Seifert, M.; Hetzel, J.; Mshaiel, H.; Suresh, A.; Lemmer, D.; Engelthaler, D. M.; Catanzaro, D. G.; Young, A. G.; Denkinger, C. M.; Rodwell, T. C., Whole-genome and targeted sequencing of drug-resistant Mycobacterium tuberculosis on the iSeq100 and MiSeq: A performance, ease-of-use, and cost evaluation. *PLoS Med* **2019,** *16* (4), e1002794.

158. Hall, M. B.; Coin, L. J. M., Assessment of the 2021 WHO Mycobacterium tuberculosis drug resistance mutation catalogue on an independent dataset. *The Lancet Microbe* **2022**.

159. Takahashi, H.; Hayakawa, I.; Akimoto, T., The history of the development and changes of quinolone antibacterial agents. *Yakushigaku Zasshi* **2003,** *38* (2), 161-79.

160. Fedorowicz, J.; Sączewski, J., Modifications of quinolones and fluoroquinolones: hybrid compounds and dual-action molecules. *Monatshefte für Chemie - Chemical Monthly* **2018,** *149* (7), 1199-1245.

161. Bergwerf, H. MolView: An attempt to Get the Cloud into Chemistry Classrooms. https://confchem.ccce.divched.org/2015FallCCCENLP9 (accessed 6th July).

162. Andriole, V. T., The Quinolones: Past, Present, and Future. *Clinical Infectious Diseases* **2005,** *41* (Supplement_2), S113-S119.

163. Kabbani, S.; Hersh, A. L.; Shapiro, D. J.; Fleming-Dutra, K. E.; Pavia, A. T.; Hicks, L. A., Opportunities to Improve Fluoroquinolone Prescribing in the United States for Adult Ambulatory Care Visits. *Clinical Infectious Diseases* **2018,** *67* (1), 134-136.

164. Aditi Sriram, E. K., Geetanjali Kapoor, Jessica Craig, Ruchita Balasubramanian, Sehr Brar, Nicola Criscuolo, Alisa Hamilton, Eili Klein, Katie Tseng, Thomas P Van Boeckel, Ramanan Laxminarayan The State of the World's Antibiotics in 2021. https://cddep.org/publications/the-state-of-the-worlds-antibiotic-in-2021/ (accessed 11 May 2022).

165. Browne, A. J.; Chipeta, M. G.; Haines-Woodhouse, G.; Kumaran, E. P. A.; Hamadani, B. H. K.; Zaraa, S.; Henry, N. J.; Deshpande, A.; Reiner, R. C., Jr.; Day, N. P. J.; Lopez, A. D.; Dunachie, S.; Moore, C. E.; Stergachis, A.; Hay, S. I.; Dolecek, C., Global

antibiotic consumption and usage in humans, 2000–18: a spatial modelling study. *The Lancet Planetary Health* **2021,** *5* (12), e893-e904.

166. Yang, P.; Chen, Y.; Jiang, S.; Shen, P.; Lu, X.; Xiao, Y., Association between the rate of fluoroquinolones-resistant gram-negative bacteria and antibiotic consumption from China based on 145 tertiary hospitals data in 2014. *BMC infectious diseases* **2020,** *20* (1), 269.

167. Adam, H. J.; Hoban, D. J.; Gin, A. S.; Zhanel, G. G., Association between fluoroquinolone usage and a dramatic rise in ciprofloxacin-resistant Streptococcus pneumoniae in Canada, 1997-2006. *Int J Antimicrob Agents* **2009,** *34* (1), 82-5.

168. Terahara, F.; Nishiura, H., Fluoroquinolone consumption and Escherichia coli resistance in Japan: an ecological study. *BMC Public Health* **2019,** *19* (1), 426.

169. Kenyon, C., Positive Association between the Use of Quinolones in Food Animals and the Prevalence of Fluoroquinolone Resistance in E. coli and K. pneumoniae, A. baumannii and P. aeruginosa: A Global Ecological Analysis. *Antibiotics* **2021,** *10* (10), 1193.

170. Hu, Y.; Coates, A. R.; Mitchison, D. A., Sterilizing activities of fluoroquinolones against rifampin-tolerant populations of Mycobacterium tuberculosis. *Antimicrob Agents Chemother* **2003,** *47* (2), 653-7.

171. Alvirez-Freites, E. J.; Carter, J. L.; Cynamon, M. H., In vitro and in vivo activities of gatifloxacin against Mycobacterium tuberculosis. *Antimicrob Agents Chemother* **2002,** *46* (4), 1022-5.

172. Zhao, B. Y.; Pine, R.; Domagala, J.; Drlica, K., Fluoroquinolone action against clinical isolates of Mycobacterium tuberculosis: effects of a C-8 methoxyl group on survival in liquid media and in human macrophages. *Antimicrob Agents Chemother* **1999,** *43* (3), 661-6.

173. Aubry, A.; Pan, X. S.; Fisher, L. M.; Jarlier, V.; Cambau, E., Mycobacterium tuberculosis DNA gyrase: interaction with quinolones and correlation with antimycobacterial drug activity. *Antimicrob Agents Chemother* **2004,** *48* (4), 1281-8.

174. Bastos, M. L.; Lan, Z.; Menzies, D., An updated systematic review and meta-analysis for treatment of multidrug-resistant tuberculosis. *European Respiratory Journal* **2017,** *49* (3), 1600803.

175. Altin, T.; Ozcan, O.; Turhan, S.; Ongun Ozdemir, A.; Akyurek, O.; Karaoguz, R.; Guldal, M., Torsade de pointes associated with moxifloxacin: a rare but potentially fatal adverse event. *Can J Cardiol* **2007,** *23* (11), 907-8.

176. Täubel, J.; Prasad, K.; Rosano, G.; Ferber, G.; Wibberley, H.; Cole, S. T.; Van Langenhoven, L.; Fernandes, S.; Djumanov, D.; Sugiyama, A., Effects of the Fluoroquinolones Moxifloxacin and Levofloxacin on the QT Subintervals: Sex Differences in Ventricular Repolarization. *J Clin Pharmacol* **2020,** *60* (3), 400-408.

177. Park-Wyllie, L. Y.; Juurlink, D. N.; Kopp, A.; Shah, B. R.; Stukel, T. A.; Stumpo, C.; Dresser, L.; Low, D. E.; Mamdani, M. M., Outpatient gatifloxacin therapy and dysglycemia in older adults. *N Engl J Med* **2006,** *354* (13), 1352-61.

178. Koh, W. J.; Lee, S. H.; Kang, Y. A.; Lee, C. H.; Choi, J. C.; Lee, J. H.; Jang, S. H.; Yoo, K. H.; Jung, K. H.; Kim, K. U.; Choi, S. B.; Ryu, Y. J.; Chan Kim, K.; Um, S.; Kwon, Y. S.; Kim, Y. H.; Choi, W. I.; Jeon, K.; Hwang, Y. I.; Kim, S. J.; Lee, Y. S.; Heo, E. Y.; Lee, J.; Ki, Y. W.; Shim, T. S.; Yim, J. J., Comparison of levofloxacin versus moxifloxacin for multidrug-resistant tuberculosis. *Am J Respir Crit Care Med* **2013,** *188* (7), 858-64.

179. Johnson, J. L.; Hadad, D. J.; Boom, W. H.; Daley, C. L.; Peloquin, C. A.; Eisenach, K. D.; Jankus, D. D.; Debanne, S. M.; Charlebois, E. D.; Maciel, E.; Palaci, M.; Dietze, R., Early and extended early bactericidal activity of levofloxacin, gatifloxacin and moxifloxacin in pulmonary tuberculosis. *Int J Tuberc Lung Dis* **2006,** *10* (6), 605-12.

180. Pienaar, E.; Sarathy, J.; Prideaux, B.; Dietzold, J.; Dartois, V.; Kirschner, D. E.; Linderman, J. J., Comparing efficacies of moxifloxacin, levofloxacin and gatifloxacin in tuberculosis granulomas using a multi-scale systems pharmacology approach. *PLoS Comput Biol* **2017,** *13* (8), e1005650.

181. Collin, F.; Karkare, S.; Maxwell, A., Exploiting bacterial DNA gyrase as a drug target: current state and perspectives. *Appl Microbiol Biotechnol* **2011,** *92* (3), 479-97.

182. Schmidt, B. H.; Burgin, A. B.; Deweese, J. E.; Osheroff, N.; Berger, J. M., A novel and unified two-metal mechanism for DNA cleavage by type II and IA topoisomerases. *Nature* **2010,** *465* (7298), 641-4.

183. Sissi, C.; Chemello, A.; Vazquez, E.; Mitchenall, L. A.; Maxwell, A.; Palumbo, M., DNA gyrase requires DNA for effective two-site coordination of divalent metal ions: further insight into the mechanism of enzyme action. *Biochemistry* **2008,** *47* (33), 8538-45.

184. Blower, T. R.; Williamson, B. H.; Kerns, R. J.; Berger, J. M., Crystal structure and stability of gyrase-fluoroquinolone cleaved complexes from Mycobacterium tuberculosis. *Proc Natl Acad Sci U S A* **2016,** *113* (7), 1706-13.

185. Aldred, K. J.; Kerns, R. J.; Osheroff, N., Mechanism of quinolone action and resistance. *Biochemistry* **2014,** *53* (10), 1565-74.

186. Hooper, D. C., Mode of action of fluoroquinolones. *Drugs* **1999,** *58 Suppl 2*, 6-10.

187. Humphrey, W.; Dalke, A.; Schulten, K., VMD: visual molecular dynamics. *J Mol Graph* **1996,** *14* (1), 33-8, 27-8.

188. Che, Y.; Song, Q.; Yang, T.; Ping, G.; Yu, M., Fluoroquinolone resistance in multidrug-resistant Mycobacterium tuberculosis independent of fluoroquinolone use. *European Respiratory Journal* **2017,** *50* (6), 1701633.

189. Shakoor, S.; Tahseen, S.; Jabeen, K.; Fatima, R.; Malik, F. R.; Rizvi, A. H.; Hasan, R., Fluoroquinolone consumption and -resistance trends in Mycobacterium tuberculosis and other respiratory pathogens: Ecological antibiotic pressure and consequences in Pakistan, 2009–2015. *International journal of mycobacteriology* **2016,** *5* (4), 412-416.

190. Migliori, G. B.; Langendam, M. W.; D'Ambrosio, L.; Centis, R.; Blasi, F.; Huitric, E.; Manissero, D.; van der Werf, M. J., Protecting the tuberculosis drug pipeline: stating the case for the rational use of fluoroquinolones. *European Respiratory Journal* **2012,** *40* (4), 814-822.

191. Gandhi, N. R.; Moll, A.; Sturm, A. W.; Pawinski, R.; Govender, T.; Lalloo, U.; Zeller, K.; Andrews, J.; Friedland, G., Extensively drug-resistant tuberculosis as a cause of death in patients co-infected with tuberculosis and HIV in a rural area of South Africa. *Lancet* **2006,** *368* (9547), 1575-80.

192. Sharma, R.; Singh, B. K.; Kumar, P.; Ramachandran, R.; Jorwal, P., Presence of Fluoroquinolone mono-resistance among drug-sensitive Mycobacterium tuberculosis isolates: An alarming trend and implications. *Clinical Epidemiology and Global Health* **2019,** *7* (3), 363-366.

193. Sharma, R.; Sharma, S. K.; Singh, B. K.; Mittal, A.; Kumar, P., High degree of fluoroquinolone resistance among pulmonary tuberculosis patients in New Delhi, India. *Indian J Med Res* **2019,** *149* (1), 62-66.

194. Xu, P.; Li, X.; Zhao, M.; Gui, X.; DeRiemer, K.; Gagneux, S.; Mei, J.; Gao, Q., Prevalence of fluoroquinolone resistance among tuberculosis patients in Shanghai, China. *Antimicrob Agents Chemother* **2009,** *53* (7), 3170-2.

195. Malik, S.; Willby, M.; Sikes, D.; Tsodikov, O. V.; Posey, J. E., New insights into fluoroquinolone resistance in Mycobacterium tuberculosis: functional genetic analysis of gyrA and gyrB mutations. *PloS one* **2012,** *7* (6), e39754.

196. Aubry, A.; Veziris, N.; Cambau, E.; Truffot-Pernot, C.; Jarlier, V.; Fisher, L. M., Novel gyrase mutations in quinolone-resistant and -hypersusceptible clinical isolates of Mycobacterium tuberculosis: functional analysis of mutant enzymes. *Antimicrob Agents Chemother* **2006,** *50* (1), 104-12.

197. Pantel, A.; Petrella, S.; Veziris, N.; Brossier, F.; Bastian, S.; Jarlier, V.; Mayer, C.; Aubry, A., Extending the definition of the GyrB quinolone resistance-determining region in Mycobacterium tuberculosis DNA gyrase for assessing fluoroquinolone resistance in *M. tuberculosis*. *Antimicrob Agents Chemother* **2012,** *56* (4), 1990-6.

