eLife

Reviewed Preprint
v1 • December 12, 2025
Not revised

Reviewed Preprint
v2 • September 1, 2026
Revised by authors

The structural context of mutations in proteins predicts their effect on antibiotic resistance

Anna G Green[1,2], **Mahbuba Tasmin**[2], **Roger Vargas Jr**[1], **Maha R Farhat**[1,3]

[1]Department of Biomedical Informatics, Harvard Medical School, Boston, United States • [2]Manning College of Information and Computer Sciences, University of Massachusetts, Amherst, United States • [3]Division of Pulmonary & Critical Care, Massachusetts General Hospital, Boston, United States

For correspondence: maha_farhat@hms.harvard.edu

Competing interests: No competing interests declared

Funding: See page 16

Reviewing editor: Bryan D. Bryson, Massachusetts Institute of Technology, United States

eLife Assessment

This **important** study leverages a large global dataset of tens of thousands of tuberculosis samples to place recurrent protein-coding mutations into their three-dimensional structural context, offering an expanded view of how antibiotic resistance emerges compared to traditional genetic analyses alone. The strength of evidence is **compelling**, supported by the scale and breadth of the dataset and the systematic structural analysis, although some of the assumptions made in the modeling approach are only partially supported. Overall, the work will be of broad interest to researchers studying microbial evolution, antibiotic resistance, and structure-function relationships in pathogens.

https://doi.org/10.7554/eLife.109450.2.sa2

Abstract

In *Mycobacterium tuberculosis*, a prevalent and deadly pathogen, resistance to antibiotics evolves primarily through non-synonymous mutations in proteins. Sequence-based analyses are currently used to understand the genetic basis of antibiotic resistance, either via genotype-phenotype association, or via signals of convergent evolution. These methods focus on primary sequence and often neglect other biological signals such as protein structural information. We hypothesize that integrating the structural context of mutations improves the prediction of effects on function and phenotype. We curate high confidence structural annotations for the *M. tuberculosis* proteome from 1,371 crystallography and 2,316 AlphaFold predictions, and combine the structures with mutations from over 31,000 clinical *M. tuberculosis* isolates. We demonstrate that mutations in proteins known to cause resistance are clustered in 3D space, even in proteins where inactivating mutations at any position are thought to cause resistance. We develop a statistic to search the *M. tuberculosis* proteome for signal of clustered mutations, finding over 450 proteins that display this signal, many of which have a known relationship with antibiotic resistance. We show that a supervised classifier trained on 3D distance to known resistance sites alone has an F1 score of 94.6% at classifying mutations as resistance-conferring across proteins. This work demonstrates that protein structure provides useful information for categorizing which variants may cause antibiotic resistance, even when the majority of structures are AI-predicted.

Introduction

The increasing prevalence of antibiotic-resistant *Mycobacterium tuberculosis* is challenging control of tuberculosis (TB), a disease responsible for the highest number of infectious disease related deaths globally.[1] Currently, diagnosis of antibiotic-resistant tuberculosis relies on time-consuming laboratory phenotype testing or the detection of resistance-conferring mutations in the *M. tuberculosis* genome through molecular assays or genetic sequencing.[2,3] Although the sensitivity

and specificity of these assays is high for several first and some second-line TB drugs, prediction is less accurate for other second-line antibiotics or novel agents like bedaquiline and pretomanid, which have recently become cornerstones of multi-drug resistant TB treatment. Improving the accuracy of resistance diagnosis relies on new computational approaches that can better link mutations with their functional impact on the resistance phenotype.

Two major computational strategies exist for identifying resistance-conferring variants: supervised approaches which associate genetic variation with resistance,[4] and unsupervised approaches that search for signal of evolutionary adaptation, which can indicate resistance development. Supervised statistical methods such as genome-wide association studies and random forest classifiers have identified mutations associated with antibiotic resistance in *M. tuberculosis*,[5–9] and machine learning approaches have built on this success to identify additional resistance-conferring variants.[10–16] But, these approaches require a large number of phenotyped isolates (both resistant and susceptible) to make accurate predictions, which presents a major limitation, especially for recently evolved mutations. Evolution-based approaches are an alternative for finding variants associated with antibiotic resistance by analyzing their mutational frequency and phylogenetic distribution: by searching for convergent positive selection in *M. tuberculosis*, studies have found mutations in the *ald* gene associated with D-cycloserine resistance[17], phase variation associated with virulence,[18] and mutations involved in host-pathogen interactions that potentiate the evolution of antibiotic resistance.[19]

