Prediction of pyrazinamide resistance in *Mycobacterium tuberculosis* using structure-based machine learning approaches.

Joshua J Carter[1], Timothy M Walker[1], A Sarah Walker[1,2], Michael G. Whitfield[3‡], Glenn P. Morlock[4], Timothy EA Peto[1,2], James E. Posey[4], Derrick W Crook[1,2,5], Philip W Fowler[1,2*]

[1] Nuffield Department of Medicine, University of Oxford, John Radcliffe Hospital, Headley Way, Oxford, OX3 9DU, UK
[2] National Institute of Health Research Oxford Biomedical Research Centre, John Radcliffe Hospital, Headley Way, Oxford, OX3 9DU, UK
[3] Division of Molecular Biology and Human Genetics, SAMRC Centre for Tuberculosis Research, DST/NRF Centre of Excellence for Biomedical Tuberculosis Research, Faculty of Medicine and Health Sciences, Stellenbosch University, Tygerberg, South Africa
[4] Division of Tuberculosis Elimination, National Center for HIV/AIDS, Viral Hepatitis, STD, and TB Prevention, Centers for Disease Control and Prevention, Atlanta, Georgia, United States
[5] NIHR Health Protection Research Unit in Healthcare Associated Infection and Antimicrobial Resistance at University of Oxford in partnership with Public Health England, Oxford, UK
‡ on behalf of the "EXIT-RIF" investigators: Prof Robin M Warren, Prof Annelies van Rie, Prof Lesley Scott, Prof Wendy Stevens

[*] Corresponding author and Lead Contact: philip.fowler@ndm.ox.ac.uk, @philipwfowler

Summary

Pyrazinamide is one of four first-line antibiotics used to treat tuberculosis, however antibiotic susceptibility testing for pyrazinamide is problematic. Resistance to pyrazinamide is primarily driven by genetic variation in *pncA,* an enzyme that converts pyrazinamide into its active form. We curated a derivation dataset of 291 non-redundant, missense amino acid mutations in *pncA* with associated high-confidence phenotypes from published studies and then trained three different machine learning models to predict pyrazinamide resistance based on sequence- and structure-based features of each missense mutation. The clinical performance of the models was estimated by predicting the binary pyrazinamide resistance phenotype of 2,292 clinical isolates harboring missense mutations in *pncA*. Overall, this work offers an approach to improve the sensitivity/specificity of pyrazinamide resistance prediction in genetics-based clinical microbiology workflows, highlights novel mutations for future biochemical investigation, and is a proof of concept for using this approach in other drugs such as bedaquiline.

Introduction

Mycobacterium tuberculosis is an evolutionarily ancient human pathogen that is the leading cause of death by infectious disease worldwide[1]. In 2017, tuberculosis was responsible for 1.6 million deaths and 10 million new infections[1]. Tuberculosis control efforts have been hampered by the evolution of resistance to antibiotics, threatening the efficacy of the standard four drug antibiotic regimen consisting of rifampicin, isoniazid, ethambutol, and pyrazinamide[1,2]. Pyrazinamide plays a critical role in tuberculosis treatment through its specific action on slow-growing, "persister" bacteria that often tolerate other drugs due to their reduced metabolism[3–6]. This unique activity has been instrumental in shortening the standard treatment duration to six months, substantially increasing the effectiveness of antibiotic therapy[5,6]. Numerous studies have also found that including pyrazinamide in treatment regimens increases sputum-conversion rates in both pan-susceptible and multidrug-resistant (defined as resistant to rifampicin and isoniazid) tuberculosis populations[7]. Due to its unique sterilizing effect and its synergy with new tuberculosis drugs such as bedaquiline, pyrazinamide is also included in most new treatment regimens targeting drug-resistant tuberculosis[8–13]. Therefore, accurately and rapidly determining whether a clinical isolate is resistant to pyrazinamide is critically important for the treatment of tuberculosis.

The majority of culture-based laboratory methods to determine pyrazinamide resistance are technically challenging, requiring highly-trained technicians. Even then, results are often not reproducible, meaning these methods are rarely employed in low-resource and/or high-burden clinical settings[14]. Even the current gold standard, the Mycobacteria Growth Indicator Tube (MGIT), which is relatively simple to use, can suffer from low precision, with false-resistance rates of 1-68% reported[15–21]. As the prevalence of multidrug-resistant and extensively drug-resistant TB increases, this lack of precision will become more of a problem and hence the WHO is evaluating the efficacy of genetics-based approaches such as line-probe assays or whole genome sequencing for all first-line tuberculosis antibiotics during 2019[22].

Resistance to rifampicin or isoniazid can be predicted in most isolates (90-95% and 50-97%, respectively) by the presence of a small number of highly-penetrant genetic variants in short and well-delineated regions of one or two genes (*rpoB* and *katG*/*fabG1*, respectively)[4]. However, despite pyrazinamide being used to treat tuberculosis since 1952, comparatively less is known about which genetic variants confer resistance compared to other first-line drugs[5]. In a recent study aimed at assessing the efficacy of whole-genome-based approaches for prediction of resistance in *M. tuberculosis*, the performance for pyrazinamide was markedly lower (75.8% sensitivity and 92.4% specificity) than either rifampicin or isoniazid (94.5% and 93.6% or 93.1% and 94.1% sensitivity and specificity respectively)[23]. While some of this poor performance is likely due to inaccuracies in phenotypic testing, a comprehensive genetic catalog for pyrazinamide resistance mutations remains elusive.

Pyrazinamide is a pro-drug that is converted to its active form of pyrazinoic acid by the action of PncA, a pyrazinamidase/nicotinamidase encoded by the *pncA* gene[24]. While other genetic loci have been implicated in pyrazinamide resistance (notably *rpsA*, *panD*, *clpC1*, and the putative efflux pumps *Rv0191*, *Rv3756c*, *Rv3008*, and *Rv1667c*), the majority (70-97%) of pyrazinamide-resistant clinical isolates harbor genetic variants in either the promoter region or coding sequence of *pncA*[25–35]. In contrast to the well-delineated and relatively restricted "resistance-determining regions" found in *rpoB* (rifampicin, 27 codons) and *katG* (isoniazid, single codon), pyrazinamide-resistant variants have been identified along the entire length of the *pncA* gene (**Figure 1A**) with no single variant predominating.

The consequence of this is that whilst line-probe assays have been successfully developed that predict resistance to rifampicin and isoniazid more quickly than culture-based methods, it is much more challenging to develop a line-probe assay with a high sensitivity for predicting pyrazinamide resistance while only targeting specific regions of *pncA*[36–38]. Alternatively, while targeted or whole-genome sequencing approaches are capable of assaying the entire *pncA* gene, the number and diversity of resistance-conferring variants in *pncA*

fundamentally limits the sensitivity and specificity of heuristic approaches that aim to predict the effectiveness of pyrazinamide based on a catalogue of previously-observed genetic variants[4,14,31,32,37,39,40].

Genetics-based clinical microbiology for tuberculosis depends on being able to predict or infer the effect of any likely occurring *pncA* mutation on pyrazinamide susceptibility. Recent studies to identify pyrazinamide-resistance determining mutations have focused on either classifying mutations from previously observed clinical isolates or discovering novel mutations through *in vitro/in vivo* screening approaches[23,32,41,42]. However, these strategies are constrained, respectively, by the relative paucity of sequenced clinical isolates compared to the number of potential resistance-causing mutations and the lack of laboratory capacity to systematically generate and test mutants. A computational modelling approach could potentially predict the effect of a significant number of missense mutations *before* they are observed in clinical isolates, allowing clinicians to more rapidly make an informed decision in the face of emerging resistance patterns as well as focusing future fundamental *in vitro* studies on the most important mutations to investigate.

In this paper, we demonstrate that machine-learning models can robustly and accurately predict the effect of missense amino acid mutations on pyrazinamide susceptibility. The models were trained using a derivation dataset of 291 non-redundant, missense amino acid mutations in PncA collected from pooled MGIT phenotypic studies and a comprehensive *in vitro*/*in vivo* pyrazinamide-resistance screen (**Methods, Table 1, S1**)[23,32,41,42]. This dataset reflects the clinically observed distribution of mutations across the *pncA* gene (**Figure 1A**), and consequently, throughout the protein structure. Since the *pncA* gene is not essential, our hypothesis is that missense mutations can confer resistance by altering the folding, stability or function of the PncA protein. This led us to consider how each mutation perturbs the local chemistry and overall structure of the protein. Hence, we used information about the structural and chemical properties and evolutionary context of the wild-type and mutant amino acids in

question as inputs for several different machine-learning models (**Methods**). The predictions from the best performing model were then re-applied to an aggregated clinical dataset to examine their clinical relevance and also validated against a smaller, independent quantitative dataset of *in vitro* pyrazinamide minimum inhibitory concentrations (MICs) to determine their capacity to also predict the degree of resistance for specific mutations (**Table 1**). Finally, the model was used to predict the effect of *all* (1105) non-synonymous substitutions possible from single nucleotide polymorphisms in *pncA* on the action of pyrazinamide. These data were then used to predict the occurrence of pyrazinamide resistance in a large dataset of 47,769 *M. tuberculosis* sequences downloaded from public genetic sequence repositories **(Table 1**)[43,44]**.** This study is a proof of principle for using computational approaches to model and predict antibiotic resistance in other drugs, such as bedaquiline, pretomanid/delamanid, isoniazid, and ethionamide, where some genes implicated in resistance pathways appear to be non-essential.