198. Takiff, H. E.; Salazar, L.; Guerrero, C.; Philipp, W.; Huang, W. M.; Kreiswirth, B.; Cole, S. T.; Jacobs, W. R., Jr.; Telenti, A., Cloning and nucleotide sequence of Mycobacterium tuberculosis gyrA and gyrB genes and detection of quinolone resistance mutations. *Antimicrobial agents and chemotherapy* **1994,** *38* (4), 773-780.

199. Malik, S.; Willby, M.; Sikes, D.; Tsodikov, O. V.; Posey, J. E., New Insights into Fluoroquinolone Resistance in Mycobacterium tuberculosis: Functional Genetic Analysis of gyrA and gyrB Mutations. *PLoS One. 2012; 7(6).* **2012,** *7* (6).

200. Farhat, M. R.; Sultana, R.; Iartchouk, O.; Bozeman, S.; Galagan, J.; Sisk, P.; Stolte, C.; Nebenzahl-Guimaraes, H.; Jacobson, K.; Sloutsky, A.; Kaur, D.; Posey, J.; Kreiswirth, B. N.; Kurepina, N.; Rigouts, L.; Streicher, E. M.; Victor, T. C.; Warren, R. M.; Soolingen, D. v.; Murray, M., Genetic Determinants of Drug Resistance in Mycobacterium tuberculosis and Their Diagnostic Value. *American Journal of Respiratory and Critical Care Medicine* **2016,** *194* (5), 621-630.

201. Aldred Katie, J.; Blower Tim, R.; Kerns Robert, J.; Berger James, M.; Osheroff, N., Fluoroquinolone interactions with Mycobacterium tuberculosis gyrase: Enhancing drug

activity against wild-type and resistant gyrase. *Proceedings of the National Academy of Sciences* **2016,** *113* (7), E839-E846.

202. Singhal, R.; Reynolds, P. R.; Marola, J. L.; Epperson, L. E.; Arora, J.; Sarin, R.; Myneedu, V. P.; Strong, M.; Salfinger, M., Sequence Analysis of Fluoroquinolone Resistance-Associated Genes gyrA and gyrB in Clinical Mycobacterium tuberculosis Isolates from Patients Suspected of Having Multidrug-Resistant Tuberculosis in New Delhi, India. *J Clin Microbiol* **2016,** *54* (9), 2298-305.

203. Farhat, M. R.; Jacobson, K. R.; Franke, M. F.; Kaur, D.; Sloutsky, A.; Mitnick, C. D.; Murray, M., Gyrase Mutations Are Associated with Variable Levels of Fluoroquinolone Resistance in Mycobacterium tuberculosis. *J Clin Microbiol* **2016,** *54* (3), 727-33.

204. Disratthakit, A.; Prammananan, T.; Tribuddharat, C.; Thaipisuttikul, I.; Doi, N.; Leechawengwongs, M.; Chaiprasert, A., Role of gyrB Mutations in Pre-extensively and Extensively Drug-Resistant Tuberculosis in Thai Clinical Isolates. *Antimicrob Agents Chemother* **2016,** *60* (9), 5189-97.

205. The CRyPTIC Consortium; Earle, S. G.; Wilson, D. J., Genome-wide association studies of global Mycobacterium tuberculosis resistance to thirteen antimicrobials in 10,228 genomes. *bioRxiv* **2021**, 2021.09.14.460272.

206. Singh, M.; Jadaun, G. P. S.; Ramdas; Srivastava, K.; Chauhan, V.; Mishra, R.; Gupta, K.; Nair, S.; Chauhan, D. S.; Sharma, V. D.; Venkatesan, K.; Katoch, V. M., Effect of efflux pump inhibitors on drug susceptibility of ofloxacin resistant Mycobacterium tuberculosis isolates. *Indian Journal of Medical Research* **2011,** *133* (5).

207. Escribano, I.; Rodríguez, J. C.; Llorca, B.; García-Pachon, E.; Ruiz, M.; Royo, G., Importance of the Efflux Pump Systems in the Resistance of Mycobacterium tuberculosis to Fluoroquinolones and Linezolid. *Chemotherapy* **2007,** *53* (6), 397-401.

208. Maringolo-Ribeiro, C.; Grecco, J. A.; Bellato, D. L.; Almeida, A. L.; Baldin, V. P.; Caleffi-Ferracioli, K. R.; Pavan, F. R., Rescue of susceptibility to second-line drugs in resistant clinical isolates of Mycobacterium tuberculosis. *Future Microbiology* **2022,** *17* (7), 511-527.

209. Maitre, T.; Petitjean, G.; Chauffour, A.; Bernard, C.; El Helali, N.; Jarlier, V.; Reibel, F.; Chavanet, P.; Aubry, A.; Veziris, N., Are moxifloxacin and levofloxacin equally effective to treat XDR tuberculosis? *Journal of Antimicrobial Chemotherapy* **2017,** *72* (8), 2326-2333.

210. Poissy, J.; Aubry, A.; Fernandez, C.; Lott, M. C.; Chauffour, A.; Jarlier, V.; Farinotti, R.; Veziris, N., Should moxifloxacin be used for the treatment of extensively drug-resistant tuberculosis? An answer from a murine model. *Antimicrob Agents Chemother* **2010,** *54* (11), 4765-71.

211. Tornheim, J. A.; Udwadia, Z. F.; Arora, P. R.; Gajjar, I.; Sharma, S.; Karane, M.; Sawant, N.; Kharat, N.; Blum, A. J.; Shivakumar, S.; Gupte, A. N.; Gupte, N.; Mullerpattan, J. B.; Pinto, L. M.; Ashavaid, T. F.; Gupta, A.; Rodrigues, C., Increased Moxifloxacin Dosing Among Patients With Multidrug-Resistant Tuberculosis With Low-Level Resistance to Moxifloxacin Did Not Improve Treatment Outcomes in a Tertiary Care Center in Mumbai, India. *Open Forum Infect Dis* **2022,** *9* (2), ofab615.

212. Cabibbe, A. M.; Walker, T. M.; Niemann, S.; Cirillo, D. M., Whole genome sequencing of Mycobacterium tuberculosis. *Eur Respir J* **2018,** *52* (5).

213. Kim, J. I.; Maguire, F.; Tsang, K. K.; Gouliouris, T.; Peacock, S. J.; McAllister, T. A.; McArthur, A. G.; Beiko, R. G., Machine Learning for Antimicrobial Resistance Prediction: Current Practice, Limitations, and Clinical Perspective. *Clinical microbiology reviews* **2022**, e0017921.

214. Liu, L.; Li, Y.; Li, S.; Hu, N.; He, Y.; Pong, R.; Lin, D.; Lu, L.; Law, M., Comparison of Next-Generation Sequencing Systems. *Journal of Biomedicine and Biotechnology* **2012,** *2012*, 251364.

215. Breiman, L., Random Forests. *Machine Learning* **2001,** *45* (1), 5-32.

216. Chen, T.; Guestrin, C., XGBoost: A Scalable Tree Boosting System. In *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, Association for Computing Machinery: San Francisco, California, USA, 2016; pp 785–794.

217. Fan, R.-E.; Chang, K.-W.; Hsieh, C.-J.; Wang, X.-R.; Lin, C.-J., LIBLINEAR: A Library for Large Linear Classification. *J. Mach. Learn. Res.* **2008,** *9*, 1871–1874.

218. Broyden, C. G., The Convergence of a Class of Double-rank Minimization Algorithms 1. General Considerations. *IMA Journal of Applied Mathematics* **1970,** *6* (1), 76-90.

219. Fletcher, R., A new approach to variable metric algorithms. *The Computer Journal* **1970,** *13* (3), 317-322.

220. Goldfarb, D., A family of variable metric updates derived by variational means. *Mathematics of Computation* **1970,** *24* (109), 23-26.

221. Shanno, D. F., Conditioning of quasi-Newton methods for function minimization. *Mathematics of Computation* **1970,** *24* (111), 647-656.

222. Rennie, J. D. M. a. S., N., Loss Functions for Preference Levels : Regression with Discrete Ordered Labels. *Proceedings of the IJCAI Multidisciplinary Workshop on Advances in Preference Handling* **2005**.

223. Pedregosa-Izquierdo, F. Feature extraction and supervised learning on fMRI : from practice to theory

Estimation de variables et apprentissage supervisé en IRMf : de la pratique à la théorie. Université Pierre et Marie Curie - Paris VI, 2015.

224. Bahl, A.; Hellack, B.; Balas, M.; Dinischiotu, A.; Wiemann, M.; Brinkmann, J.; Luch, A.; Renard, B. Y.; Haase, A., Recursive feature elimination in random forest classification supports nanomaterial grouping. *NanoImpact* **2019,** *15*, 100179.

225. Kohavi, R.; John, G. H., Wrappers for feature subset selection. *Artificial Intelligence* **1997,** *97* (1), 273-324.

226. Miao, J.; Niu, L., A Survey on Feature Selection. *Procedia Computer Science* **2016,** *91*, 919-926.

227. Dhal, P.; Azad, C., A comprehensive survey on feature selection in the various fields of machine learning. *Applied Intelligence* **2022,** *52* (4), 4543-4581.

228. Guyon, I.; Weston, J.; Barnhill, S.; Vapnik, V., Gene Selection for Cancer Classification using Support Vector Machines. *Machine Learning* **2002,** *46* (1), 389-422.

229. Larochelle, H.; Erhan, D.; Courville, A.; Bergstra, J.; Bengio, Y., An empirical evaluation of deep architectures on problems with many factors of variation. In *Proceedings of the 24th international conference on Machine learning*, Association for Computing Machinery: Corvalis, Oregon, USA, 2007; pp 473–480.

230. Bergstra, J.; Bengio, Y., Random search for hyper-parameter optimization. *J. Mach. Learn. Res.* **2012,** *13* (null), 281–305.

231. Guvench, O.; MacKerell, A. D., Jr., Computational evaluation of protein-small molecule binding. *Curr Opin Struct Biol* **2009,** *19* (1), 56-61.

232. Kitchen, D. B.; Decornez, H.; Furr, J. R.; Bajorath, J., Docking and scoring in virtual screening for drug discovery: methods and applications. *Nat Rev Drug Discov* **2004,** *3* (11), 935-49.

233. Wang, E.; Sun, H.; Wang, J.; Wang, Z.; Liu, H.; Zhang, J. Z. H.; Hou, T., End-Point Binding Free Energy Calculation with MM/PBSA and MM/GBSA: Strategies and Applications in Drug Design. *Chemical Reviews* **2019,** *119* (16), 9478-9508.

234. Genheden, S.; Ryde, U., The MM/PBSA and MM/GBSA methods to estimate ligand-binding affinities. *Expert Opin Drug Discov* **2015,** *10* (5), 449-61.

235. Gilson, M. K.; Given, J. A.; Bush, B. L.; McCammon, J. A., The statistical-thermodynamic basis for computation of binding affinities: a critical review. *Biophys J* **1997,** *72* (3), 1047-69.

236. Mey, A. S. J. S., Allen, B. K., Bruce McDonald, H. E., Chodera, J. D., Hahn, D. F., Kuhn, M., Michel, J., Mobley, D. L., Naden, L. N., Prasad, S., Rizzi, A., Scheen, J., Shirts, M. R., Tresadern, G., & Xu, H, Best Practices for Alchemical Free Energy Calculations [Article v1.0]. *Living Journal of Computational Molecular Science* **2020,** *2* (1), 18378.

237. Klimovich, P. V.; Shirts, M. R.; Mobley, D. L., Guidelines for the analysis of free energy calculations. *J Comput Aided Mol Des* **2015,** *29* (5), 397-411.

238. Zwanzig, R. W., High-Temperature Equation of State by a Perturbation Method. I. Nonpolar Gases. *The Journal of Chemical Physics* **1954,** *22* (8), 1420-1426.

239. Bennett, C. H., Efficient estimation of free energy differences from Monte Carlo data. *Journal of Computational Physics* **1976,** *22* (2), 245-268.

240. Hornak, V.; Simmerling, C., Development of softcore potential functions for overcoming steric barriers in molecular dynamics simulations. *Journal of Molecular Graphics and Modelling* **2004,** *22* (5), 405-413.

241. Beutler, T. C.; Mark, A. E.; van Schaik, R. C.; Gerber, P. R.; van Gunsteren, W. F., Avoiding singularities and numerical instabilities in free energy calculations based on molecular simulations. *Chemical Physics Letters* **1994,** *222* (6), 529-539.

242. Woods, C. J.; Essex, J. W.; King, M. A., Enhanced Configurational Sampling in Binding Free-Energy Calculations. *The Journal of Physical Chemistry B* **2003,** *107* (49), 13711-13718.

243. Fukunishi, H.; Watanabe, O.; Takada, S., On the Hamiltonian replica exchange method for efficient sampling of biomolecular systems: Application to protein structure prediction. *The Journal of Chemical Physics* **2002,** *116* (20), 9058-9067.