Both evolution-based and supervised approaches consider only the DNA or protein sequence as input. They generally assume that all sites are equally likely to affect the phenotype, and consequently many examples of a mutation are needed to infer significant effects. This assumption is not true: proteins have three-dimensional shapes and functional regions, and not every mutation is equally likely to impact function. Analyzing mutations in their three-dimensional context has uncovered hotspots of mutation in cancer,[20–24] and provided post-hoc rationale for resistance-conferring variants in *Mycobacterium tuberculosis*.[19,25] Protein three-dimensional structure has shown utility as an input feature for identifying resistance-conferring variants in known resistance-conferring proteins such as RpoB,[26,27] PncA[28–30], and AtpE.[31]While past work has sought to reannotate parts of the *M. tuberculosis* proteome with computationally predicted protein structures using older structure prediction methods,[32] we can now infer a protein structure for nearly every protein in the proteome using AlphaFold,[33] leading to new works examining the 3D location of mutations in known and suspected resistance-conferring proteins.[34,35]

In this paper, we use an unsupervised method to discover clustering of mutations in *M. tuberculosis* proteins, by integrating a proteome wide-structural database with mutations from over 31,000 clinical isolates. We show that mutations display statistically significant clustering in resistance-conferring genes, even in non-essential proteins such as PncA where individually rare inactivating mutations are thought to cause resistance.[36] We identify over 450 proteins in the *M. tuberculosis* proteome that have significant clustering of mutations in their structures. Finally, we show that protein structural information provides a useful feature to predict whether variants are associated with antibiotic resistance, across all proteins with resistance variants

Results

Characterization of protein-modifying mutations in M. tuberculosis

We aimed to study how acquired missense variation in the *M. tuberculosis* proteome distributes in the structure of each protein. We chose to analyze the number of independent arisals of each mutation (homoplasy) rather than their population-level frequency, because analyzing the frequency of alleles in a population can be biased by oversampling of particular lineages, and by evolutionary recency,. This ensures that more recent evolutionary events are not under-represented due to lack of time to spread in the population. To accomplish this, we used a previously compiled a dataset of genomes of 31,428 isolates from the Mycobacterium tuberculosis

complex (MTBC), with ancestral sequence reconstruction to determine the number of independent arisals of each mutation (Supplementary Data 1 ↗).[18,19,37] The dataset represents a diversity of MTBC isolates: with 2,815 isolates from lineage 1; 8,090 from Lineage 2; 3,398 from Lineage 3; 16,931 from Lineage 4; 98 from Lineage 5; and 96 from Lineage 6.

We filtered the 782,565 unique SNVs and 47,425 insertion/deletion mutations (INDELs) found in the original dataset to focus only missense mutations in protein coding genes, or INDELS that preserve the original translation frame. We count the number of independent arisals (homoplasy score) for each SNV and INDEL. As we are looking for positive selection that alters but does not completely ablate protein function, we do not consider frameshift mutations. After excluding mutations occurring in regions where variant calling is computationally challenging with short-read sequencing,[38,39] we observe a total of 469,942 total unique missense mutations and 5,104 total inframe indels (**Methods**, Supplementary Data 2 ↗). On a per-protein basis, a mean of 32.0% of sites have at least one mutation across the dataset of 31,428 isolates, with a mean of 1.4 mutations per mutated site.

We observe that proteins with the highest total number of mutations are those known to be involved in resistance to first- and second-line antibiotics (Figure 1 ↗). We also observe a large number of homoplastic mutations in the protein RpoC, in which mutations can compensate for fitness loss after evolution of rifampicin resistance via mutations in RpoB.[40] Thus, this is a marker of but not a direct cause of antibiotic resistance. Another protein with many homoplastic mutations is Cas10, a component of the CRISPR-Cas system. The vast majority of mutations events (N=1539) in this protein are due to a short inframe indel at position 3,131,469 in the H37Rv reference genome, found in over 5,000 isolates, which given the repetitive nature of the sequence region may indicate phase variation (Supplementary Figure S1 ↗). Our analysis does not return ribosomal RNA genes, which are important causes of aminoglycoside resistance, because we are restricted to protein-coding genes. Finally, we note that we do not find large numbers of homoplastic mutations in genes encoding resistance to newly repurposed and introduced antibiotics linezolid, pretomanid, delamanid, bedaquiline, and clofazimine, because the majority of our dataset was sequenced before wide adoption of these drugs