Results

Dataset	Phenotype	Isolate Sources	# Isolates	Non-redundant Missense Mutations
Exploratory	R/S/F	Ref [23]	2,651	254
Derivation	R/S	Ref [23,32,41,42,45]	1,792 + *in vitro* isolates[42]	291
Clinical	R/S	Ref [23,32,41]	2,292	272
Quantitative	MIC	EXIT-RIF Study + US CDC	71	27
Prevalence	None	European Nucleotide Archive and ref [23]	47,769	480

Table 1. Description of datasets employed in this study. (R=resistant to antibiotic, S=susceptible, F=test failed to return a result, MIC=Minimum inhibitory concentration)

An in depth analysis of the genetic variation reported in the initial exploratory dataset revealed 2,651 strains with variants in the open reading frame of *pncA*, a substantial majority (2,400) of which only harbored a single variant (**Table 1**)[23]. Of these, 2,261 (94.2%) were substitutions, with the remaining 139 (5.8%) comprising insertions, deletions and frameshifts. Insertions, deletions, and frameshifts, along with nonsense substitutions (collectively 190, 7.9% of the single variant strains), were all associated with pyrazinamide resistance, consistent with their likely disruption of the PncA enzyme, thereby preventing pyrazinamide activation. Most synonymous substitutions (present in 758, 31.6% of the single variant strains**, Figure 1B**) were not associated with resistance, however four variants were observed in resistant isolates. S65S (8 resistant isolates) is a phylogenetic SNP present in Lineage 1; however, it is susceptible in 607 strains, suggesting that these eight isolates are either phenotyping errors or that there is an alternative mechanism of pyrazinamide resistance at play in these strains. The remaining

mutations—P62P, A102A, and T168T—are each present a single time, limiting our ability to confidently associate these variants with resistance. Thus, non-synonymous substitution variants (present in 1,452, 60.5% of single variant strains) appear to represent the majority of the potential pyrazinamide resistance-causing variation in *M. tuberculosis*.

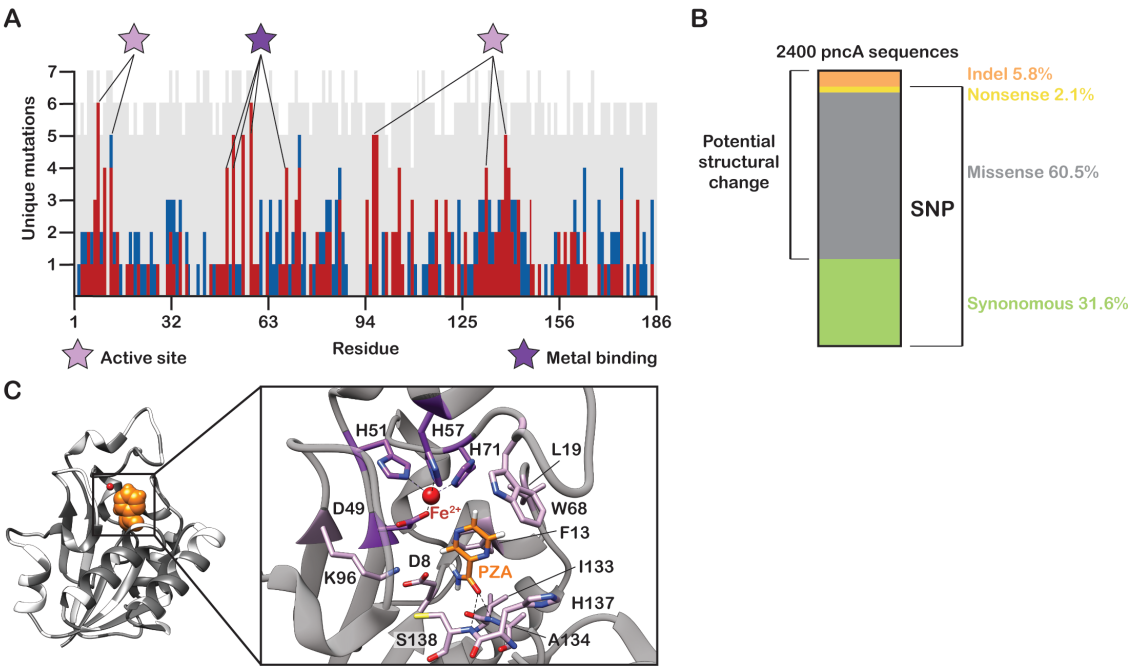

Figure 1. Distribution of PncA mutations from published datasets. (A) Barplot of the impact of possible missense mutations in PncA by amino acid position. High confidence resistant (red) and susceptible (blue) mutations are overlaid on the possible missense mutations whose effect on resistance is unknown or unclear (grey). (**B**) Distribution of the types of mutations reported by the CRyPTIC consortium *et al*[23]. (**C**) Missense mutations from the dataset plotted onto the PncA structure (PDB ID: 3PL1) in dark grey. A pyrazinamide molecule (orange) has been modeled into the active site.

Structural and evolutionary traits correlate with mutational impact on pyrazinamide susceptibility

Recent studies have implicated synonymous substitutions in altering protein folding and thus function through changes in the rate of translation[46–48]. This process is mediated through the different availabilities of complementary tRNA in the intracellular milieu[46,48]. Indeed, Yadon *et al* found that some synonymous substitutions were enriched in both *in vivo* and *in vitro* screens, suggesting that changes in the rate of protein translation offer a *bona fide* route to pyrazinamide resistance[42]. As previous studies have noted that functional synonymous mutations are enriched for at conserved codons, we tested to see if these variants were similarly enriched at conserved sites in *pncA*, but we found no significant difference. Given the dearth of isolates harboring sSNPs that have resistant phenotypes, the lack of data on tRNA concentrations in *M. tuberculosis* and, the lack of any association between resistance-causing sSNPs and conserved codons, we have no information from which to predict the effect of synonymous mutations. Further work to measure pyrazinamide MICs on large numbers of clinical strains is needed, as the effects of these sSNPs may result in sub-MIC shifts in resistance that could be clinically relevant[49].

We built a preliminary set of 722 non-synonymous substitutions in *pncA* that had either been observed multiple times in clinical isolates for which antibiotic susceptibility testing data was available or were generated in a high-precision laboratory screening study of *pncA* resistance variants (**Methods**). To create a high confidence dataset for model training, we discarded mutations for which there was any uncertainty around whether they conferred resistance (or not), which left us with a final derivation dataset of 291 missense mutations (**Table 1, Methods**). To understand the structural features that determine a mutation's effect on pyrazinamide susceptibility, we mapped our derivation dataset onto the PncA structure. No obvious clustering was revealed, consistent with the previously observed distribution of resistant mutations across the gene sequence and protein structure (**Figure 1A,C**)[14,31,32,39,42]. Interestingly, there were a significant number of *pncA* codons where different mutations

associated with either resistance or susceptibility were seen, suggesting that the change in local chemistry introduced by the mutant amino acid is an important factor in determining resistance (**Figure 1A**). The amino acid positions with the highest mutational diversity in the dataset were all residues involved in active site formation or metal binding, suggesting that, consistent with our hypothesis, loss or alteration of these functions is a common mechanism for gaining pyrazinamide resistance. Indeed, previous studies have noted a negative correlation between a mutation's distance from the active site and its tendency to cause resistance (**Figure S1**)[32,42,50].

Examining the PncA structure also suggested that resistant mutations were more likely to be buried in the protein core, consistent with findings from previous *in vitro* and *in vivo* screens (**Figure 2A**)[32,42]. Alterations in the hydrophobic core of a protein are likely to be destabilizing[51–54]. Indeed, some pyrazinamide-resistant mutations result in reduced production of functional PncA, perhaps due to impaired protein folding/stability[42,55]. To assess a *pncA* mutation's impact on the stability of PncA, we employed a meta-predictor that calculates the predicted change in free energy of protein unfolding *in silico*[56]. This is a fast, heuristic method; although other more accurate methods exist, these require vastly greater computational resources[57].

In addition to these chemical and structural properties, we also included information on the evolutionary variation at each position obtained from multiple sequence alignments of related orthologs (**Methods**). Unsurprisingly, increased conservation at a position was associated with a higher potential of a mutation at that position to confer resistance (**Figure S1**). Finally, we applied a recent computational method, called Multivariate Analysis of Protein Polymorphism (MAPP), that quantifies the evolutionary constraints imposed on a given position in a protein[58]. MAPP does this by combining the range of physicochemical amino acid properties observed at a particular position in a multiple sequence alignment with weights generated from the branch lengths of a phylogenetic tree[58]. Resistant mutations had

significantly higher MAPP scores, indicating that resistance-conferring mutations in *pncA* are less conservative in amino acid chemistry and function (**Figure 2B, S1**).