244. McCammon, J. A.; Gelin, B. R.; Karplus, M., Dynamics of folded proteins. *Nature* **1977,** *267* (5612), 585-90.

245. Berendsen, H. J. C.; van der Spoel, D.; van Drunen, R., GROMACS: A message-passing parallel molecular dynamics implementation. *Computer Physics Communications* **1995,** *91* (1), 43-56.

246. D.A. Case, H. M. A., K. Belfon, I.Y. Ben-Shalom, J.T. Berryman, S.R. Brozell, D.S. Cerutti, T.E. Cheatham, III, G.A. Cisneros, V.W.D. Cruzeiro, T.A. Darden, R.E. Duke, G. Giambasu, M.K. Gilson, H. Gohlke, A.W. Goetz, R. Harris, S. Izadi, S.A. Izmailov, K. Kasavajhala, M.C. Kaymak, E. King, A. Kovalenko, T. Kurtzman, T.S. Lee, S. LeGrand, P. Li, C. Lin, J. Liu, T. Luchko, R. Luo, M. Machado, V. Man, M. Manathunga, K.M. Merz, Y. Miao, O. Mikhailovskii, G. Monard, H. Nguyen, K.A. O'Hearn, A. Onufriev, F. Pan, S. Pantano, R. Qi, A. Rahnamoun, D.R. Roe, A. Roitberg, C. Sagui, S. Schott-Verdugo, A. Shajan, J. Shen, C.L. Simmerling, N.R. Skrynnikov, J. Smith, J. Swails, R.C. Walker, J Wang, J. Wang, H. Wei, R.M. Wolf, X. Wu, Y. Xiong, Y. Xue, D.M. York, S. Zhao, and P.A. Kollman, *Amber 2022*. University of California, San Francisco: 2022.

247. Phillips, J. C.; Hardy, D. J.; Maia, J. D. C.; Stone, J. E.; Ribeiro, J. V.; Bernardi, R. C.; Buch, R.; Fiorin, G.; Hénin, J.; Jiang, W.; McGreevy, R.; Melo, M. C. R.; Radak, B. K.; Skeel, R. D.; Singharoy, A.; Wang, Y.; Roux, B.; Aksimentiev, A.; Luthey-Schulten, Z.; Kalé, L. V.; Schulten, K.; Chipot, C.; Tajkhorshid, E., Scalable molecular dynamics on CPU and GPU architectures with NAMD. *J Chem Phys* **2020,** *153* (4), 044130.

248. Brooks, B. R.; Brooks, C. L., 3rd; Mackerell, A. D., Jr.; Nilsson, L.; Petrella, R. J.; Roux, B.; Won, Y.; Archontis, G.; Bartels, C.; Boresch, S.; Caflisch, A.; Caves, L.; Cui, Q.; Dinner, A. R.; Feig, M.; Fischer, S.; Gao, J.; Hodoscek, M.; Im, W.; Kuczera, K.; Lazaridis, T.; Ma, J.; Ovchinnikov, V.; Paci, E.; Pastor, R. W.; Post, C. B.; Pu, J. Z.; Schaefer, M.; Tidor, B.; Venable, R. M.; Woodcock, H. L.; Wu, X.; Yang, W.; York, D. M.; Karplus, M., CHARMM: the biomolecular simulation program. *J Comput Chem* **2009,** *30* (10), 1545-614.

249. Bash, P. A.; Singh, U. C.; Brown, F. K.; Langridge, R.; Kollman, P. A., Calculation of the Relative Change in Binding Free Energy of a Protein-Inhibitor Complex. *Science* **1987,** *235* (4788), 574-576.

250. Hollingsworth, S. A.; Dror, R. O., Molecular Dynamics Simulation for All. *Neuron* **2018,** *99* (6), 1129-1143.

251. Cornell, W. D.; Cieplak, P.; Bayly, C. I.; Gould, I. R.; Merz, K. M.; Ferguson, D. M.; Spellmeyer, D. C.; Fox, T.; Caldwell, J. W.; Kollman, P. A., A Second Generation Force Field for the Simulation of Proteins, Nucleic Acids, and Organic Molecules. *Journal of the American Chemical Society* **1995,** *117* (19), 5179-5197.

252. Wang, J.; Wolf, R. M.; Caldwell, J. W.; Kollman, P. A.; Case, D. A., Development and testing of a general amber force field. *Journal of Computational Chemistry* **2004,** *25* (9), 1157-1174.

253. Brooks, B. R.; Bruccoleri, R. E.; Olafson, B. D.; States, D. J.; Swaminathan, S.; Karplus, M., CHARMM: A program for macromolecular energy, minimization, and dynamics calculations. *Journal of Computational Chemistry* **1983,** *4* (2), 187-217.

254. Vanommeslaeghe, K.; Hatcher, E.; Acharya, C.; Kundu, S.; Zhong, S.; Shim, J.; Darian, E.; Guvench, O.; Lopes, P.; Vorobyov, I.; Mackerell, A. D., Jr., CHARMM general force field: A force field for drug-like molecules compatible with the CHARMM all-atom additive biological force fields. *J Comput Chem* **2010,** *31* (4), 671-90.

255. Christen, M.; Hünenberger, P. H.; Bakowies, D.; Baron, R.; Bürgi, R.; Geerke, D. P.; Heinz, T. N.; Kastenholz, M. A.; Kräutler, V.; Oostenbrink, C.; Peter, C.; Trzesniak, D.; van Gunsteren, W. F., The GROMOS software for biomolecular simulation: GROMOS05. *Journal of Computational Chemistry* **2005,** *26* (16), 1719-1751.

256. Maier, J. A.; Martinez, C.; Kasavajhala, K.; Wickstrom, L.; Hauser, K. E.; Simmerling, C., ff14SB: Improving the Accuracy of Protein Side Chain and Backbone Parameters from ff99SB. *Journal of Chemical Theory and Computation* **2015,** *11* (8), 3696-3713.

257. Guvench, O.; MacKerell, A. D., Jr., Comparison of protein force fields for molecular dynamics simulations. *Methods Mol Biol* **2008,** *443*, 63-88.

258. Verlet, L., Computer "Experiments" on Classical Fluids. I. Thermodynamical Properties of Lennard-Jones Molecules. *Physical Review* **1967,** *159* (1), 98-103.

259. Swope, W. C. H., C.; Andersen, H.C.; Berens, P.H.; Wilson, K.R. , A Computer Simulation Method for the Calculation of Equilibrium Constants for the Formation of Physical Clusters of Molecules: Application to Small Water Clusters. . *Journal of Chemical Physics* **1982,** *76*, 637-649.

260. Hockney, R. W.; Goel, S. P.; Eastwood, J. W., Quiet high-resolution computer models of a plasma. *Journal of Computational Physics* **1974,** *14* (2), 148-158.

261. Loncharich, R. J.; Brooks, B. R.; Pastor, R. W., Langevin dynamics of peptides: the frictional dependence of isomerization rates of N-acetylalanyl-N'-methylamide. *Biopolymers* **1992,** *32* (5), 523-35.

262. Parrinello, M.; Rahman, A., Polymorphic transitions in single crystals: A new molecular dynamics method. *Journal of Applied Physics* **1981,** *52* (12), 7182-7190.

263. Hess, B.; Bekker, H.; Berendsen, H. J. C.; Fraaije, J. G. E. M., LINCS: A linear constraint solver for molecular simulations. *Journal of Computational Chemistry* **1997,** *18* (12), 1463-1472.

264. Deelder, W.; Christakoudi, S.; Phelan, J.; Benavente, E. D.; Campino, S.; McNerney, R.; Palla, L.; Clark, T. G., Machine Learning Predicts Accurately Mycobacterium tuberculosis Drug Resistance From Whole Genome Sequencing Data. *Front Genet* **2019,** *10*, 922.

265. Kouchaki, S.; Yang, Y.; Walker, T. M.; Sarah Walker, A.; Wilson, D. J.; Peto, T. E. A.; Crook, D. W.; Consortium, C.; Clifton, D. A., Application of machine learning techniques to tuberculosis drug resistance analysis. *Bioinformatics* **2018,** *35* (13), 2276-2282.

266. Chen, M. L.; Doddi, A.; Royer, J.; Freschi, L.; Schito, M.; Ezewudo, M.; Kohane, I. S.; Beam, A.; Farhat, M., Beyond multidrug resistance: Leveraging rare variants with machine and statistical learning models in Mycobacterium tuberculosis resistance prediction. *EBioMedicine* **2019,** *43*, 356-369.

267. Merker, M.; Kohl, T. A.; Barilar, I.; Andres, S.; Fowler, P. W.; Chryssanthou, E.; Angeby, K.; Jureen, P.; Moradigaravand, D.; Parkhill, J.; Peacock, S. J.; Schon, T.; Maurer, F. P.; Walker, T.; Koser, C.; Niemann, S., Phylogenetically informative mutations in genes implicated in antibiotic resistance in Mycobacterium tuberculosis complex. *Genome Med* **2020,** *12* (1), 27.

268. Köser, C. U.; Bryant, J. M.; Parkhill, J.; Peacock, S. J., Consequences of whiB7 (Rv3197A) mutations in Beijing genotype isolates of the Mycobacterium tuberculosis complex. *Antimicrob Agents Chemother* **2013,** *57* (7), 3461.

269. Walker, T. M.; Merker, M.; Knoblauch, A. M.; Helbling, P.; Schoch, O. D.; van der Werf, M. J.; Kranzer, K.; Fiebig, L.; Kroger, S.; Haas, W.; Hoffmann, H.; Indra, A.; Egli, A.; Cirillo, D. M.; Robert, J.; Rogers, T. R.; Groenheit, R.; Mengshoel, A. T.; Mathys, V.; Haanpera, M.; Soolingen, D. V.; Niemann, S.; Bottger, E. C.; Keller, P. M.; The CRyPTIC Consortium, A cluster of multidrug-resistant Mycobacterium tuberculosis among patients arriving in Europe from the Horn of Africa: a molecular epidemiological study. *Lancet Infect Dis* **2018,** *18* (4), 431-440.

270. Salaam-Dreyer, Z.; Streicher, E. M.; Sirgel, F. A.; Menardo, F.; Borrell, S.; Reinhard, M.; Doetsch, A.; Cudahy, P. G. T.; Mohr-Holland, E.; Daniels, J.; Dippenaar, A.; Nicol, M. P.; Gagneux, S.; Warren, R. M.; Cox, H., Rifampicin-Monoresistant Tuberculosis Is Not the Same as Multidrug-Resistant Tuberculosis: a Descriptive Study from Khayelitsha, South Africa. *Antimicrobial Agents and Chemotherapy* **2021,** *65* (11), e00364-21.

271. Modlin, S. J.; Marbach, T.; Werngren, J.; Mansjö, M.; Hoffner, S. E.; Valafar, F., Atypical Genetic Basis of Pyrazinamide Resistance in Monoresistant Mycobacterium tuberculosis. *Antimicrob Agents Chemother* **2021,** *65* (6).

272. The CRyPTIC Consortium, Epidemiological cutoff values for a 96-well broth microdilution plate for high-throughput research antibiotic susceptibility testing of *M. tuberculosis*. *European Respiratory Journal* **2022**, 2200239.

273. Hunt, M.; Letcher, B.; Malone, K. M.; Nguyen, G.; Hall, M. B.; Colquhoun, R. M.; Lima, L.; Schatz, M. C.; Ramakrishnan, S.; consortium, C.; Iqbal, Z., Minos: variant adjudication and joint genotyping of cohorts of bacterial genomes. *bioRxiv* **2021**, 2021.09.15.460475.

274. Hunt, M.; Bradley, P.; Lapierre, S. G.; Heys, S.; Thomsit, M.; Hall, M. B.; Malone, K. M.; Wintringer, P.; Walker, T. M.; Cirillo, D. M.; Comas, I.; Farhat, M. R.; Fowler, P.; Gardy, J.; Ismail, N.; Kohl, T. A.; Mathys, V.; Merker, M.; Niemann, S.; Omar, S. V.; Sintchenko, V.; Smith, G.; van Soolingen, D.; Supply, P.; Tahseen, S.; Wilcox, M.; Arandjelovic, I.; Peto, T. E. A.; Crook, D. W.; Iqbal, Z., Antibiotic resistance prediction for Mycobacterium tuberculosis from genome sequence data with Mykrobe. *Wellcome Open Res* **2019,** *4*, 191.

275. Rancoita, P. M. V.; Cugnata, F.; Gibertoni Cruz, A. L.; Borroni, E.; Hoosdally, S. J.; Walker, T. M.; Grazian, C.; Davies, T. J.; Peto, T. E. A.; Crook, D. W.; Fowler, P. W.; Cirillo, D. M., Validating a 14-Drug Microtiter Plate Containing Bedaquiline and Delamanid for Large-Scale Research Susceptibility Testing of Mycobacterium tuberculosis. *Antimicrob Agents Chemother* **2018,** *62* (9).