Structures of the whole M. tuberculosis proteome

We identified an experimentally measured or predicted structure for each protein in the *M. tuberculosis* proteome. We used a sensitive pipeline to search the RCSB Protein structure databank (**Methods**) for structures homologous to *M. tuberculosis* proteins. We find a protein structure that represents at least 90% of the protein sequence for 34% of proteins (N = 1,371). For the remainder of proteins, we use the structure downloaded from the AlphaFold protein structure database, removing residues where the structure prediction is uncertain (pLDDT < 70). We exclude proteins with known low accuracy of variant calling (**Methods**),[32,38] or with fewer than 30 residues represented in the filtered protein structure, for a final dataset of 3,687 proteins out of 3,996 annotated ORFs in the *M. tuberculosis* H37Rv reference proteome (Supplementary Data 3 ↗).

Antibiotic resistance proteins demonstrate mutational clustering

To calculate clustering of mutations, we use an approach from geographic analysis that calculates spatial autocorrelation between statistics, called the Getis-Ord G-statistic.[45,46] Briefly, for each residue *i* in the protein, it calculates a *z*-score that computes if residue *i* is more proximal to highly mutated residues than would be expected by chance. This approach was previously applied to three-dimensional structure data by PIVOTAL to prioritize mutations associated with human disease.[47]

We examine the raw Getis-Ord score, which can be interpreted as a z-score, on the structures of 9 proteins in which mutations are known to confer antibiotic resistance[44] – RpoB, EmbB, KatG, InhA, PncA, GyrA, RsmG (GidB), EthA, and Rs12 (**Methods**). For each protein, we observe a region with residue G-scores greater than 5 that are clustered in a single location (Figure 2 ↗). This finding is expected for essential drug target proteins where resistance-conferring mutations preserve protein function while occluding antibiotic binding, *e.g.* RpoB, EmbB, InhA, GyrA, as

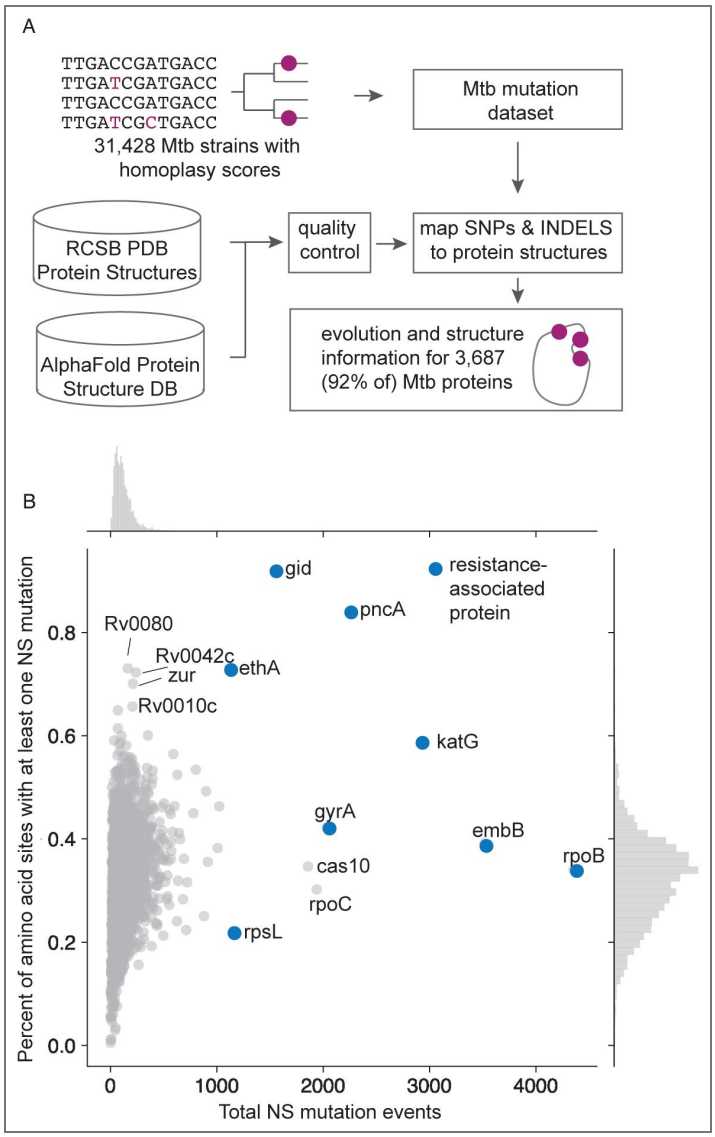

Figure 1. Proteins with the highest frequency of mutation events are associated with antibiotic resistance.