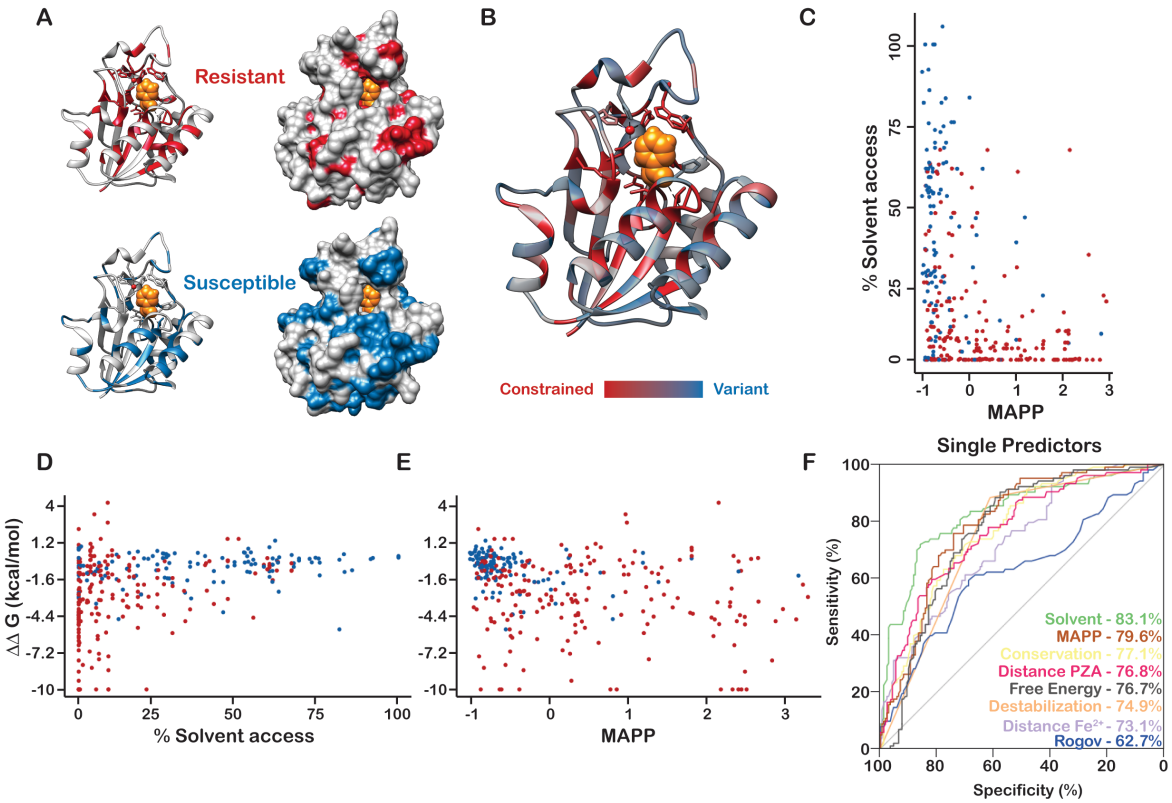

Figure 2: Structural and evolutionary traits correlate with mutational impact on pyrazinamide susceptibility. (A) PncA with resistant (red) or susceptible (blue) missense mutations shown. (**B)** PncA residue's median MAPP scores are shown (red denotes more deleterious). (**C-E)** Distributions of structural features that separate resistant (red) and susceptible (blue) missense mutations. (**F)** Performance of the individual features for prediction of pyrazinamide resistance plotted on a ROC curve. Grey line denotes random guessing.

Machine-learning models accurately predict pyrazinamide resistance

Univariable logistic regression over the derivation dataset revealed that most of the individual predictors were associated with resistant phenotypes (**Table S2**, **Figure 2F**). The MAPP score and solvent accessibility proved to be the most discriminatory individual features. As PncA can be inactivated through defects in protein folding, reduced stability, distortion of

active site geometry, abrogation of metal binding, or some combination of these, we expected a machine-learning approach to be ideally suited to simultaneously consider all these possible mechanisms of PncA inactivation, and hence more accurately predict pyrazinamide resistance/susceptibility.

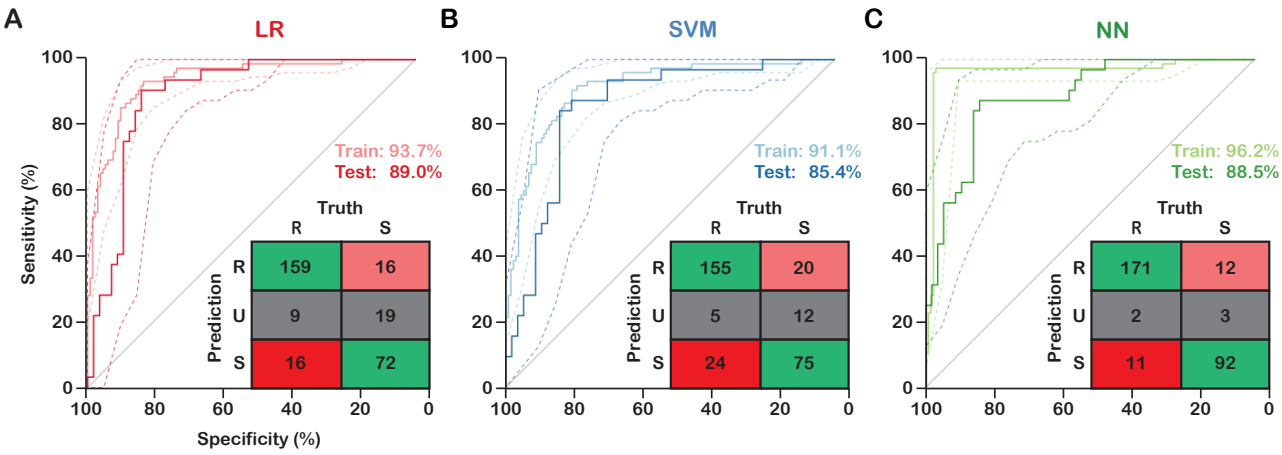

Figure 3: Machine learning models predict pyrazinamide resistance from structural features. Performance of (**A**) logistic regression (LR), (**B**) support vector machine with radial kernel (SVM) and (**C**) neural network (NN) models for prediction of pyrazinamide resistance. Dotted lines represent 95% confidence intervals from bootstrapping (n=10,000) and the area under the curve is reported for training and testing sets. Truth tables are shown for the combined training and test sets.

To evaluate the different models, we randomly divided our derivation dataset mutations (184 resistant, 107 susceptible) into a 70% training set and a 30% testing set, preserving the overall distribution of resistant and susceptible mutations. Models were then trained using repeated 10-fold cross validation (**Methods**). Since the models output a probability of resistance between 0 and 1, we defined three categories; resistant (R, $p<0.4$), susceptible (S, $p>0.6$) and uncertain (U, $0.4 \leq p \leq 0.6$) (**Figure S2A**). A model call of uncertain (U) was considered a minor error for the purposes of comparing to binary (R/S) phenotypic data. The models were able to call ~87-99% (183-190) of the mutations in the training set using these thresholds. As expected,

drops in performance (as measured by area under the ROC curve) were observed for all models when applied to the independent testing set, however the 95% bootstrap confidence intervals for training and testing overlapped for all three methods (**Figure 3, Table S3**). The neural network (NN) model had the highest diagnostic odds ratio (119), followed by logistic regression (LR, 45) and then the support vector machine (SVM, 24, **Figure 3**). As the best performing model, the predictions from the neural network model were used for all further analyses.

Analysis of model errors on the derivation set

Clinical diagnostic errors for antibiotic resistance are categorized into three classes: very major errors, which represent truly resistant isolates that are called susceptible, major errors (true susceptible cases called resistant), and minor errors, which are not called by the method being tested but are determined as resistant or susceptible by the reference method[59]. Collectively, the three models made 107 incorrect calls (28 very major errors, 29 major errors, and 50 minor errors); however, only 14 of these were shared between all three models (8 very major errors and 6 major errors, **Figure 4A**, **Table S4**). The best performing model (neural network) had a sensitivity of 93% (89-97%), a specificity of 86% (79-93%), and a positive predictive value of 94% (91-97%) with minor errors considered equivalent to major and very major errors where appropriate (**Table S3**).

Although the mutations responsible for the very major errors (predicted susceptible, phenotypically resistant) of the neural network model were dispersed throughout the protein structure, most (10/11) were surface exposed (**Figure 4B**). All these mutations were predicted to either not affect or slightly increase the stability of PncA, suggesting these errors may be due to inaccuracies in the predicted free energy change of unfolding (**Figure S2**). Major errors (predicted resistant, phenotypically susceptible) were typically driven by overestimation of a mutation's potential to effect PncA structure or function. Mapping the mutations responsible for

these major errors onto the sequence and structure revealed one cluster near the active site and the coiled turn between the α–1 and α–2 helix (residues 15, 21, 23, 131, and 133) (**Figure 4B, C**). Major errors appear to be caused in part by a combination of overestimation of the MAPP score at invariant positions that are near the active site or buried in the protein core.

Interestingly, several major errors occurred in a region of the active site termed the "oxyanion hole" (residues 131-138) which coordinates the carbonyl group of pyrazinamide in the active site[50]. The effects of mutations that lie in this region could be over-estimated due to their proximity to the active site and relatively lower solvent exposure. As the interaction between the oxyanion hole and pyrazinamide is mediated by the peptide backbone, and is therefore sidechain-independent, there is likely to be less stringent selection of the residues at these positions as long as the overall peptide backbone structure is maintained. Interestingly however, Gly132 and Ala134 are invariant in the alignment used to generate the MAPP score, which would suggest that these sites are under strong selective pressures. It has been shown previously that specific residues are favored in the positions surrounding *cis* peptide bonds, so

future work could attempt to model the mutations occurring in this functional region more specifically[60].