276. Fowler, P. W.; Gibertoni Cruz, A. L.; Hoosdally, S. J.; Jarrett, L.; Borroni, E.; Chiacchiaretta, M.; Rathod, P.; Lehmann, S.; Molodtsov, N.; Walker, T. M.; Robinson, E.; Hoffmann, H.; Peto, T. E. A.; Cirillo, D. M.; Smith, G. E.; Crook, D. W., Automated detection of bacterial growth on 96-well plates for high-throughput drug susceptibility testing of Mycobacterium tuberculosis. *Microbiology* **2018**.

277. Fowler, P. W.; Wright, C.; Spiers, H.; Zhu, T.; Baeten, E. M.; Hoosdally, S. W.; Cruz, A. L. G.; Roohi, A.; Kouchaki, S.; Walker, T. M.; Peto, T. E.; Miller, G.; Lintott, C.; Clifton, D.; Crook, D. W.; Walker, A. S.; Community, T. Z. V.; Consortium, T. C., BashTheBug: a crowd of volunteers reproducibly and accurately measure the minimum inhibitory concentrations of 13 antitubercular drugs from photographs of 96-well broth microdilution plates. *bioRxiv* **2021**, 2021.07.20.453060.

278. Testing, E. C. f. A. S., MIC distributions and epidemiological cut-off value (ECOFF) setting. In *EUCAST SOP* 2017; Vol. 10.0, pp 1-17.

279. Newcombe, R. G., Improved confidence intervals for the difference between binomial proportions based on paired data. *Stat Med* **1998,** *17* (22), 2635-50.

280. Grange, J. M., Mycobacterium bovis infection in human beings. *Tuberculosis (Edinb)* **2001,** *81* (1-2), 71-7.

281. Falzon, D.; Schunemann, H. J.; Harausz, E.; Gonzalez-Angulo, L.; Lienhardt, C.; Jaramillo, E.; Weyer, K., World Health Organization treatment guidelines for drug-resistant tuberculosis, 2016 update. *Eur Respir J* **2017,** *49* (3).

282. Ghimire, S.; Karki, S.; Maharjan, B.; Kosterink, J. G. W.; Touw, D. J.; van der Werf, T. S.; Shrestha, B.; Alffenaar, J. W., Treatment outcomes of patients with MDR-TB in Nepal on a current programmatic standardised regimen: retrospective single-centre study. *BMJ Open Respir Res* **2020,** *7* (1).

283. Vilchèze, C.; Wang, F.; Arai, M.; Hazbón, M. H.; Colangeli, R.; Kremer, L.; Weisbrod, T. R.; Alland, D.; Sacchettini, J. C.; Jacobs, W. R., Jr., Transfer of a point mutation in Mycobacterium tuberculosis inhA resolves the target of isoniazid. *Nat Med* **2006,** *12* (9), 1027-9.

284. Morlock, G. P.; Metchock, B.; Sikes, D.; Crawford, J. T.; Cooksey, R. C., ethA, inhA, and katG loci of ethionamide-resistant clinical Mycobacterium tuberculosis isolates. *Antimicrob Agents Chemother* **2003,** *47* (12), 3799-805.

285. Mvelase, N. R.; Balakrishna, Y.; Lutchminarain, K.; Mlisana, K., Evolving rifampicin and isoniazid mono-resistance in a high multidrug-resistant and extensively drug-resistant tuberculosis region: a retrospective data analysis. *BMJ Open* **2019,** *9* (11), e031663.

286. Heysell, S. K.; Ahmed, S.; Rahman, M. T.; Akhanda, M. W.; Gleason, A. T.; Ebers, A.; Houpt, E. R.; Banu, S., Hearing loss with kanamycin treatment for multidrug-resistant tuberculosis in Bangladesh. *European Respiratory Journal* **2018,** *51* (3), 1701778.

287. Nimmo, C.; Millard, J.; van Dorp, L.; Brien, K.; Moodley, S.; Wolf, A.; Grant, A. D.; Padayatchi, N.; Pym, A. S.; Balloux, F.; O'Donnell, M., Population-level emergence of bedaquiline and clofazimine resistance-associated variants among patients with drug-resistant tuberculosis in southern Africa: a phenotypic and phylogenetic analysis. *Lancet Microbe* **2020,** *1* (4), e165-e174.

288. Huitric, E.; Verhasselt, P.; Koul, A.; Andries, K.; Hoffner, S.; Andersson, D. I., Rates and mechanisms of resistance development in Mycobacterium tuberculosis to a novel diarylquinoline ATP synthase inhibitor. *Antimicrob Agents Chemother* **2010,** *54* (3), 1022-8.

289. Nasiri, M. J.; Zamani, S.; Pormohammad, A.; Feizabadi, M. M.; Aslani, H. R.; Amin, M.; Halabian, R.; Imani Fooladi, A. A., The reliability of rifampicin resistance as a proxy for multidrug-resistant tuberculosis: a systematic review of studies from Iran. *Eur J Clin Microbiol Infect Dis* **2018,** *37* (1), 9-14.

290. Manson, A. L.; Cohen, K. A.; Abeel, T.; Desjardins, C. A.; Armstrong, D. T.; Barry, C. E., 3rd; Brand, J.; Consortium, T. B. G. G.; Chapman, S. B.; Cho, S. N.; Gabrielian, A.; Gomez, J.; Jodals, A. M.; Joloba, M.; Jureen, P.; Lee, J. S.; Malinga, L.; Maiga, M.; Nordenberg, D.; Noroc, E.; Romancenco, E.; Salazar, A.; Ssengooba, W.; Velayati, A. A.; Winglee, K.; Zalutskaya, A.; Via, L. E.; Cassell, G. H.; Dorman, S. E.; Ellner, J.; Farnia, P.; Galagan, J. E.; Rosenthal, A.; Crudu, V.; Homorodean, D.; Hsueh, P. R.; Narayanan, S.; Pym, A. S.; Skrahina, A.; Swaminathan, S.; Van der Walt, M.; Alland, D.; Bishai, W. R.; Cohen, T.; Hoffner, S.; Birren, B. W.; Earl, A. M., Genomic analysis of globally diverse Mycobacterium tuberculosis strains provides insights into the emergence and spread of multidrug resistance. *Nat Genet* **2017,** *49* (3), 395-402.

291. Dean, A. S.; Zignol, M.; Cabibbe, A. M.; Falzon, D.; Glaziou, P.; Cirillo, D. M.; Köser, C. U.; Gonzalez-Angulo, L. Y.; Tosas-Auguet, O.; Ismail, N.; Tahseen, S.; Ama, M. C. G.; Skrahina, A.; Alikhanova, N.; Kamal, S. M. M.; Floyd, K., Prevalence and genetic profiles of isoniazid resistance in tuberculosis patients: A multicountry analysis of cross-sectional data. *PLOS Medicine* **2020,** *17* (1), e1003008.

292. Nguyen, L., Antibiotic resistance mechanisms in *M. tuberculosis*: an update. *Arch Toxicol* **2016,** *90* (7), 1585-604.

293. Autoimmun Diagnostika., PCR DETECTION OF MYCOBACTERIUM TUBERCULOSIS COMPLEX AND ITS RESISTANCES AGAINST FLUOROQUINOLONES AND ETHAMBUTOL. https://www.aid-diagnostika.com/en/kits/molecular-biologic-assay/infectious-diseases/antibiotic-resistances/tb-fluoroquinolone-ethambutol (accessed 12 May 2022).

294. Pérez-García, F.; Ruiz-Serrano, M. J.; López Roa, P.; Acosta, F.; Pérez-Lago, L.; García-De-Viedma, D.; Bouza, E., Diagnostic performance of Anyplex II MTB/MDR/XDR for detection of resistance to first and second line drugs in Mycobacterium tuberculosis. *Journal of Microbiological Methods* **2017,** *139*, 74-78.

295. Mitarai, S.; Kato, S.; Ogata, H.; Aono, A.; Chikamatsu, K.; Mizuno, K.; Toyota, E.; Sejimo, A.; Suzuki, K.; Yoshida, S.; Saito, T.; Moriya, A.; Fujita, A.; Sato, S.; Matsumoto, T.; Ano, H.; Suetake, T.; Kondo, Y.; Kirikae, T.; Mori, T., Comprehensive multicenter evaluation of a new line probe assay kit for identification of Mycobacterium species and detection of drug-resistant Mycobacterium tuberculosis. *J Clin Microbiol* **2012,** *50* (3), 884-90.

296. Theron, G.; Peter, J.; Richardson, M.; Warren, R.; Dheda, K.; Steingart, K. R., GenoType((R)) MTBDRsl assay for resistance to second-line anti-tuberculosis drugs. *Cochrane Database Syst Rev* **2016,** *9*, CD010705.

297. Machado, D.; Couto, I.; Viveiros, M., Advances in the molecular diagnosis of tuberculosis: From probes to genomes. *Infect Genet Evol* **2019,** *72*, 93-112.

298. Initiative, G. L. Line probe assays for drug-resistant tuberculosis detection Interpretation and reporting guide for laboratory staff and clinicians. https://stoptb.org/wg/gli/assets/documents/LPA_test_web_ready.pdf (accessed 14/01/2022).

299. Zeesan MeltPro® MTB/FQ Test Kit Product Details. http://www.zeesandx.com/products/meltpro-mtbfq-test-kit.html (accessed 12 May 2022).

300. Lee, Y. S.; Kang, M. R.; Jung, H.; Choi, S. B.; Jo, K.-W.; Shim, T. S., Performance of REBA MTB-XDR to detect extensively drug-resistant tuberculosis in an intermediate-burden country. *Journal of Infection and Chemotherapy* **2015,** *21* (5), 346-351.

301. Chakravorty, S.; Roh, S. S.; Glass, J.; Smith, L. E.; Simmons, A. M.; Lund, K.; Lokhov, S.; Liu, X.; Xu, P.; Zhang, G.; Via, L. E.; Shen, Q.; Ruan, X.; Yuan, X.; Zhu, H. Z.; Viazovkina, E.; Shenai, S.; Rowneki, M.; Lee, J. S.; Barry, C. E., 3rd; Gao, Q.; Persing, D.; Kwiatkawoski, R.; Jones, M.; Gall, A.; Alland, D., Detection of Isoniazid-, Fluoroquinolone-, Amikacin-, and Kanamycin-Resistant Tuberculosis in an Automated, Multiplexed 10-Color Assay Suitable for Point-of-Care Use. *J Clin Microbiol* **2017,** *55* (1), 183-198.

302. Cepheid Xpert® MTB/XDR Package Insert. https://www.cepheid.com/Package%20Insert%20Files/Xpert%20MTB-XDR%20ENGLISH%20Package%20Insert%20302-3514%20Rev%20C.pdf (accessed 14/01/22).

303. Pantel, A.; Petrella, S.; Veziris, N.; Matrat, S.; Bouige, A.; Ferrand, H.; Sougakoff, W.; Mayer, C.; Aubry, A., Description of compensatory gyrA mutations restoring fluoroquinolone susceptibility in Mycobacterium tuberculosis. *Journal of Antimicrobial Chemotherapy* **2016,** *71* (9), 2428-2431.

304. Castro, R. A. D.; Ross, A.; Kamwela, L.; Reinhard, M.; Loiseau, C.; Feldmann, J.; Borrell, S.; Trauner, A.; Gagneux, S., The Genetic Background Modulates the Evolution of Fluoroquinolone-Resistance in Mycobacterium tuberculosis. *Mol Biol Evol* **2020,** *37* (1), 195-207.

305. Rufai, S. B.; Singh, J.; Kumar, P.; Mathur, P.; Singh, S., Association of gyrA and rrs gene mutations detected by MTBDRsl V1 on Mycobacterium tuberculosis strains of diverse genetic background from India. *Scientific Reports* **2018,** *8* (1), 9295.

306. Li, Q.; Gao, H.; Zhang, Z.; Tian, Y.; Liu, T.; Wang, Y.; Lu, J.; Liu, Y.; Dai, E., Mutation and Transmission Profiles of Second-Line Drug Resistance in Clinical Isolates of Drug-Resistant Mycobacterium tuberculosis From Hebei Province, China. *Frontiers in Microbiology* **2019,** *10*.

307. Chen, L.; Li, H.; Chen, T.; Yu, L.; Guo, H.; Chen, Y.; Chen, M.; Li, Z.; Wu, Z.; Wang, X.; Zhao, J.; Yan, H.; Wang, X.; Zhou, L.; Zhou, J., Genome-wide DNA methylation and transcriptome changes in Mycobacterium tuberculosis with rifampicin and isoniazid resistance. *Int J Clin Exp Pathol* **2018,** *11* (6), 3036-3045.