Minor errors are cases where the model could not confidently call a mutation resistant or susceptible and represented 34% (47/104) of errors collectively made by all three models. The model features for mutations that were called U tended to be intermediate compared to those of resistant and susceptible isolates, which raised the intriguing possibility that these are mutations with an intermediate effect on protein stability and/or enzyme activity (**Figure S2**).

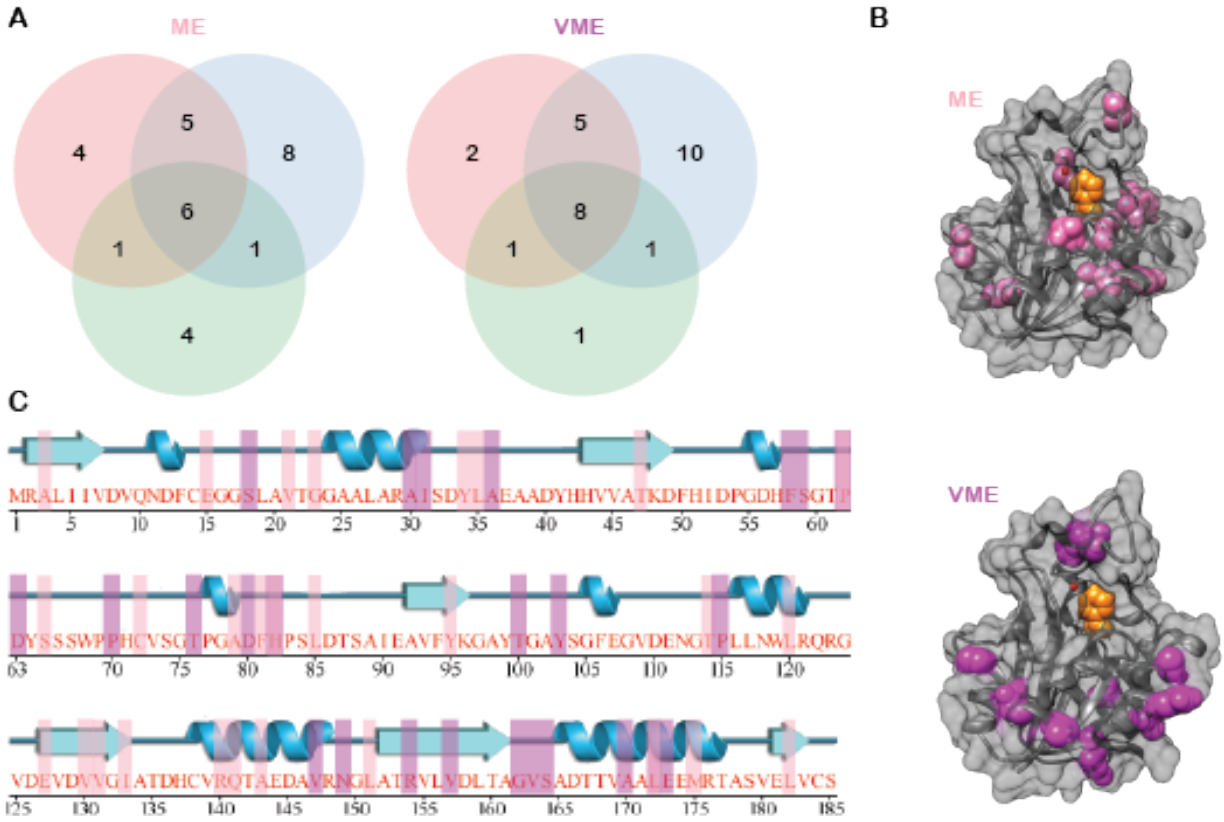

Figure 4. Very major errors are concentrated on the surface of PncA. (**A**) Major (ME) and very major (VME) errors shared between the models. Errors from logistic regression are shown in red, support vector machines in blue, and neural network in green (**B**) PncA with major (pink) and very major (magenta) errors shown as spheres. (**C**) Major (pink) and very major (magenta) errors mapped onto the PncA primary sequence.

As the neural network model does not clearly indicate which PncA features drive its predictions, we used logistic regression with backwards elimination (**Methods**) on the training dataset to gain further insight into the complex interplay between these factors. This analysis revealed that solvent accessibility, distance to the active site, the evolutionary conservation of the wild-type amino acid, the number of hydrogen bonds formed by the wild-type residue, significant changes in protein stability (measured as a change in free energy of protein unfolding of >2 kcal/mol), and the MAPP score were all independent explanatory factors. In addition, interactions between both the MAPP score and number of hydrogen bonds were found to moderate the effect of the distance to the active site. A higher MAPP score increased the deleterious effect of a mutation near the active site, while the importance of the number of hydrogen bonds a residue was involved in decreased with greater distance from the active site. This may be due to the requirement of hydrogen bonding interactions for proper active site geometry. There was weak evidence for two other interactions (between protein destabilization and either the number of hydrogen bonds or solvent accessibility, $p=0.073$ and $p=0.054$ respectively). These results suggest that interactions between model features are important for the prediction of resistance, which may be why the neural network model outperforms logistic regression in classifying mutations.

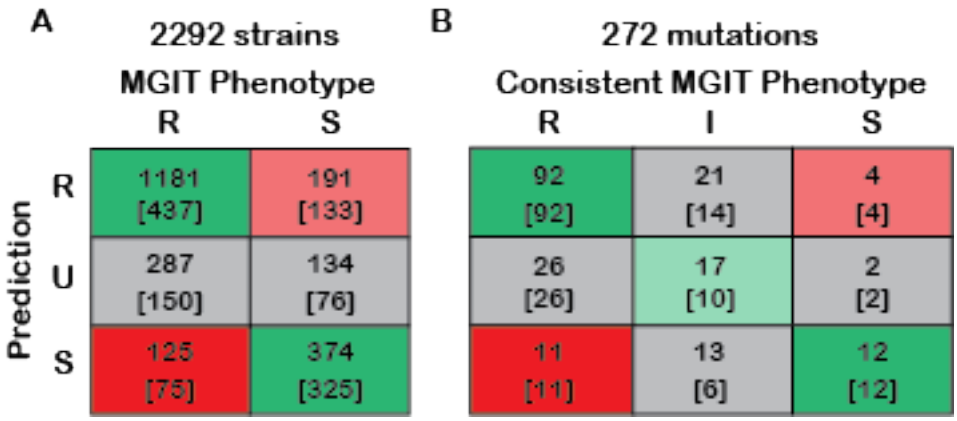

Figure 5. Model predictions based on mutations generalize to MGIT phenotypes. (**A**) Truth table of the predictions' performance on a dataset of *M. tuberculosis* strains tested by MGIT. Brackets denote predictions based on missense mutations not in the training set. (**B**) Truth table of model predictions versus average mutation phenotypes. "I" is defined as mutations that are not R or S in >75% of isolates tested (n≥4). Brackets denote phenotypes for mutations for which there were enough MGIT results to be confident in the assigned phenotype (177/272, 65%, **Methods**).

Neural network predictions generalize to a large clinical dataset

To assess the generalizability of our best model, the neural network, we applied its predictions to a clinical dataset of 2,292 *pncA* gene sequences with MGIT antibiotic susceptibility results (**Table 1**), each representing a unique isolate collated from published studies of clinical isolates[23,32,41,42]. In addition to the clinical isolates that were used to build the derivation dataset, this dataset also included 500 isolates with mutations that were only observed once or isolates with mutations whose phenotype appeared to vary between isolates tested. Thus, this dataset includes noisier phenotypic data that encapsulates both the uncertain phenotypes of some mutations and the real-world variability of culture-based phenotypic methods for pyrazinamide susceptibility testing. As the models were trained on high-confidence phenotypes derived from pooled MGIT phenotypes for each mutation, the 204 mutations used in the training set represented 1,096 isolates (48%) of the clinical dataset, while 1,196 isolates (52% of the clinical dataset) harbored mutations that were not used in model training. Predicting

resistance/susceptibility based on the mutation present, the model correctly predicted 74.1% (1,181/1,593) of MGIT-resistant isolates with *pncA* mutations; however, it performed more poorly for MGIT-susceptible isolates, correctly predicting only 53.5% (374/699) of strains (**Figure 5A**). 68.1% (287/421) of U calls made by the model were associated with resistance, which suggests that clinically a U call could be interpreted as possible pyrazinamide resistance contingent upon further testing (**Figure 5A**).

Intriguingly, 30 mutations (present in 385 strains, 16.8% of strains classified) had variable MGIT results (defined as having a resistant MGIT phenotype in 25-75% of cases with at least 4 observations). To understand our model performance with these possible “intermediate” mutations, we compared the model predictions with the average phenotypes of the 272 unique missense mutations found in clinical dataset, which we classified as resistant (MGIT resistant in >75% of isolates with this mutation tested), susceptible (MGIT susceptible in >75% of isolates with this mutation tested) and intermediate (I), the remainder. The model classified 14 (47%) of these high-confidence “intermediate” mutations as resistant, leading to 76 strains (40% of major errors made by the model) with susceptible MGIT phenotypes being misclassified as resistant (**Figure 5B**). While the model predicted U for 10 (33%) of the intermediate mutations, there was no clear relationship between U calls and intermediate mutations, which is consistent with the fact that the model was trained on binary data (**Figure 5B**). Overall, the model more accurately (as assessed by the diagnostic odds ratio) predicted the average phenotypes of mutations than the individual MGIT phenotypes of clinical isolates. The model still made 11 very major errors on the average phenotypes; however, some of these errors may be due to errors in MGIT antibiotic susceptibility testing. Alternatively, these errors could result from resistance that is determined by factors upstream of protein folding and function and is therefore outside the scope of our model. These results collectively suggest that the model predictions could be used to confirm MGIT susceptibility test results, with a discordant result suggesting that culture-based testing should be repeated.

Comparison of model predictions with pyrazinamide minimum inhibitory concentrations in vitro

Since it is difficult to assess how much of the discordance in the previous section can be attributed to either error in the measured clinical phenotype or deficiencies in our model, we compared its predictions to minimum inhibitory concentration (MIC) data taken from a small but high-quality dataset of 71 *M. tuberculosis* isolates (59 unique missense mutations, quantitative dataset, **Table 1**, **Methods**). This also enabled us to test the model's capacity to predict the *degree* of pyrazinamide resistance conferred by a particular mutation, by comparing the calls and predicted probabilities of our model with the pyrazinamide MICs. Overall, our model correctly predicted the binary (R/S) phenotype for 56 of 71 isolates with single missense mutations in PncA, and, crucially, predicted the correct phenotype for 11 out of 14 mutations (18 isolates in total) that were not in either the derivation or clinical datasets (**Table S6**). One of these errors is likely due to a MGIT testing error, as the same mutation (Gly97Asp) is observed in another resistant isolate and has been classified as conferring resistance in other studies (**Table S1**). Interestingly, of the seven isolates with mutations previously cataloged as susceptible (**Table S1**), four had MICs that were above the R/S cutoff of 100 μg/mL, highlighting the variability in MIC determination for pyrazinamide using MGIT. Out of the 18 isolates with mutations that lacked previously classified R/S phenotypes, 10 harbored mutations with disputed effects, and 8 harbored mutations that had previously lacked sufficient evidence to be classified. In addition, several samples with large deletions in *pncA* were also observed; these strains had extremely high (>900 μg/mL) MICs, which is consistent with the loss of functional PncA protein (**Table S6**). These data, taken together with the results for isolates harboring insertions/deletions in the study by the CRyPTIC Consortium *et al*, confirm it could be reasonable to assume that nearly all large insertion/deletion mutations and frameshifts in *pncA* confer resistance to pyrazinamide[23].

Predicting the effect of all possible non-synonymous pncA mutations on pyrazinamide susceptibility

Since trained machine learning models require very little computational resource, we applied our model to the set of missense mutations resulting from every possible non-synonymous single nucleotide polymorphism (SNP) in *pncA* (coding for 1,105 unique amino acid changes, 814 previously unclassified missense mutations), thereby estimating the probability that each mutation confers pyrazinamide resistance (**Table S5**). Overall, 22% (244) of missense mutations were predicted to confer resistance, while 63% (691) were predicted to have no effect on the action of pyrazinamide and the remaining 14% (158) were predicted to have an uncertain effect. Interestingly, the proportion of predicted resistant mutations from all possible non-synonymous SNPs was much lower than that in the derivation set (22% versus 63% respectively). This may be due to an increased likelihood of sequencing pyrazinamide-resistant clinical isolates, leading to an over-representation of resistance-conferring mutations in our derivation dataset as opposed to susceptible ones. This estimate is also more consistent with the proportion of resistant mutations identified in the Yadon *et al* screen (31%)[42]. As more unselected studies of whole genome sequencing are conducted, we expect this bias to unwind and consequently more susceptible mutations will be found for most established drugs. Alternatively, it could be caused by a global underestimation of resistance by our model, which underpredicted resistance by ~10% in the clinical dataset. Finally, this difference could represent the fact that phenotypically intermediate mutations classified as U by our model are classified as resistant in the catalogs/screens used to develop the dataset.

To try and get an understanding of how these predictions could improve our capacity to identify resistant mutations in *pncA*, we queried a bacterial index of the European Nucleotide Archive (ENA) to identify all single nucleotide polymorphisms (SNPs) coding for missense mutations in the *M. tuberculosis pncA* coding sequence (**Table S5**)[43]. We found 37,560 sequences classified as *M. tuberculosis*, with 4,102 strains harboring single missense mutations

in PncA (**Methods**). We supplemented this ENA dataset with the *pncA* sequences collected by CRyPTIC *et al*[23] as these sequences were not deposited when the index of the ENA was built, bringing the total number of strains with single missense mutations to 6,193 (47,769 total sequences, prevalence dataset, **Table 1**). We are using this dataset as the largest available sample of the genetic diversity in PncA existing in clinical infections; this is almost certainly biased to some degree due to oversampling of outbreak strains and other factors, however, until very large unselected clinical datasets are collected, it is the best dataset available.

Out of the 480 unique missense mutations observed in the prevalence dataset, 237 were observed in at least two lineages. After applying our predictions, we found that homoplastic (observed in at least two lineages) missense mutations in PncA were associated with resistance (odds ratio = 4.4, $p<0.0001$). 327 (68%) mutations were observed 10 or fewer times and 111 (23%) only once, highlighting again the need for approaches capable of predicting the effect of rare missense mutations. We classified the prevalence dataset using a published heuristic catalog[45], supplemented with our resistant and susceptible model predictions, to quantify how much our machine learning model improves our capacity to screen for potential pyrazinamide resistance using whole genome sequencing. While the heuristic catalog alone was able to classify 4,022 strains (73%, 291 mutations), our model classified an additional 983 strains (17%, 47 mutations), allowing us to provisionally classify 90% of the strains with missense mutations in the prevalence dataset.

Discussion

De novo prediction of 814 *pncA* mutations' effects on pyrazinamide resistance constitutes a significant step forward in our ability to predict pyrazinamide resistance from genetics and a proof of concept for using structural approaches to infer the effects of *pncA* SNPs on pyrazinamide resistance. While improvements to the model are necessary to achieve the sensitivity and specificity required for routine clinical use, this work increases our ability to classify rare resistance mutations, thereby potentially increasing the capability of whole genome sequencing based diagnostic susceptibility testing to respond to emerging and rare resistance patterns, as well as prioritizing rare resistance mutations for *in vitro* validation. Additionally, improving the classification of susceptible *pncA* mutations will allow us to begin to disentangle the involvement of other genes in pyrazinamide resistance, including determining the effect of mutations in other pyrazinamide resistance-associated genes such as *panD* and *rpsA*.

There are two principal limitations of this approach: (1) since the training set uses a binary resistant/susceptible phenotype, the models can only predict whether a mutation confers high-level resistance (>100 μg/mL) or not and (2) it can only make predictions for missense mutations in the coding sequence of *pncA*. It is known that genetic variation can lead to small changes in MIC for pyrazinamide and other first-line antitubercular compounds and that, whilst these may not change the binary phenotype, they do affect clinical outcome[49,61]. Models that predict the change in MIC due to genetic variation would be very helpful and interesting, however, until very large datasets of pyrazinamide MICs and their associated genomes become available, we must be content with predicting binary phenotypes. In addition, while we have shown that missense mutations represent most (60.5%) of the possible resistant genetic variants in *pncA*, insertions/deletions and nonsense mutations (7.9%) must also be considered, as they are generally associated with resistance. Likewise, promoter mutations that result in reduced transcription of *pncA* will likely also lead to resistance. While no synonymous mutation has yet been observed to cause clinical pyrazinamide resistance, the possibility remains that a

synonymous mutation could have an effect on RNA stability, ribosomal stalling, or codon usage and confer resistance. Indeed, Yadon *et al* do pick up 18 sSNPs that are enriched both *in vivo* and *in vitro* in their pyrazinamide resistance screens[42]. The model also does not take into account the introduction of protease cleavage sites or other processing abnormalities. Finally, while most pyrazinamide resistance is caused by mutations in *pncA*, recent studies have also implicated other genes, notably *rpsA*, *panD*, and the putative efflux pumps *Rv0191*, *Rv3756c*, *Rv3008*, and *Rv1667c* in pyrazinamide resistance[4,25-31]. Further research is needed to determine if mutations in these genes can be reliably inferred to confer pyrazinamide resistance.

Several predictive features used in the model could be improved upon in future work. The MAPP score relies on the maintenance of function between diverse homologs to determine the evolutionary constraints on each position in a protein. While we selected sequences that contained the residues involved in active site formation and metal binding, we did not experimentally confirm pyrazinamide conversion by each homolog. Additionally, the *in silico* method that we employed to estimate each mutation's effect on protein stability could be improved by comparison and calibration with *in vitro* biochemical data. Finally, as the active site of PncA is formed in part by a *cis* peptide bond between Ile133 and Ala134, more detailed modeling of the evolutionary constraints at this site could more accurately assess the functional impact of a mutation at these positions. Despite the fact that most features we investigated were associated with pyrazinamide resistance, not all were retained as independent predictors in our final logistic regression model. The change in hydropathy and sidechain volume as well as the Rogov score are all likely encapsulated by the MAPP score, as this takes into account the wild-type and mutant amino acids in its calculation.