308. Chu, H.; Hu, Y.; Zhang, B.; Sun, Z.; Zhu, B., DNA Methyltransferase HsdM Induce Drug Resistance on Mycobacterium tuberculosis via Multiple Effects. *Antibiotics (Basel)* **2021,** *10* (12).

309. Nimmo, C.; Brien, K.; Millard, J.; Grant, A. D.; Padayatchi, N.; Pym, A. S.; O'Donnell, M.; Goldstein, R.; Breuer, J.; Balloux, F., Dynamics of within-host Mycobacterium tuberculosis diversity and heteroresistance during treatment. *EBioMedicine* **2020,** *55*, 102747.

310. Said Mohammed, K.; Kibinge, N.; Prins, P.; Agoti, C. N.; Cotten, M.; Nokes, D. J.; Brand, S.; Githinji, G., Evaluating the performance of tools used to call minority variants from whole genome short-read data. *Wellcome Open Res* **2018,** *3*, 21.

311. Bradley, P.; Gordon, N. C.; Walker, T. M.; Dunn, L.; Heys, S.; Huang, B.; Earle, S.; Pankhurst, L. J.; Anson, L.; de Cesare, M.; Piazza, P.; Votintseva, A. A.; Golubchik, T.; Wilson, D. J.; Wyllie, D. H.; Diel, R.; Niemann, S.; Feuerriegel, S.; Kohl, T. A.; Ismail, N.; Omar, S. V.; Smith, E. G.; Buck, D.; McVean, G.; Walker, A. S.; Peto, T. E.; Crook, D. W.; Iqbal, Z., Rapid antibiotic-resistance predictions from genome sequence data for Staphylococcus aureus and Mycobacterium tuberculosis. *Nat Commun* **2015,** *6*, 10063.

312. Rigouts, L.; Miotto, P.; Schats, M.; Lempens, P.; Cabibbe, A. M.; Galbiati, S.; Lampasona, V.; de Rijk, P.; Cirillo, D. M.; de Jong, B. C., Fluoroquinolone heteroresistance in Mycobacterium tuberculosis: detection by genotypic and phenotypic assays in experimentally mixed populations. *Sci Rep* **2019,** *9* (1), 11760.

313. Georghiou, S. B.; Penn-Nicholson, A.; de Vos, M.; Macé, A.; Syrmis, M. W.; Jacob, K.; Mape, A.; Parmar, H.; Cao, Y.; Coulter, C.; Ruhwald, M.; Pandey, S. K.; Schumacher, S. G.; Denkinger, C. M., Analytical performance of the Xpert MTB/XDR® assay for tuberculosis and expanded resistance detection. *Diagnostic Microbiology and Infectious Disease* **2021,** *101* (1), 115397.

314. Seabold, S.; Perktold, J. In *Statsmodels: Econometric and Statistical Modeling with Python*, 2010.

315. Zhang, X.; Zhao, B.; Liu, L.; Zhu, Y.; Zhao, Y.; Jin, Q., Subpopulation analysis of heteroresistance to fluoroquinolone in Mycobacterium tuberculosis isolates from Beijing, China. *J Clin Microbiol* **2012,** *50* (4), 1471-4.

316. Avalos, E.; Catanzaro, D.; Catanzaro, A.; Ganiats, T.; Brodine, S.; Alcaraz, J.; Rodwell, T., Frequency and geographic distribution of gyrA and gyrB mutations associated with fluoroquinolone resistance in clinical Mycobacterium tuberculosis isolates: a systematic review. *PloS one* **2015,** *10* (3), e0120470.

317. Luo, T.; Yuan, J.; Peng, X.; Yang, G.; Mi, Y.; Sun, C.; Wang, C.; Zhang, C.; Bao, L., Double mutation in DNA gyrase confers moxifloxacin resistance and decreased fitness of Mycobacterium smegmatis. *J Antimicrob Chemother* **2017,** *72* (7), 1893-1900.

318. Li, J.; Gao, X.; Luo, T.; Wu, J.; Sun, G.; Liu, Q.; Jiang, Y.; Zhang, Y.; Mei, J.; Gao, Q., Association of gyrA/B mutations and resistance levels to fluoroquinolones in clinical isolates of Mycobacterium tuberculosis. *Emerg Microbes Infect* **2014,** *3* (3), e19.

319. Van Deun, A.; Chiang, C.-Y., Shortened multidrug-resistant tuberculosis regimens overcome low-level fluoroquinolone resistance. *European Respiratory Journal* **2017,** *49* (6), 1700223.

320. World Health Organization (WHO), Target product profile for next-generation tuberculosis drug-susceptibility testing at peripheral centres. **2021**.

321. Zhang, D.; Gomez, J. E.; Chien, J. Y.; Haseley, N.; Desjardins, C. A.; Earl, A. M.; Hsueh, P. R.; Hung, D. T., Genomic Analysis of the Evolution of Fluoroquinolone Resistance in Mycobacterium tuberculosis Prior to Tuberculosis Diagnosis. *Antimicrob Agents Chemother* **2016,** *60* (11), 6600-6608.

322. Borrell, S.; Teo, Y.; Giardina, F.; Streicher, E. M.; Klopper, M.; Feldmann, J.; Müller, B.; Victor, T. C.; Gagneux, S., Epistasis between antibiotic resistance mutations drives the evolution of extensively drug-resistant tuberculosis. *Evol Med Public Health* **2013,** *2013* (1), 65-74.

323. Gupta, A. K.; Katoch, V. M.; Chauhan, D. S.; Sharma, R.; Singh, M.; Venkatesan, K.; Sharma, V. D., Microarray analysis of efflux pump genes in multidrug-resistant Mycobacterium tuberculosis during stress induced by common anti-tuberculous drugs. *Microbial drug resistance* **2010,** *16* (1), 21-8.

324. Pi, R.; Liu, Q.; Takiff, H. E.; Gao, Q., Fitness Cost and Compensatory Evolution in Levofloxacin-Resistant Mycobacterium aurum. *Antimicrob Agents Chemother* **2020,** *64* (8).

325. Goossens, S. N.; Heupink, T. H.; De Vos, E.; Dippenaar, A.; De Vos, M.; Warren, R.; Van Rie, A., Detection of minor variants in Mycobacterium tuberculosis whole genome sequencing data. *Brief Bioinform* **2022,** *23* (1).

326. Kavvas, E. S.; Catoiu, E.; Mih, N.; Yurkovich, J. T.; Seif, Y.; Dillon, N.; Heckmann, D.; Anand, A.; Yang, L.; Nizet, V.; Monk, J. M.; Palsson, B. O., Machine learning and structural

analysis of Mycobacterium tuberculosis pan-genome identifies genetic signatures of antibiotic resistance. *Nature communications* **2018,** *9* (1), 4306.

327. Pires, D. E. V.; Blundell, T. L.; Ascher, D. B., mCSM-lig: quantifying the effects of mutations on protein-small molecule affinity in genetic disease and emergence of drug resistance. *Scientific Reports* **2016,** *6* (1), 29575.

328. Chowdhury, A. S.; Khaledian, E.; Broschat, S. L., Capreomycin resistance prediction in two species of Mycobacterium using a stacked ensemble method. *Journal of Applied Microbiology* **2019,** *127* (6), 1656-1664.

329. Carter, J. J.; Walker, T. M.; Walker, A. S.; Whitfield, M. G.; Morlock, G. P.; Peto, T. E.; Posey, J. E.; Crook, D. W.; Fowler, P. W., Prediction of pyrazinamide resistance in Mycobacterium tuberculosis using structure-based machine learning approaches. *bioRxiv* **2019**, 518142.

330. Jamal, S.; Khubaib, M.; Gangwar, R.; Grover, S.; Grover, A.; Hasnain, S. E., Artificial Intelligence and Machine learning based prediction of resistant and susceptible mutations in Mycobacterium tuberculosis. *Scientific Reports* **2020,** *10* (1), 5487.

331. World Health Organization (WHO). Updated interim critical concentrations for first-line and second-line DST 2012, p. 1. http://www.stoptb.org/wg/gli/assets/documents/Updated%20critical%20concentration%20table_1st%20and%202nd%20line%20drugs.pdf (accessed 11/30/2016).

332. Hecht, M.; Bromberg, Y.; Rost, B., Better prediction of functional effects for sequence variants. *BMC Genomics* **2015,** *16 Suppl 8* (Suppl 8), S1.

333. Touw, W. G.; Baakman, C.; Black, J.; te Beek, T. A.; Krieger, E.; Joosten, R. P.; Vriend, G., A series of PDB-related databanks for everyday needs. *Nucleic acids research* **2015,** *43* (Database issue), D364-8.

334. Salentin, S.; Schreiber, S.; Haupt, V. J.; Adasme, M. F.; Schroeder, M., PLIP: fully automated protein-ligand interaction profiler. *Nucleic acids research* **2015,** *43* (W1), W443-7.

335. Tan, K. P.; Varadarajan, R.; Madhusudhan, M. S., DEPTH: a web server to compute depth and predict small-molecule binding cavities in proteins. *Nucleic acids research* **2011,** *39* (Web Server issue), W242-8.

336. Pires, D. E.; Ascher, D. B.; Blundell, T. L., mCSM: predicting the effects of mutations in proteins using graph-based signatures. *Bioinformatics* **2014,** *30* (3), 335-42.

337. Michaud-Agrawal, N.; Denning, E. J.; Woolf, T. B.; Beckstein, O., MDAnalysis: a toolkit for the analysis of molecular dynamics simulations. *J Comput Chem* **2011,** *32* (10), 2319-27.

338. Xu, P.; Li, X.; Zhao, M.; Gui, X.; DeRiemer, K.; Gagneux, S.; Mei, J.; Gao, Q., Prevalence of fluoroquinolone resistance among tuberculosis patients in Shanghai, China. *Antimicrobial agents and chemotherapy* **2009,** *53* (7), 3170-3172.

339. Pedregosa, F.; Varoquaux, G.; Gramfort, A.; Michel, V.; Thirion, B.; Grisel, O.; Blondel, M.; Prettenhofer, P.; Weiss, R.; Dubourg, V.; Vanderplas, J.; Passos, A.; Cournapeau, D.; Brucher, M.; Perrot, M.; Duchesnay, É., Scikit-learn: Machine Learning in Python. *J. Mach. Learn. Res.* **2011,** *12* (null), 2825–2830.

340. Antoniuk, K.; Franc, V.; Hlaváč, V. In *MORD: Multi-class Classifier for Ordinal Regression*, Berlin, Heidelberg, Springer Berlin Heidelberg: Berlin, Heidelberg, 2013; pp 96-111.

341. Germe, T.; Vörös, J.; Jeannot, F.; Taillier, T.; Stavenger, R. A.; Bacqué, E.; Maxwell, A.; Bax, B. D., A new class of antibacterials, the imidazopyrazinones, reveal structural transitions involved in DNA gyrase poisoning and mechanisms of resistance. *Nucleic acids research* **2018,** *46* (8), 4114-4128.

342. Robinson, A.; Causer, R. J.; Dixon, N. E., Architecture and conservation of the bacterial DNA replication machinery, an underexploited drug target. *Curr Drug Targets* **2012,** *13* (3), 352-72.

343. Standardization, I. O. f., Clinical laboratory testing and in vitro diagnostic test systems — Susceptibility testing of infectious agents and evaluation of performance of antimicrobial susceptibility test devices — Part 2: Evaluation of performance of antimicrobial susceptibility test devices. 2007.

344. Johnsen, C. H.; Clausen, P. T. L. C.; Aarestrup, F. M.; Lund, O., Improved Resistance Prediction in Mycobacterium tuberculosis by Better Handling of Insertions and Deletions, Premature Stop Codons, and Filtering of Non-informative Sites. *Frontiers in Microbiology* **2019,** *10*.

345. Ren, Y.; Chakraborty, T.; Doijad, S.; Falgenhauer, L.; Falgenhauer, J.; Goesmann, A.; Hauschild, A.-C.; Schwengers, O.; Heider, D., Prediction of antimicrobial resistance based on whole-genome sequencing and machine learning. *Bioinformatics* **2021,** *38* (2), 325-334.

346. Lau, R. W. T.; Ho, P.-L.; Kao, R. Y. T.; Yew, W.-W.; Lau, T. C. K.; Cheng, V. C. C.; Yuen, K.-Y.; Tsui, S. K. W.; Chen, X.; Yam, W.-C., Molecular Characterization of Fluoroquinolone Resistance in Mycobacterium tuberculosis: Functional Analysis of gyrA Mutation at Position 74. *Antimicrobial Agents and Chemotherapy* **2011,** *55* (2), 608-614.

347. Nosova, E. Y.; Bukatina, A. A.; Isaeva, Y. D.; Makarova, M. V.; Galkina, K. Y.; Moroz, A. M., Analysis of mutations in the gyrA and gyrB genes and their association with the resistance of Mycobacterium tuberculosis to levofloxacin, moxifloxacin and gatifloxacin. *Journal of medical microbiology* **2013,** *62* (1), 108-113.