The accuracy of model predictions based on structural features suggests that the underlying hypothesis of predicting pyrazinamide resistance based on predicted PncA function is valid. Mapping the potential of each position to harbor a predicted or bona fide resistance mutation onto the structure reveals that many resistance-prone positions are associated with the

active site or metal binding, as noted previously (**Figure 1C**). Interestingly, however, most of the other resistance-prone positions are involved in packing interactions between secondary structure elements in PncA, supporting the assertion that a major mechanism of pyrazinamide resistance is loss of protein stability. All susceptibility-prone positions are highly solvent-exposed and many are on flexible loops, consistent with our expectation that these regions experience lower selective pressures and have a lower/negligible effect on PncA stability and function. The effect of perturbing the protein core appears to be more dependent on the specific amino acid chemistries involved, as many codons harbor nearly equal numbers of resistant and susceptible mutations, which is consistent with the ability of a conservative, hydrophobic mutation to be tolerated in a region that relies on non-specific, volume-mediated packing driven by the hydrophobic effect.

One major question that remains is whether the mutations not called by the model (U) represent inaccuracies in the calculation of model features, breakdowns in the model, or have an intermediate effect on protein stability and/or enzyme activity. Most of the un-called mutations have intermediate features that lie in between the resistant and susceptible distributions (**Figure S2**). The MAPP score has been shown to be capable of delineating mutations that have mild and severely deleterious effects in other genes, suggesting that mutations with intermediate MAPP scores may indeed be intermediate in effect[58]. In addition, some of the mutations that are called U appear to not be reproducible when experimentally tested using the chosen gold-standard culture-based method, MGIT, supporting the possibility of an intermediate class (**Figure 5B**). One previous study has shown associations between reductions in PncA stability/function *in vitro* and outcomes in infected mice, but more work is necessary to fully understand whether this relationship extends to clinical outcomes in patients[42]. While mutations have historically been classified using a binary system, this study supports the view of mutations as conferring a spectrum of resistance. This is evidenced by the variable effects on MIC for some of the mutations investigated in this study, three of which

range from susceptible to above the MGIT breakpoint for pyrazinamide (100 μg/mL, **Table S6**). Future approaches could examine either probabilistic modelling or multi-class classification to attempt to encapsulate the uncertainty in phenotype associated with certain *pncA* mutations. Additionally, mutations could also be weighted in training by their relative prevalence in the ENA, which would prioritize accuracy for abundant mutations. However, this may result in bias toward highly sequenced variants that are a part of outbreak strains.

Predictions made by this model could potentially provide clinicians with an initial estimate of pyrazinamide susceptibility after a novel mutation is observed but before traditional phenotypic testing has been completed. Given the latter can take weeks or even months, this could help guide initial therapy and further antibiotic susceptibility testing. In addition, the putative classification of additional *pncA* mutations potentially enables genetic variants conferring pyrazinamide resistance that do not involve the *pncA* gene to be discovered. The identification of pyrazinamide-susceptible mutations is also crucial, as it has been suggested that any non-synonymous mutation in *pncA* that is not cataloged as susceptible confers resistance, an incorrect assumption that would lead to overprediction of pyrazinamide resistance[62].

This study constitutes a proof-of-concept for the computational prediction of pyrazinamide resistance, a critically important drug in the treatment of tuberculosis. However, this approach is not limited to *pncA* but should in theory be extensible to any pro-drug system where the converting enzyme is non-essential, such as delaminid, protaminid, or ethionamide as well as to pro-drug systems outside of *M. tuberculosis*. Interestingly, a recent study has highlighted similar trends in the features used in this study for resistance-conferring mutations in *katG* (isoniazid), *rpoB* (rifampicin), and *alr* (D-cycloserine), suggesting that this approach may even be applicable to non-prodrug systems[63]. One promising area for future work is in the tuberculosis drug bedaquiline, where resistance is caused in part by mutations in a transcriptional repressor (*Rv0678*) that cause loss of DNA binding and upregulation of efflux

pumps[64,65]. *Rv0678* has shown a high degree of mutational promiscuity in published sequencing studies, which would highlight the value of a computational approach[66–70]. The ability of this approach to identify the major mechanisms of resistance to pyrazinamide highlights the need for continued basic research to determine the structures of other bacterial proteins implicated in antibiotic resistance. Additionally, the efficacy of this approach highlights the value of including evolutionary constraints for prediction of mutational effects. Further understanding of the effect of *pncA* mutations also increases the ability of whole genome sequencing approaches to move to the forefront of global tuberculosis control efforts.

Acknowledgements

The study was funded by the National Institute for Health Research Health Protection Research Unit (NIHR HPRU) in Healthcare Associated Infections and Antimicrobial Resistance at Oxford University in partnership with Public Health England (PHE) [HPRU-2012-10041]; the National Institute for Health Research (NIHR) Oxford Biomedical Research Centre (BRC); the CRyPTIC consortium, which is funded by a Wellcome Trust/Newton Fund-MRC Collaborative Award [200205/Z/15/Z] and the Bill and Melinda Gates Foundation Trust [OPP1133541]; and the South African Medical Research Council. The EXIT-RIF project (Prof Annelies Van; Prof Rob Warren, Prof Lesley Scott, Prof Wendy Stevens, Dr Michael Whitfield) is funded by National Institutes of Health grant [#R01 AI099026]. J.J.C. is supported by a fellowship from the Rhodes Trust. T.E.A.P. and D.W.C. are NIHR Senior Investigators. T.M.W. is an NIHR Academic Clinical Lecturer. J.J.C would like to thank Spencer Dunleavy for thoughtful discussions on statistical analysis and modeling. The content is the solely the responsibility of the authors and does not necessarily represent the official views of the South African Medical Research Council. MGW is supported by TORCH funding through the Flemish Fund for Scientific Research (FWO G0F8316N) and a fellowship from the Claude Leon Foundation. The findings and conclusions in this report are solely the responsibility of the authors and do not necessarily represent the official views of the NHS, the NIHR, the Department of Health, the Centers for Disease Control and Prevention (CDC), or the US Department of Health and Human Services. References in this manuscript to any specific commercial products, process, service, manufacturer, or company do not constitute its endorsement or recommendation by the US government or CDC.

Author Contributions

JJC, PWF, and TMW designed experiments, JJC carried out the experiments, JJC and ASW performed statistical analyses, JJC, PWF, ASW, and TEAP wrote the manuscript. PWF and DWC supervised the work. MGW, JEP and GPM contributed samples and edited the manuscript.

Declaration of Interests

The authors have no interests to declare.

Methods

Mutation phenotype thresholds

Mutations included in the derivation (training and testing) dataset were selected from four large studies/reviews that included phenotypic diagnostic susceptibility testing of clinical isolates for pyrazinamide and one *in vitro*/*in vivo* phenotypic screening study[23,32,41,42,45]. Briefly, phenotypes for strains with single missense mutations in *pncA* from the four studies of clinical isolates were aggregated by mutation, tallying the results of the phenotypic testing. Mutations that were resistant or susceptible at least 75% of the time and that had been phenotyped at least 4 times were included as derivation phenotypes. Additionally, mutations that had been phenotyped at least twice with no discrepancies were also included. These mutations were then cross-referenced and supplemented with mutations from Yadon *et al.* that were either enriched (resistant) or depleted (susceptible) in both the *in vitro* and *in vivo* screens performed in that study[42]. Mutations that had conflicting clinical and laboratory phenotypes (n=2) were removed from the derivation dataset. Mutations that were only present in either clinical isolates or *in vitro* isolates but that met the criteria for inclusion from that set were included. This led to a final total of 291 mutations with high-confidence phenotypes of which 184 were resistant and 107 were susceptible to pyrazinamide.

In silico structural measurements

The change in mass, volume, charge, hydrophobicity, distance from the Fe^{2+} ion and pyrazinamide molecule, solvent accessibility, MAPP score, Rogov score, degree of hydrogen bonding, and predicted change in the free energy of protein unfolding were determined for each mutation. Hydrophobicity was estimated using the Kyte-Doolittle scale. Distances were calculated as the minimum distance between each residue and the Fe^{2+} ion or pyrazinamide molecule using UCSF Chimera. Solvent accessibility and predicted number of hydrogen bonds

were calculated in UCSF Chimera. *In silico* calculation of the change in free energy of protein unfolding was calculated using a meta-predictor as described in Broom *et al*[56]. The MAPP score was calculated using software available at (http://mendel.stanford.edu/SidowLab/downloads/MAPP/index.html) using related orthologs.

Statistical modelling procedures

Logistic regression, support vector machines with radial kernels, and multi-layer perceptron neural networks were implemented using the R caret package. Briefly, 70% of the derivation dataset was randomly selected (maintaining the ratio of resistant to susceptible phenotypes) as a training set and 30% was reserved as a test set. All three model types were trained using repeated (n=10) 10-fold cross validation with class weights to compensate for the class imbalance. Model performance was evaluated by the area under the receiver-operating characteristic curve. After model selection for each method, generalizability was determined using an independent test set. The final model was selected by calculating the diagnostic odds ratio for each method over the entire derivation set. In order to select the variables used for logistic regression, backwards stepwise elimination (exit p=0.15) was performed on the entire derivation dataset to select the main effects and then interactions between the significant terms were manually investigated, retaining any with heterogeneity $p<0.05$. Two additional weak interactions (between the protein destabilization factor and either number of hydrogen bonds ($p=0.073$) or solvent accessibility ($p=0.054$)) were not included in the final model. The final logistic regression model was trained using the identified main effects with the two significant interactions on the training set using repeated (n=10) 10-fold cross validation to select hyperparameters before being applied to the test set. The training set along with all relevant R code is publicly-available (https://github.com/carterjosh/PZA-machine-learner). This allows the reader to either load the three machine-learning models described, or to repeat the training process thereby creating their own models.