348. Yoshida, M.; Nakata, N.; Miyamoto, Y.; Fukano, H.; Ato, M.; Hoshino, Y., A rapid and non-pathogenic assay for association of Mycobacterium tuberculosis gyrBA mutations and fluoroquinolone resistance using recombinant Mycobacterium smegmatis. *FEMS Microbiology Letters* **2018,** *365* (23).

349. Wang, L.; Wu, Y.; Deng, Y.; Kim, B.; Pierce, L.; Krilov, G.; Lupyan, D.; Robinson, S.; Dahlgren, M. K.; Greenwood, J.; Romero, D. L.; Masse, C.; Knight, J. L.; Steinbrecher, T.; Beuming, T.; Damm, W.; Harder, E.; Sherman, W.; Brewer, M.; Wester, R.; Murcko, M.; Frye, L.; Farid, R.; Lin, T.; Mobley, D. L.; Jorgensen, W. L.; Berne, B. J.; Friesner, R. A.; Abel,

R., Accurate and Reliable Prediction of Relative Ligand Binding Potency in Prospective Drug Discovery by Way of a Modern Free-Energy Calculation Protocol and Force Field. *Journal of the American Chemical Society* **2015,** *137* (7), 2695-2703.

350. Gapsys, V.; Pérez-Benito, L.; Aldeghi, M.; Seeliger, D.; van Vlijmen, H.; Tresadern, G.; de Groot, B. L., Large scale relative protein ligand binding affinities using non-equilibrium alchemy. *Chemical Science* **2020,** *11* (4), 1140-1152.

351. Steinbrecher, T. B.; Dahlgren, M.; Cappel, D.; Lin, T.; Wang, L.; Krilov, G.; Abel, R.; Friesner, R.; Sherman, W., Accurate Binding Free Energy Predictions in Fragment Optimization. *Journal of Chemical Information and Modeling* **2015,** *55* (11), 2411-2420.

352. Ponder, J. W.; Case, D. A., Force fields for protein simulations. *Adv Protein Chem* **2003,** *66*, 27-85.

353. Gapsys, V.; Michielssens, S.; Seeliger, D.; de Groot, B. L., pmx: Automated protein structure and topology generation for alchemical perturbations. *J Comput Chem* **2015,** *36* (5), 348-54.

354. Fowler, P. W.; Cole, K.; Gordon, N. C.; Kearns, A. M.; Llewelyn, M. J.; Peto, T. E. A.; Crook, D. W.; Walker, A. S., Robust Prediction of Resistance to Trimethoprim in Staphylococcus aureus. *Cell Chem Biol* **2018,** *25* (3), 339-349 e4.

355. Hauser, K.; Negron, C.; Albanese, S. K.; Ray, S.; Steinbrecher, T.; Abel, R.; Chodera, J. D.; Wang, L., Predicting resistance of clinical Abl mutations to targeted kinase inhibitors using alchemical free-energy calculations. *Commun Biol* **2018,** *1*, 70.

356. Aldeghi, M.; Gapsys, V.; de Groot, B. L., Predicting Kinase Inhibitor Resistance: Physics-Based and Data-Driven Approaches. *ACS Central Science* **2019,** *5* (8), 1468-1474.

357. Bhati, A. P.; Wan, S.; Coveney, P. V., Ensemble-Based Replica Exchange Alchemical Free Energy Methods: The Effect of Protein Mutations on Inhibitor Binding. *Journal of Chemical Theory and Computation* **2019,** *15* (2), 1265-1277.

358. Eastman, P.; Swails, J.; Chodera, J. D.; McGibbon, R. T.; Zhao, Y.; Beauchamp, K. A.; Wang, L. P.; Simmonett, A. C.; Harrigan, M. P.; Stern, C. D.; Wiewiora, R. P.; Brooks, B. R.; Pande, V. S., OpenMM 7: Rapid development of high performance algorithms for molecular dynamics. *PLoS Comput Biol* **2017,** *13* (7), e1005659.

359. Kutzner, C.; Páll, S.; Fechner, M.; Esztermann, A.; de Groot, B. L.; Grubmüller, H., More bang for your buck: Improved use of GPU nodes for GROMACS 2018. *Journal of Computational Chemistry* **2019,** *40* (27), 2418-2431.

360. Cournia, Z.; Allen, B.; Sherman, W., Relative Binding Free Energy Calculations in Drug Discovery: Recent Advances and Practical Considerations. *Journal of Chemical Information and Modeling* **2017,** *57* (12), 2911-2937.

361. Fowler, P. W., How quickly can we predict trimethoprim resistance using alchemical free energy methods? *Interface Focus* **2020,** *10* (6).

362. The CRyPTIC Consortium; Carter, J. J., Quantitative measurement of antibiotic resistance in Mycobacterium tuberculosis reveals genetic determinants of resistance and susceptibility in a target gene approach. *bioRxiv* **2021**.

363. Wohlkonig, A.; Chan, P. F.; Fosberry, A. P.; Homes, P.; Huang, J.; Kranz, M.; Leydon, V. R.; Miles, T. J.; Pearson, N. D.; Perera, R. L.; Shillings, A. J.; Gwynn, M. N.; Bax, B. D., Structural basis of quinolone inhibition of type IIA topoisomerases and target-mediated resistance. *Nat Struct Mol Biol* **2010,** *17* (9), 1152-3.

364. Fiser, A.; Sali, A., ModLoop: automated modeling of loops in protein structures. *Bioinformatics* **2003,** *19* (18), 2500-1.

365. Bas, D. C.; Rogers, D. M.; Jensen, J. H., Very fast prediction and rationalization of pKa values for protein-ligand complexes. *Proteins* **2008,** *73* (3), 765-83.

366. Sousa da Silva, A. W.; Vranken, W. F., ACPYPE - AnteChamber PYthon Parser interfacE. *BMC Research Notes* **2012,** *5* (1), 367.

367. Jefferys, E.; Sands, Z. A.; Shi, J.; Sansom, M. S.; Fowler, P. W., Alchembed: A Computational Method for Incorporating Multiple Proteins into Complex Lipid Geometries. *J Chem Theory Comput* **2015,** *11* (6), 2743-2754.

368. Pham, T. T.; Shirts, M. R., Identifying low variance pathways for free energy calculations of molecular transformations in solution phase. *J Chem Phys* **2011,** *135* (3), 034114.

369. Galindo-Murillo, R.; Roe, D. R.; Cheatham, T. E., 3rd, Convergence and reproducibility in molecular dynamics simulations of the DNA duplex d(GCACGAACGAACGAACGC). *Biochim Biophys Acta* **2015,** *1850* (5), 1041-1058.

370. Lundborg, M.; Lindahl, E., Automatic GROMACS topology generation and comparisons of force fields for solvation free energy calculations. *J Phys Chem B* **2015,** *119* (3), 810-23.

371. Rocklin, G. J.; Mobley, D. L.; Dill, K. A.; Hünenberger, P. H., Calculating the binding free energies of charged species based on explicit-solvent simulations employing lattice-sum methods: an accurate correction scheme for electrostatic finite-size effects. *J Chem Phys* **2013,** *139* (18), 184103.

372. Öhlknecht, C.; Lier, B.; Petrov, D.; Fuchs, J.; Oostenbrink, C., Correcting electrostatic artifacts due to net-charge changes in the calculation of ligand binding free energies. *Journal of Computational Chemistry* **2020,** *41* (10), 986-999.

373. Lin, Y.-L.; Aleksandrov, A.; Simonson, T.; Roux, B., An Overview of Electrostatic Free Energy Computations for Solutions and Proteins. *Journal of Chemical Theory and Computation* **2014,** *10* (7), 2690-2709.

374. Richter, S. N.; Giaretta, G.; Comuzzi, V.; Leo, E.; Mitchenall, L. A.; Fisher, L. M.; Maxwell, A.; Palumbo, M., Hot-spot consensus of fluoroquinolone-mediated DNA cleavage by Gram-negative and Gram-positive type II DNA topoisomerases. *Nucleic acids research* **2007,** *35* (18), 6075-6085.

375. Sutormin, D.; Rubanova, N.; Logacheva, M.; Ghilarov, D.; Severinov, K., Single-nucleotide-resolution mapping of DNA gyrase cleavage sites across the Escherichia coli genome. *Nucleic acids research* **2019,** *47* (3), 1373-1388.

376. Kutzner, C.; Kniep, C.; Cherian, A.; Nordstrom, L.; Grubmüller, H.; de Groot, B. L.; Gapsys, V., GROMACS in the Cloud: A Global Supercomputer to Speed Up Alchemical Drug Design. *Journal of Chemical Information and Modeling* **2022**.

377. Bax, B. D.; Chan, P. F.; Eggleston, D. S.; Fosberry, A.; Gentry, D. R.; Gorrec, F.; Giordano, I.; Hann, M. M.; Hennessy, A.; Hibbs, M.; Huang, J.; Jones, E.; Jones, J.; Brown, K. K.; Lewis, C. J.; May, E. W.; Saunders, M. R.; Singh, O.; Spitzfaden, C. E.; Shen, C.; Shillings, A.; Theobald, A. J.; Wohlkonig, A.; Pearson, N. D.; Gwynn, M. N., Type IIA topoisomerase inhibition by a new class of antibacterial agents. *Nature* **2010,** *466* (7309), 935-40.

378. Laponogov, I.; Pan, X. S.; Veselkov, D. A.; Cirz, R. T.; Wagman, A.; Moser, H. E.; Fisher, L. M.; Sanderson, M. R., Exploring the active site of the Streptococcus pneumoniae topoisomerase IV-DNA cleavage complex with novel 7,8-bridged fluoroquinolones. *Open Biol* **2016,** *6* (9).

379. Vanden Broeck, A.; Lotz, C.; Ortiz, J.; Lamour, V., Cryo-EM structure of the complete E. coli DNA gyrase nucleoprotein complex. *Nature communications* **2019,** *10* (1), 4935.

380. Eastman, P.; Swails, J.; Chodera, J. D.; McGibbon, R. T.; Zhao, Y.; Beauchamp, K. A.; Wang, L.-P.; Simmonett, A. C.; Harrigan, M. P.; Stern, C. D.; Wiewiora, R. P.; Brooks, B. R.; Pande, V. S., OpenMM 7: Rapid development of high performance algorithms for molecular dynamics. *PLoS Comput. Biol* **2017,** *13* (7), e1005659-e1005659.

381. Brankin, A. E.; Fowler, P. W., Predicting Resistance Is (Not) Futile. *ACS Central Science* **2019,** *5* (8), 1312-1314.

9 Appendix

Appendix Table 1 Features selected by recursive feature elimination for binary machine learning classifiers to predict levofloxacin resistance. 1 indicates that a feature was selected.

Feature	Logistic Regression	Random Forest	XGBoost
Atom change	1	1	
Volume change	1	1	1
Hydropathy change		1	
Charge change	1	1	
H-donor change		1	
H-acceptor change	1	1	
Rotable bond change		1	
Aromatic ring change	1	1	1
Sulphur change	1	1	
Snap2 score		1	
Residue depth		1	
Ligand distance	1	1	1
Ligand Mg distance	1	1	1
Catalytic Mg distance	1	1	1
DNA distance	1	1	1
Relative solvent accessibility		1	1
Stability change	1	1	
Protein H-bonds		1	1
DNA H-bonds	1	1	
Protein-protein affinity change	1	1	
In gyrA	1	1	
In gyrB	1	1	
Secondary structure B			
Secondary structure E	1	1	
Secondary structure G	1		
Secondary structure H	1	1	1
Secondary structure I			
Secondary structure unknown	1	1	
Secondary structure T	1	1	
Secondary structure S			
Lineage 1		1	
Lineage 2	1	1	1
Lineage 3	1	1	
Lineage 4	1	1	1
Lineage 6			

Ligand interaction none	1	1	
Ligand interaction unknown	1	1	
Ligand interaction salt bridge			
Ligand interaction water bridge	1	1	
Mixed allele	1	1	1
Isoniazid and rifampicin susceptible	1	1	1
Isoniazid mono resistant	1	1	
Rifampicin mono resistant	1	1	
Multi drug resistant	1	1	1
Isoniazid and rifampicin resistance unknown	1	1	
Sample from Brazil	1	1	
Sample from China	1	1	1
Sample from Germany	1	1	
Sample from India	1	1	1
Sample from Italy		1	
Sample from Kyrgyzstan	1	1	
Sample from Nepal	1	1	1
Sample from Pakistan		1	
Sample from Peru	1	1	1
Sample from Tajikistan	1	1	
Sample from Turkmenistan	1	1	
Sample from Ukraine	1	1	
Sample from Vietnam	1	1	1
Sample from South Africa	1	1	1
Total features selected	**44**	**53**	**20**

Appendix Table 2 Features selected by recursive feature elimination for binary machine learning classifiers to predict moxifloxacin resistance. 1 indicates that a feature was selected.

	Logistic Regression	Random Forest	XGBoost
Atom change	1	1	1
Volume change	1	1	1
Hydropathy change	1	1	1
Charge change		1	1
H-donor change	1	1	1
H-acceptor change	1	1	1
Rotable bond change	1	1	
Aromatic ring change	1		1
Sulphur change	1		
Snap2 score		1	1
Residue depth	1	1	
Ligand distance		1	1

Ligand Mg distance	1	1	1
Catalytic Mg distance		1	1
DNA distance	1	1	1
Relative solvent accessibility		1	
Stability change	1	1	1
Protein H-bonds	1	1	1
DNA H-bonds	1	1	
Protein-protein affinity change	1	1	1
In gyrA	1		1
In gyrB	1		
Secondary structure B			
Secondary structure E	1		
Secondary structure G			
Secondary structure H		1	
Secondary structure I			
Secondary structure unknown	1	1	
Secondary structure T	1		
Secondary structure S	1		
Lineage 1		1	1
Lineage 2	1	1	1
Lineage 3	1	1	1
Lineage 4	1	1	1
Lineage 6			
Ligand interaction none		1	
Ligand interaction unknown	1	1	
Ligand interaction salt bridge			
Ligand interaction water bridge			
Mixed allele	1	1	1
Isoniazid and rifampicin susceptible	1	1	1
Isoniazid mono resistant	1	1	1
Rifampicin mono resistant	1		1
Multi drug resistant	1	1	1
Isoniazid and rifampicin resistance unknown	1	1	1
Sample from Brazil			
Sample from China	1	1	1
Sample from Germany	1	1	1
Sample from India	1	1	1
Sample from Italy	1	1	1
Sample from Kyrgyzstan	1	1	
Sample from Nepal	1	1	1
Sample from Pakistan		1	1
Sample from Peru	1	1	1
Sample from Tajikistan			

Sample from Turkmenistan	1	1	
Sample from Ukraine			
Sample from Vietnam	1		
Sample from South Africa	1	1	1
Total features selected	**41**	**41**	**34**

Appendix Table 3 *gyrA* mutations predicted by an xgboost classifier model to be resistant to levofloxacin.

POSITION	PREDICTED R MUTATIONS
R26	F, H, Y
D30	F, H, W, Y
Y31	W
I36	F
V37	H
R39	F, H, I, K, L, M, W, Y
A40	C, D, E, F, H, I, K, L, M, P, Q, R, S, V, W, Y
L41	F
E43	H
L48	F
K49	F, H, I, L
V51	A, C, D, E, F, G, H, I, K, L, M, P, Q, R, S, W, Y
H52	A, C, D, E, F, G, I, K, L, M, N, P, Q, R, S, T, V, W, Y
V55	F, H, Y
D61	F, H, W, Y
H70	I, K, L, M, R
K72	F
S73	H
A74	C, D, E, F, G, H, I, K, L, M, N, P, Q, R, S, T, V, W, Y
V77	E, F, H, I, K, L, M, P, Q, R, W, Y
A78	C, D, F, H, S, W, Y
M81	A, C, D, E, F, G, H, I, K, L, N, P, Q, R, S, T, V, W, Y
H85	I, K, L, M, R, W
P86	C, D, E, F, H, I, K, L, M, N, Q, R, T, V, W, Y
H87	A, C, D, E, F, G, I, K, L, M, N, P, Q, R, S, T, V, W, Y
G88	A, C, D, E, F, H, I, K, L, M, N, P, Q, R, S, T, V, W, Y
D89	C, E, F, H, I, K, L, M, N, P, Q, R, T, V, W, Y
A90	C, D, E, F, G, H, I, K, L, M, N, P, Q, R, S, T, V, W, Y
S91	A, C, D, E, F, G, H, I, K, L, M, N, P, Q, R, T, V, W, Y
I92	F, H, K, L, M, R, W, Y
Y93	W
D94	A, C, E, F, G, H, I, K, L, M, N, P, Q, R, S, T, V, W, Y

S95	A, C, D, F, H, I, K, L, M, P, Q, W, Y
L96	F, H, I, K, M, R, W, Y
V97	E, F, H, I, K, L, M, P, Q, R, W, Y
R98	A, C, D, E, F, G, H, I, K, L, M, N, P, Q, S, T, V, W, Y
Q101	F, H, I, L, M, Y
L105	F
R106	F, Y
L109	F
F116	W, Y
D122	C, F, H, N, P, T
P123	D, F, H, I, K, L, M, N, R, T, W, Y
P124	D, F, H, I, K, L, M, N, Q, R, T, W, Y
A125	C, D, E, F, G, H, I, K, L, M, N, P, Q, R, S, T, V, W, Y
A126	C, D, E, F, H, I, K, L, M, N, P, Q, R, S, T, V, W, Y
M127	F, H, I, K, L, R, W, Y
R128	A, C, D, E, F, G, H, I, K, L, M, N, P, Q, S, T, V, W, Y
Y129	A, C, D, E, F, G, H, I, K, L, M, N, P, Q, R, S, T, V, W
T130	E, F, N, Y
R133	F, Y
L134	F
R143	F, Y
E162	H
L173	F
I181	F, K, L, R, W, Y
V278	H
N279	D
H280	W
D281	F, H
T285	F, H, W, Y
E289	F, H, W, Y
Q305	H
L346	F
Y364	W
D367	F, H, W, Y
H490	I, K, L, M, R, W

Appendix Table 4 *gyrB* mutations predicted by an xgboost classifier model to be resistant to levofloxacin.

POSITION	PREDICTED R MUTATIONS
R429	F, Y
K441	F

R446	F, H, Y
E454	H
E459	F, H, Q, V, W, Y
G460	A, C, D, E, F, H, I, K, L, M, N, P, Q, R, S, T, V, W, Y
D461	A, C, E, F, G, H, I, K, L, M, N, P, Q, R, S, T, V, W, Y
S462	H
K468	F
L481	F, H, I, K, R, W, Y
R482	A, C, D, E, F, G, H, I, K, L, M, N, P, Q, S, T, V, W, Y
G483	A, C, D, E, F, H, I, K, L, M, N, P, Q, R, S, T, V, W, Y
K484	A, C, D, E, F, G, H, I, L, M, N, P, Q, R, S, T, V, W, Y
I485	F, H, K, L, R, W, Y
R495	F, Y
N499	C, D, E, F, H, I, K, L, M, P, R, T, W, Y
Q503	H
D536	C, F, H, I, K, L, M, N, P, Q, R, T, V, W, Y
H539	I, K, L, M, R, W
I540	F, H, K, L, M, R, W, Y
R550	F, Y
E557	H, W, Y
K570	F
K572	F
R587	F, Y
D588	W, Y
G589	F, H, V, W, Y
L595	F, W, Y
K611	F
E615	H
K619	F
E623	H
R631	F, Y
R634	F, Y
A642	F, H, W, Y
R659	F, H, W, Y
S661	F, H, W, Y
R665	W

Appendix Table 5 *gyrA* mutations predicted by a random forest classifier model to be resistant to moxifloxacin.

POSITION	PREDICTED R MUTATIONS
Y31	A, C, D, E, G, H, I, K, L, M, N, P, Q, R, S, T, V, W

R39	G
A40	D, E, F, H, K, L, N, P, Q, R, W, Y
V51	D
H85	D, E, G, K, P, R, S, T
H87	D
G88	A, C, D, E, F, H, I, K, L, M, N, P, Q, R, S, T, V, W, Y
D89	A, C, E, F, G, H, I, K, L, M, N, P, Q, R, S, T, V, W, Y
A90	C, D, E, F, G, H, I, K, L, M, N, P, Q, R, V, W, Y
S91	C, D, E, F, G, H, I, K, L, M, N, P, Q, R, T, V, W, Y
I92	A, D, E, F, G, H, K, N, P, Q, R, S, T, W, Y
Y93	A, D, E, G, H, K, N, P, Q, R, S, T, W
D94	A, C, E, F, G, H, I, K, L, M, N, P, Q, R, S, T, V, W, Y
S95	D, E, F, H, K, N, Q, R, W, Y
M99	K
N115	W
P124	K, R
A125	D, E, K, P, R, W
A126	D, E, F, H, K, P, R, W, Y
R128	A, C, D, E, F, G, H, I, L, M, N, P, Q, S, T, V, W, Y
Y129	A, C, D, E, F, G, H, I, K, L, M, N, P, Q, R, S, T, V, W

Appendix Table 6 *gyrB* mutations predicted by a random forest classifier model to be resistant to moxifloxacin.

POSITION	PREDICTED R MUTATIONS
E459	G, H, K, N, P, R, S, T, W, Y
G460	A, C, D, E, F, H, I, K, L, M, N, P, Q, R, S, T, V, W, Y
D461	P, W, Y
S462	D, E, F, I, K, L, P, R, V, W, Y
L481	D, E, G, H, K, N, P, R, W
R482	A, C, D, E, F, G, H, I, L, M, N, P, Q, S, T, V, Y
G483	A, C, D, E, F, H, I, K, L, M, N, P, Q, R, S, T, V, W, Y
K484	A, C, D, E, F, G, H, I, L, M, N, P, Q, R, S, T, V, W, Y
I485	D, E, G, H, K, N, R, W, Y
N499	D, E, F, G, K, Q, R, W
E501	A, C, D, F, G, H, I, K, L, M, N, P, Q, R, S, V, W, Y
D536	K, R, W, Y
G537	K, R
H539	A, C, D, E, G, K, L, N, P, Q, R, S, T
I540	A, D, E, F, G, H, K, M, N, P, Q, R, S, T, W, Y
L543	K, R

Appendix Table 7 Features selected by recursive feature elimination for machine learning classifiers to predict levofloxacin MIC. 1 indicates that a feature was selected.

	Ordinal All Threshold	Ordinal Intermediate Threshold	Ordinal Squared Error	Ordinal Ridge	Random Forest	XGBoost
Atom change	1	1	1	1	1	1
Volume change			1		1	1
Hydropathy change		1	1		1	1
Charge change	1	1	1	1		1
H-donor change	1	1	1	1		1
H-acceptor change	1	1	1	1	1	1
Rotable bond change	1	1	1	1	1	1
Aromatic ring change	1	1	1	1		1
Sulphur change	1	1	1	1		
Snap2 score			1		1	1
Residue depth	1	1	1	1	1	1
Ligand distance	1	1	1	1	1	1
Ligand Mg distance	1	1	1	1	1	1
Catalytic Mg distance	1	1	1	1	1	1
DNA distance		1	1	1	1	1
Relative solvent accessibility	1				1	
Stability change	1	1	1	1	1	1
Protein H-bonds	1	1	1	1		1
DNA H-bonds	1	1	1	1		
Protein-protein affinity change	1	1	1	1	1	1
In gyrA	1	1	1			1
In gyrB	1	1	1			
Secondary structure B	1	1	1	1		
Secondary structure E	1	1	1	1		
Secondary structure G	1	1	1	1		
Secondary structure H	1		1	1		1
Secondary structure I						
Secondary structure unknown	1	1	1	1		
Secondary structure T	1	1	1	1		1
Secondary structure S	1	1	1	1		
Lineage 1	1	1	1	1		
Lineage 2		1	1	1	1	1
Lineage 3	1	1	1	1	1	
Lineage 4	1	1	1	1	1	1
Lineage 6						

Ligand interaction none	1	1	1	1		
Ligand interaction unknown	1	1	1	1		
Ligand interaction salt bridge						
Ligand interaction water bridge		1	1	1		
Mixed allele	1	1	1	1	1	
Isoniazid and rifampicin susceptible	1	1	1	1	1	1
Isoniazid mono resistant	1		1	1	1	1
Rifampicin mono resistant	1	1	1	1		1
Multi drug resistant	1	1	1	1	1	1
Isoniazid and rifampicin resistance unknown	1	1	1	1		1
Sample from Brazil	1	1	1	1		
Sample from China	1	1	1	1	1	1
Sample from Germany	1	1	1	1		
Sample from India	1	1	1	1	1	1
Sample from Italy	1	1	1	1		
Sample from Kyrgyzstan	1	1	1	1		
Sample from Nepal	1	1	1	1		1
Sample from Pakistan	1	1	1	1	1	1
Sample from Peru	1	1	1	1	1	
Sample from Tajikistan	1	1	1	1		
Sample from Turkmenistan	1	1	1	1		
Sample from Ukraine	1	1	1	1		1
Sample from Vietnam	1	1	1	1		1
Sample from South Africa	1	1	1	1	1	1
Total features selected	**50**	**51**	**55**	**50**	**26**	**34**

Appendix Table 8 Features selected by recursive feature elimination for machine learning classifiers to predict moxifloxacin MIC. 1 indicates that a feature was selected.