Prevalence Dataset Construction

The European Nucleotide Archive was queried using pygsi and the H37Rv reference sequence for *pncA* with 10 nucleotides padded on either side[43,44]. This allowed for sequence search of all possible single nucleotide mutations in *pncA* that have been deposited in the ENA. In total, 38,440 *pncA* sequences classified as originating from *Mycobacterium tuberculosis* were extracted. These were combined with *pncA* sequences determined by the CRyPTIC Consortium *et al* and filtered for single missense mutations to form the final prevalence dataset[23].

Pyrazinamide minimum inhibitory concentration determination

Isolates used for MIC determination came from two sources, the EXIT-RIF study and US Centers for Disease Control. All EXIT-RIF isolates were collected in South Africa. Of the 366 *Mycobacterium tuberculosis* clinical isolates, 333 were collected as part of a prospective cohort study ("EXIT-RIF") aimed at comparing the outcome of patients diagnosed with rifampicin resistant tuberculosis by MTBDR*plus* (Hain LifeSciences) or Xpert MTB/RIF between November 2012 and December 2013 in three South African provinces (Free State, Eastern Cape and Gauteng). A *Mycobacterium tuberculosis* databank housed at the SAMRC Centre for Tuberculosis Research, consisting of ~45,000 drug resistant isolates collected in the Western Cape province since 2001, was queried to identify isolates containing both PZA MIC data and *pncA* genotypic data, this produced the remaining 33 *Mycobacterium tuberculosis* clinical isolates. Isolates that harbored single amino acid substitutions in PncA (39 out of 366 total) were selected for comparison to model predictions. An additional 32 clinical isolates (collected from 2000 to 2008) harboring single missense mutations in *pncA* came from the culture collection at the Laboratory Branch, Division of Tuberculosis Elimination, US CDC.

All MICs were determined using the non-radiometric BACTEC MGIT 960 method (BD Diagnostic Systems, NJ, USA) with manufactured supplied pyrazinamide medium/supplement

as previously described[71]. This system makes use of modified test media which supports the growth of mycobacteria at a pH of 5.9. MICs determined for isolates from the EXIT-RIF study were tested at 900, 600 and 300 µg/ml for the large deletion isolates and 100, 75, 50, 25 µg/ml for the rest. MICs determined by the Center for Disease Control were determined using PZA concentrations of 50, 100, 200, 300, 400, 600 and 800 µg/ml. A fully susceptible MTB laboratory strain H37Rv (ATCC 27294) was included as a control for all isolates tested.

References

1. The World Health Organization. *Global Tuberculosis Report 2017*. (2017).
2. The World Health Organization. *Treatment of tuberculosis: guidelines*. (2010). doi:10.1164/rccm.201012-1949OC
3. Njire, M. *et al.* Pyrazinamide resistance in Mycobacterium tuberculosis: Review and update. *Advances in Medical Sciences* **61**, 63–71 (2016).
4. Zhang, Y. & Yew, W. W. Mechanisms of drug resistance in Mycobacterium tuberculosis: Update 2015. *Int. J. Tuberc. Lung Dis.* **19**, 1276–1289 (2015).
5. Zhang, Y. & Mitchison, D. The curious characteristics of pyrazinamide: A review. *Int. J. Tuberc. Lung Dis.* **7**, 6–21 (2003).
6. Mitchison, D. A. The action of antituberculosis drugs in short-course chemotherapy. *Tubercle* **66**, 219–225 (1985).
7. Fox, W., Ellard, G. A. & Mitchison, D. A. Studies on the treatment of tuberculosis undertaken by the British Medical Research Council Tuberculosis Units, 1946-1986, with relevant subsequent publications. *Int. J. Tuberc. Lung Dis.* **3**, S232–S279 (1999).
8. Dawson, R. *et al.* Efficiency and safety of the combination of moxifloxacin, pretomanid (PA-824), and pyrazinamide during the first 8 weeks of antituberculosis treatment: A phase 2b, open-label, partly randomised trial in patients with drug-susceptible or drug-resistant pul. *Lancet* **385**, 1738–1747 (2015).
9. Chang, K. C. *et al.* Pyrazinamide may improve fluoroquinolone-based treatment of multidrug-resistant tuberculosis. *Antimicrob. Agents Chemother.* **56**, 5465–5475 (2012).
10. Zumla, A. I. *et al.* New antituberculosis drugs, regimens, and adjunct therapies: Needs, advances, and future prospects. *Lancet Infect. Dis.* **14**, 327–340 (2014).
11. Nuermberger, E. *et al.* Powerful bactericidal and sterilizing activity of a regimen containing PA-824, moxifloxacin, and pyrazinamide in a murine model of tuberculosis. *Antimicrob. Agents Chemother.* **52**, 1522–1524 (2008).

12. Rosenthal, I. M. *et al.* Daily dosing of rifapentine cures tuberculosis in three months or less in the murine model. *PLoS Med.* **4**, 1931–1939 (2007).

13. Veziris, N. *et al.* A once-weekly R207910-containing regimen exceeds activity of the standard daily regimen in murine tuberculosis. *Am. J. Respir. Crit. Care Med.* **179**, 75–79 (2009).

14. Whitfield, M. G. *et al.* A global perspective on pyrazinamide resistance: Systematic review and meta-analysis. *PLoS One* **10**, 1–16 (2015).

15. Chang, K. C., Yew, W. W. & Zhang, Y. Pyrazinamide Susceptibility Testing in Mycobacterium tuberculosis: a Systematic Review with Meta-Analyses. *Antimicrob. Agents Chemother.* **55**, 4499–4505 (2011).

16. Jr., H. D., Horn, D. L. & Alfalla, C. Drug-Resistant Tuberculosis: Inconsistent Results of Pyrazinamide Susceptibility Testing. *JAMA J. Am. Med. Assoc.* **273**, 916–917 (1995).

17. Miller, M. A., Thibert, L., Desjardins, F., Siddiqi, S. H. & Dascal, A. Testing of susceptibility of Mycobacterium tuberculosis to pyrazinamide: Comparison of Bactec method with pyrazinamidase assay. *J. Clin. Microbiol.* **33**, 2468–2470 (1995).

18. Hoffner, S. *et al.* Proficiency of drug susceptibility testing of Mycobacterium tuberculosis against pyrazinamide: the Swedish experience. *Int J Tuberc Lung Dis* **17**, 1486–1490 (2013).

19. Pandey, S., Newton, S., Upton, A., Roberts, S. & Drinkovi, D. Characterisation of pncA mutations in clinical Mycobacterium tuberculosis isolates in New Zealand. *Pathology* **41**, 582–584 (2009).

20. Simons, S. O. *et al.* Validation of pncA gene sequencing in combination with the mycobacterial growth indicator tube method to test susceptibility of Mycobacterium tuberculosis to pyrazinamide. *J. Clin. Microbiol.* **50**, 428–434 (2012).

21. Chedore, P., Bertucci, L., Wolfe, J., Sharma, M. & Jamieson, F. Potential for erroneous results indicating resistance when using the bactec MGIT 960 system for testing

susceptibility of Mycobacterium tuberculosis to pyrazinamide. *J. Clin. Microbiol.* **48**, 300–301 (2010).

22. The World Health Organization. *Global Tuberculosis Report 2018*. (2018).
23. The CRyPTIC Consortium & The 100000 Genomes Project. Prediction of Susceptibility to First-Line Tuberculosis Drugs by DNA Sequencing. *N. Engl. J. Med.* **379**, 1403–1415 (2018).
24. Yüksel, P. & Tansel, Ö. Characterization of pncA mutations of pyrazinamide-resistant Mycobacterium tuberculosis in Turkey. *New Microbiol.* **32**, 153–158 (2009).
25. Ramirez-Busby, S. M. *et al.* A Multinational Analysis of Mutations and Heterogeneity in PZase, RpsA, and PanD Associated with Pyrazinamide Resistance in M/XDR Mycobacterium tuberculosis. *Sci. Rep.* **7**, 1–9 (2017).
26. Sheen, P. *et al.* A multiple genome analysis of Mycobacterium tuberculosis reveals specific novel genes and mutations associated with pyrazinamide resistance. *BMC Genomics* **18**, 1–11 (2017).
27. Zhang, S. *et al.* Mutation in clpC1 encoding an ATP-dependent ATPase involved in protein degradation is associated with pyrazinamide resistance in Mycobacterium tuberculosis. *Emerg. Microbes Infect.* **6**, e8-2 (2017).
28. Gopal, P. *et al.* Pyrazinamide resistance is caused by two distinct mechanisms: Prevention of coenzyme a depletion and loss of virulence factor synthesis. *ACS Infect. Dis.* **2**, 616–626 (2016).
29. Zhang, Y., Zhang, J., Cui, P., Zhang, Y. & Zhang, W. Identification of Novel Efflux Proteins Rv0191, Rv3756c, Rv3008, and Rv1667c Involved in Pyrazinamide Resistance in Mycobacterium tuberculosis. **61**, 1–10 (2017).
30. Hirano, K., Takahashi, M., Kazumi, Y., Fukasawa, Y. & Abe, C. Mutation in pncA is a major mechanism of pyrazinamide resistance in Mycobacterium tuberculosis. *Tuber. Lung Dis.* **78**, 117–122 (1997).

31. Stoffels, K., Mathys, V., Fauville-Dufaux, M., Wintjens, R. & Bifania, P. Systematic analysis of pyrazinamide-resistant spontaneous mutants and clinical isolates of Mycobacterium tuberculosis. *Antimicrob. Agents Chemother.* **56**, 5186–5193 (2012).

32. Paolo, M. *et al.* Mycobacterium tuberculosis pyrazinamide resistance determinants: a multicenter study. *MBio* **5**, e01819-14 (2014).

33. Scorpio, A. & Zhang, Y. Mutations in pncA, a gene encoding pyrazinamidase/nicotinamidase, cause resistance to the antituberculous drug pyrazinamide in tubercle bacillus. *Nat. Med.* **2**, 662–667 (1996).

34. Kim, N., Petingi, L. & Schlick, T. Network theory tools for RNA modeling. *WSEAS Trans. Math.* **12**, 941–955 (2013).

35. Yee, M., Gopal, P. & Dick, T. Missense mutations in the unfoldase ClpC1 of the caseinolytic protease complex are associated with pyrazinamide resistance in Mycobacterium tuberculosis. *Antimicrob. Agents Chemother.* **61**, 1–6 (2017).

36. World Health Organization. WHO Guideline: The use of molecular line probe assays for the detection of resistance to isoniazid and rifampicin. 21–23 (2016).

37. Driesen, M. *et al.* Evaluation of a novel line probe assay to detect resistance to pyrazinamide, a key drug used for tuberculosis treatment. *Clin. Microbiol. Infect.* **24**, 60–64 (2018).

38. Willby, M. J. *et al.* Detection of Mycobacterium tuberculosis pncA Mutations by the Nipro Genoscholar PZA-TB II Assay Compared to Conventional Sequencing. *Antimicrob Agents Chemother.* **62**, 1–9 (2018).

39. Ramirez-Busby, S. M. & Valafar, F. Systematic review of mutations in pyrazinamidase associated with pyrazinamide resistance in mycobacterium tuberculosis clinical isolates. *Antimicrob. Agents Chemother.* **59**, 5267–5277 (2015).

40. Kalokhe, A. S. *et al.* Multidrug-resistant tuberculosis drug susceptibility and molecular diagnostic testing: a review of the literature. *Am J Med Sci.* **345**, 143–148 (2013).

41. Whitfield, M. G. *et al.* Mycobacterium tuberculosis pncA polymorphisms that do not confer pyrazinamide resistance at a breakpoint concentration of 100 micrograms per milliliter in MGIT. *J. Clin. Microbiol.* **53**, 3633–3635 (2015).

42. Yadon, A. N. *et al.* A comprehensive characterization of PncA polymorphisms that confer resistance to pyrazinamide. *Nat. Commun.* **8**, 1–10 (2017).

43. Bradley, P., den Bakker, H. C., Rocha, E. P. C., McVean, G. & Iqbal, Z. Ultrafast search of all deposited bacterial and viral genomic data. *Nat. Biotechnol.* **37**, 152–159 (2019).

44. Fowler, P. W. pygsi: a Python class to interrogate BIGSI. (2018). doi:10.5281/zenodo.1407085

45. Walker, T. M. *et al.* Whole-genome sequencing for prediction of Mycobacterium tuberculosis drug susceptibility and resistance: A retrospective cohort study. *Lancet Infect. Dis.* **15**, 1193–1202 (2015).

46. Kirchner, S. *et al.* Alteration of protein function by a silent polymorphism linked to tRNA abundance. *PLoS Biol.* **15**, (2017).

47. Rauscher, R. & Ignatova, Z. Timing during translation matters: synonymous mutations in human pathologies influence protein folding and function. *Biochem. Soc. Trans.* **46**, 937–944 (2018).

48. Im, E. H., Hahn, Y. & Choi, S. S. Functional relevance of synonymous alleles reflected in allele rareness in the population. *Genomics* **110**, 347–354 (2018).

49. Gumbo, T. *et al.* The pyrazinamide susceptibility breakpoint above which combination therapy fails. *J. Antimicrob. Chemother.* **69**, 2420–2425 (2014).

50. Petrella, S. *et al.* Crystal structure of the pyrazinamidase of mycobacterium tuberculosis: Insights into natural and acquired resistance to pyrazinamide. *PLoS One* **6**, 1–8 (2011).

51. Faure, G. & Koonin, E. V. Universal distribution of mutational effects on protein stability, uncoupling of protein robustness from sequence evolution and distinct evolutionary modes of prokaryotic and eukaryotic proteins. *Phys. Biol.* **12**, 1–15 (2015).

52. Lim, W. A., Farruggio, D. C. & Sauer, R. T. Structural and Energetic Consequences of Disruptive Mutations in a Protein Core. *Biochemistry* **31**, 4324–4333 (1992).

53. Guo, H. H., Choe, J. & Loeb, L. A. Protein tolerance to random amino acid change. *Proc. Natl. Acad. Sci.* **101**, 9205–9210 (2004).

54. Chen, H. & Zhou, H. X. Prediction of solvent accessibility and sites of deleterious mutations from protein sequence. *Nucleic Acids Res.* **33**, 3193–3199 (2005).

55. Yoon, J. H., Nam, J. S., Kim, K. J. & Ro, Y. T. Characterization of pncA mutations in pyrazinamide-resistant Mycobacterium tuberculosis isolates from Korea and analysis of the correlation between the mutations and pyrazinamidase activity. *World J. Microbiol. Biotechnol.* **30**, 2821–2828 (2014).

56. Broom, A., Jacobi, Z., Trainor, K. & Meiering, E. M. Computational tools help improve protein stability but with a solubility tradeoff. *J. Biol. Chem.* **292**, 14349–14361 (2017).

57. Fowler, P. W. *et al.* Robust Prediction of Resistance to Trimethoprim in Staphylococcus aureus. *Cell Chem. Biol.* **25**, 339-349.e4 (2018).

58. Stone, E. A. & Sidow, A. Physicochemical constraint violation by missense substitutions mediates impairment of protein function and disease severity. *Genome Res.* **15**, 978–986 (2005).

59. U.S. Department of Health and Human Services Food and Drug Administration. Guidance for Industry and FDA. Class II Special Controls Guidance Document : Antimicrobial Susceptibility Test Systems. 1–42 (2009).

60. Pal, D. & Chakrabarti, P. Cis peptide bonds in proteins: residues involved, their conformations,\interactions and locations. *J. Mol. Biol.* **294**, 271–288 (1999).

61. Colangeli, R. *et al.* Bacterial factors that predict relapse after tuberculosis therapy. *N. Engl. J. Med.* **379**, 823–833 (2018).

62. Zignol, M. *et al.* Population-based resistance of Mycobacterium tuberculosis isolates to pyrazinamide and fluoroquinolones: results from a multicountry surveillance project.

Lancet Infect. Dis. **16**, 1185–1192 (2016).

63. Portelli, S., Phelan, J. E., Ascher, D. B., Clark, T. G. & Furnham, N. Understanding molecular consequences of putative drug resistant mutations in Mycobacterium tuberculosis. *Sci. Rep.* 1–12 (2018). doi:10.1038/s41598-018-33370-6

64. Milano, A. *et al.* Azole resistance in Mycobacterium tuberculosis is mediated by the MmpS5-MmpL5 efflux system. *Tuberculosis* **89**, 84–90 (2009).

65. Nguyen, T. V. A., Anthony, R. M., Bañuls, A. L., Vu, D. H. & Alffenaar, J. W. C. Bedaquiline Resistance: Its Emergence, Mechanism, and Prevention. *Clin. Infect. Dis.* **66**, 1625–1630 (2018).

66. Zhang, S. *et al.* Identification of novel mutations associated with clofazimine resistance in Mycobacterium tuberculosis. *J. Antimicrob. Chemother.* **70**, 2507–2510 (2015).

67. Villellas, C. *et al.* Unexpected high prevalence of resistance-associated Rv0678 variants in MDR-TB patients without documented prior use of clofazimine or bedaquiline. *J. Antimicrob. Chemother.* **72**, 684–690 (2017).

68. Xu, J. *et al.* Primary clofazimine and bedaquiline resistance among isolates from patients with multidrug-resistant tuberculosis. *Antimicrob. Agents Chemother.* **61**, 1–8 (2017).

69. Hartkoorn, R. C., Uplekar, S. & Cole, S. T. Cross-resistance between clofazimine and bedaquiline through upregulation of mmpl5 in mycobacterium tuberculosis. *Antimicrob. Agents Chemother.* **58**, 2979–2981 (2014).

70. Somoskovi, A., Bruderer, V., Hömke, R., Bloemberg, G. V. & Böttger, E. C. A mutation associated with clofazimine and bedaquiline cross-resistance in MDR-TB following bedaquiline treatment. *Eur. Respir. J.* **45**, 554–557 (2015).

71. Piersimoni, C. *et al.* Prevention of false resistance results obtained in testing the susceptibility of Mycobacterium tuberculosis to pyrazinamide with the bactec MGIT 960 system using a reduced inoculum. *J. Clin. Microbiol.* **51**, 291–294 (2013).