	Ordinal All Threshold	Ordinal Intermediate Threshold	Ordinal Squared Error	Ordinal Ridge	Random Forest	XGBoost
Atom change	1	1	1		1	1
Volume change			1		1	1
Hydropathy change		1	1	1	1	1
Charge change	1	1	1	1	1	1
H-donor change	1	1	1	1	1	1

H-acceptor change	1	1	1	1	1	1
Rotable bond change	1	1	1	1	1	1
Aromatic ring change	1	1	1	1	1	1
Sulphur change	1	1	1	1	1	1
Snap2 score					1	1
Residue depth	1		1	1	1	1
Ligand distance	1	1	1	1	1	1
Ligand Mg distance	1	1	1	1	1	1
Catalytic Mg distance		1	1		1	1
DNA distance	1	1	1	1	1	1
Relative solvent accessibility		1	1		1	1
Stability change	1	1	1	1	1	1
Protein H-bonds	1		1	1	1	1
DNA H-bonds	1	1	1	1	1	
Protein-protein affinity change	1	1	1	1	1	1
In gyrA	1	1	1		1	1
In gyrB	1	1	1		1	
Secondary structure B	1	1	1	1		
Secondary structure E	1	1	1	1	1	1
Secondary structure G	1	1	1	1	1	
Secondary structure H	1	1	1	1	1	1
Secondary structure I						
Secondary structure unknown	1	1	1	1	1	
Secondary structure T	1	1	1	1	1	1
Secondary structure S	1	1	1	1	1	1
Lineage 1		1	1	1	1	1
Lineage 2	1	1	1	1	1	1
Lineage 3	1	1	1	1	1	1
Lineage 4	1	1	1	1	1	1
Lineage 6			1			
Ligand interaction none		1	1	1	1	1
Ligand interaction unknown	1	1	1	1	1	
Ligand interaction salt bridge						
Ligand interaction water bridge	1	1	1	1	1	
Mixed allele	1	1	1	1	1	1
Isoniazid and rifampicin susceptible			1	1	1	1
Isoniazid mono resistant	1	1	1	1	1	1
Rifampicin mono resistant	1	1	1	1	1	1

Multi drug resistant	1	1	1	1	1	1
Isoniazid and rifampicin resistance unknown	1	1	1	1	1	1
Sample from Brazil	1	1	1	1	1	
Sample from China	1	1	1	1	1	1
Sample from Germany	1	1	1	1	1	1
Sample from India	1	1	1	1	1	1
Sample from Italy		1	1	1	1	1
Sample from Kyrgyzstan	1	1	1	1	1	
Sample from Nepal	1	1	1	1	1	1
Sample from Pakistan	1	1	1	1	1	1
Sample from Peru	1	1	1	1	1	1
Sample from Tajikistan	1	1	1	1	1	
Sample from Turkmenistan	1	1	1	1	1	
Sample from Ukraine	1	1	1	1	1	
Sample from Vietnam	1	1		1	1	1
Sample from South Africa	1	1	1	1	1	1
Total features selected	**47**	**51**	**55**	**49**	**55**	**44**

Appendix Table 9 Individual alchemical free energy values used to calculate RBFE for DNA gyrase mutations. * indicates a *gyrB* mutation.

MUTATION	LEG	STEP	ΔG
S95T	apo	vdw	3.51
S95T	apo	vdw	7.87
S95T	apo	vdw	4.03
S95T	apo	vdw	3.12
S95T	apo	vdw	6.74
S95T	apo	vdw	6.37
S95T	apo	vdw	7.25
S95T	apo	vdw	7.12
S95T	apo	vdw	8.57
S95T	apo	vdw	7.51
S95T	apo	vdw	5.42
S95T	apo	vdw	3.09
S95T	apo	vdw	3.65
S95T	apo	vdw	6.97
S95T	apo	vdw	7.03
S95T	apo	qoff	-14.54
S95T	apo	qoff	-14.44
S95T	apo	qoff	-14.37

S95T	apo	qoff	-14.68
S95T	apo	qoff	-14.69
S95T	apo	qon	-29.78
S95T	apo	qon	-29.97
S95T	apo	qon	-29.69
S95T	apo	qon	-29.67
S95T	apo	qon	-29.96
S95T	mfx-bound	vdw	6.40
S95T	mfx-bound	vdw	5.59
S95T	mfx-bound	vdw	7.57
S95T	mfx-bound	vdw	3.36
S95T	mfx-bound	vdw	1.46
S95T	mfx-bound	vdw	5.14
S95T	mfx-bound	vdw	6.50
S95T	mfx-bound	vdw	2.37
S95T	mfx-bound	vdw	2.73
S95T	mfx-bound	vdw	5.72
S95T	mfx-bound	vdw	3.77
S95T	mfx-bound	vdw	3.76
S95T	mfx-bound	vdw	3.27
S95T	mfx-bound	vdw	1.17
S95T	mfx-bound	vdw	5.28
S95T	mfx-bound	qoff	-14.48
S95T	mfx-bound	qoff	-14.41
S95T	mfx-bound	qoff	-13.38
S95T	mfx-bound	qoff	-14.95
S95T	mfx-bound	qoff	-14.90
S95T	mfx-bound	qon	-29.77
S95T	mfx-bound	qon	-29.86
S95T	mfx-bound	qon	-30.90
S95T	mfx-bound	qon	-29.21
S95T	mfx-bound	qon	-29.66
A90V	apo	vdw	10.07
A90V	apo	vdw	9.26
A90V	apo	vdw	6.91
A90V	apo	vdw	9.54
A90V	apo	vdw	8.92
A90V	apo	vdw	9.67
A90V	apo	vdw	7.84
A90V	apo	vdw	8.89
A90V	apo	vdw	8.43
A90V	apo	vdw	8.26
A90V	apo	qoff	-8.62

A90V	apo	qoff	-8.46
A90V	apo	qoff	-8.59
A90V	apo	qoff	-8.55
A90V	apo	qoff	-8.56
A90V	apo	qon	-38.00
A90V	apo	qon	-38.45
A90V	apo	qon	-37.44
A90V	apo	qon	-36.85
A90V	mfx-bound	vdw	8.70
A90V	mfx-bound	vdw	9.81
A90V	mfx-bound	vdw	9.67
A90V	mfx-bound	vdw	11.11
A90V	mfx-bound	vdw	12.11
A90V	mfx-bound	vdw	11.73
A90V	mfx-bound	vdw	10.60
A90V	mfx-bound	vdw	7.66
A90V	mfx-bound	vdw	9.69
A90V	mfx-bound	vdw	9.67
A90V	mfx-bound	vdw	11.60
A90V	mfx-bound	vdw	8.38
A90V	mfx-bound	vdw	12.65
A90V	mfx-bound	vdw	11.48
A90V	mfx-bound	vdw	11.48
A90V	mfx-bound	qoff	-8.74
A90V	mfx-bound	qoff	-8.76
A90V	mfx-bound	qoff	-8.62
A90V	mfx-bound	qoff	-8.66
A90V	mfx-bound	qoff	-8.79
A90V	mfx-bound	qon	-36.62
A90V	mfx-bound	qon	-36.71
A90V	mfx-bound	qon	-37.95
A90V	mfx-bound	qon	-36.91
A90V	mfx-bound	qon	-37.89
A90S	apo	vdw	-0.84
A90S	apo	vdw	-1.30
A90S	apo	vdw	-0.73
A90S	apo	vdw	-1.01
A90S	apo	vdw	-0.99
A90S	apo	qoff	-7.81
A90S	apo	qoff	-7.74
A90S	apo	qoff	-7.80
A90S	apo	qoff	-7.78
A90S	apo	qoff	-7.82

A90S	apo	qon	-13.13
A90S	apo	qon	-13.70
A90S	apo	qon	-15.00
A90S	apo	qon	-14.21
A90S	apo	qon	-13.86
A90S	mfx-bound	vdw	-2.52
A90S	mfx-bound	vdw	-1.33
A90S	mfx-bound	vdw	-1.14
A90S	mfx-bound	vdw	-1.45
A90S	mfx-bound	vdw	-0.85
A90S	mfx-bound	qoff	-7.82
A90S	mfx-bound	qoff	-8.02
A90S	mfx-bound	qoff	-7.85
A90S	mfx-bound	qoff	-7.96
A90S	mfx-bound	qoff	-7.93
A90S	mfx-bound	qon	-13.08
A90S	mfx-bound	qon	-13.15
A90S	mfx-bound	qon	-13.86
A90S	mfx-bound	qon	-15.88
A90S	mfx-bound	qon	-15.80
A90S	mfx-bound	qon	-14.84
A90S	mfx-bound	qon	-14.72
A90S	mfx-bound	qon	-14.54
E501D*	apo	vdw	-35.05
E501D*	apo	vdw	-37.30
E501D*	apo	vdw	-36.18
E501D*	apo	vdw	-32.85
E501D*	apo	vdw	-38.11
E501D*	apo	vdw	-39.04
E501D*	apo	vdw	-38.01
E501D*	apo	vdw	-36.34
E501D*	apo	vdw	-32.35
E501D*	apo	vdw	-36.64
E501D*	apo	qoff	29.72
E501D*	apo	qoff	21.65
E501D*	apo	qoff	24.29
E501D*	apo	qoff	27.40
E501D*	apo	qoff	23.99
E501D*	apo	qoff	28.72
E501D*	apo	qoff	24.31
E501D*	apo	qoff	23.77
E501D*	apo	qoff	26.67
E501D*	apo	qoff	24.51

E501D*	apo	qon	9.45
E501D*	apo	qon	9.14
E501D*	apo	qon	13.19
E501D*	apo	qon	6.54
E501D*	apo	qon	8.93
E501D*	apo	qon	9.72
E501D*	apo	qon	9.42
E501D*	apo	qon	10.37
E501D*	apo	qon	7.53
E501D*	apo	qon	6.78
E501D*	mfx-bound	vdw	-36.06
E501D*	mfx-bound	vdw	-36.01
E501D*	mfx-bound	vdw	-39.56
E501D*	mfx-bound	vdw	-38.26
E501D*	mfx-bound	vdw	-34.47
E501D*	mfx-bound	vdw	-38.97
E501D*	mfx-bound	vdw	-37.02
E501D*	mfx-bound	vdw	-39.79
E501D*	mfx-bound	vdw	-36.32
E501D*	mfx-bound	vdw	-37.67
E501D*	mfx-bound	qoff	24.27
E501D*	mfx-bound	qoff	25.29
E501D*	mfx-bound	qoff	26.94
E501D*	mfx-bound	qoff	28.36
E501D*	mfx-bound	qoff	24.64
E501D*	mfx-bound	qoff	26.96
E501D*	mfx-bound	qoff	25.18
E501D*	mfx-bound	qoff	25.42
E501D*	mfx-bound	qoff	24.30
E501D*	mfx-bound	qoff	26.08
E501D*	mfx-bound	qon	8.66
E501D*	mfx-bound	qon	7.70
E501D*	mfx-bound	qon	8.07
E501D*	mfx-bound	qon	9.19
E501D*	mfx-bound	qon	8.06
E501D*	mfx-bound	qon	8.55
E501D*	mfx-bound	qon	7.32
E501D*	mfx-bound	qon	9.39
E501D*	mfx-bound	qon	5.50
D94G	apo	vdw	-156.98
D94G	apo	vdw	-161.34
D94G	apo	vdw	-160.90
D94G	apo	vdw	-160.15

D94G	apo	qoff	55.38
D94G	apo	qoff	53.18
D94G	apo	qoff	48.45
D94G	apo	qoff	49.72
D94G	apo	qoff	57.07
D94G	apo	qoff	52.66
D94G	apo	qoff	48.43
D94G	apo	qoff	56.44
D94G	apo	qon	388.72
D94G	apo	qon	392.77
D94G	apo	qon	391.86
D94G	apo	qon	405.55
D94G	apo	qon	390.88
D94G	apo	qon	390.47
D94G	apo	qon	407.33
D94G	apo	qon	389.67
D94G	mfx-bound	vdw	-156.69
D94G	mfx-bound	vdw	-159.38
D94G	mfx-bound	vdw	-160.30
D94G	mfx-bound	vdw	-161.84
D94G	mfx-bound	vdw	-156.19
D94G	mfx-bound	qoff	46.90
D94G	mfx-bound	qoff	57.64
D94G	mfx-bound	qoff	57.21
D94G	mfx-bound	qoff	51.80
D94G	mfx-bound	qoff	52.53
D94G	mfx-bound	qoff	44.54
D94G	mfx-bound	qoff	47.36
D94G	mfx-bound	qoff	47.85
D94G	mfx-bound	qoff	51.54
D94G	mfx-bound	qoff	52.53
D94G	mfx-bound	qon	380.66
D94G	mfx-bound	qon	396.02
D94G	mfx-bound	qon	395.83
D94G	mfx-bound	qon	395.47
D94G	mfx-bound	qon	392.75
D94G	mfx-bound	qon	393.71
D94G	mfx-bound	qon	392.72
D94G	mfx-bound	qon	393.59
D94G	mfx-bound	qon	392.30
D94G	mfx-bound	qon	393.13