(A) Workflow used to create our combined dataset of missense substitutions and inframe INDEL mutations mapped to protein 3D structures, for 92% of the *M. tuberculosis* H37Rv proteome. With a dataset of homoplastic mutations from 31,428 MTBC isolates[19,41], we mapped these mutations to protein sequences, and 3D structures based on a combination of experimentally determined (RCSB PDB[42]) and computationally predicted (Alphafold[43]) structures. **(B)** The total number of mutation events in our dataset per protein, versus the percent of the amino acids in the protein's structure that have been mutated at least once. Marginal histograms are displayed along both axes. Proteins with the highest frequency of mutations are those associated with resistance to antibiotics, according to the WHO catalogue of resistance-associated mutations[44].

mutations will tend to cluster in the antibiotic binding sites of those proteins. Notably we also identify significant clustering for proteins that are not essential to the cell, where any mutation that disrupts the protein can be resistance-conferring, *e.g.* PncA. We observe that the residue G-score distribution demonstrates a periodic pattern across the gene length in which higher scores alternate with lower scores even when the high G-score residues cluster in a single protein location (Figure 2), a signal which emerges due to the 3D structure of the protein.

A protein-level statistic to test for mutational clustering

The G-score provides a residue-level z-score measuring three-dimensional proximity to other residues with high numbers of amino acid substitutions. We postulate that randomly accumulating substitutions would be evenly distributed in the 3-D structure of the protein. Uneven distribution of substitutions in 3D space may indicate selection; purifying selection leads to a depletion of substitutions in the hydrophobic cores of proteins, which must maintain stability,[52] whereas positive selection may lead to mutations that cluster in protein-protein interaction interfaces or ligand binding pockets.

We evaluated five candidate protein-level statistics to quantify clustering of mutations possibly indicative of positive selection (**Methods**, Supplementary Data 4). We selected as a positive control set of proteins the nine above proteins known to be under selection for antibiotic resistance, with demonstrated residue-level clustering– RpoB, EmbB, KatG, InhA, PncA, GyrA, RsmG (GidB), EthA, and Rs12. Because each of these proteins in an outlier in terms of the overall number of mutations observed (Figure 1B), we reasoned that downsampling the number of mutations would produce controls that were more similar to the average protein in the proteome. Hence for each of the nine proteins, we generate 100 positive control examples by downsampling the number of mutations observed while keeping the same relative probabilities of mutations at each site (Figure 3, **Methods**), thus preserving the signal for clustering while forcing their total number of mutations to be similar to that observed across all proteins in the proteome. We generated negative controls for the same nine proteins by simulating mutations from a uniform distribution across all residues, setting the mutation rate to the mean per-site mutation rate across the proteome, and generating 100 negative controls per protein. We find the Kolmogorov-Smirnov statistic to have the highest discrimination between positive and negative controls (95% precision, 68% recall, at significance level $p < 0.01$).

Significant mutational clustering across MTB proteome

We test all 3,687 proteins in the *M. tuberculosis* structure and homoplasy dataset, and identify 499 proteins with significant clustering (Figure 4, Supplementary Data 5).

We assay whether all proteins known to cause antibiotic resistance demonstrate mutational clustering, using the WHO catalog of resistance-conferring mutations[44]. The majority of proteins with Tier-1 resistance conferring mutations (8 of 12, 67%) demonstrate significant clustering. The four proteins without clustering include TlyA and RplC, in which there is only one missense substitution known to confer resistance, and MmpR5 a transcriptional regular protein involved in resistance to the newly administered drug bedaquiline. The fourth protein is RpsL (RS12_MYCTU), whose G-score is driven by just two highly mutated residue positions, K43 and K88 (Supplementary Figure S2). In a small protein such as RpsL (124 amino acids), having two highly mutated residues close in 3D structure occurs with some frequency in the negative control examples.

For 90.6% (452 of 499) of the significant hits, the top G-score pair of residues are within 15 Ångstroms in 3-D space, confirming that the identified clusters are in a single location within the protein. Decreasing the distance threshold to 8 and 5 Ångstroms results in 74.7% (373) and 61.1% (305) hits whose top pairs are close in 3-D, respectively (Supplementary Figure S3). We find that distance between top pairs of residues does not increase with protein length, a proxy for size (Supplementary Figure S4). Proteins with a high number of mutations that are not spatially clustered are not significant hits. This includes Cas10, which has 1861 independent mutation events, 1489 of which occur in the same amino acid position.

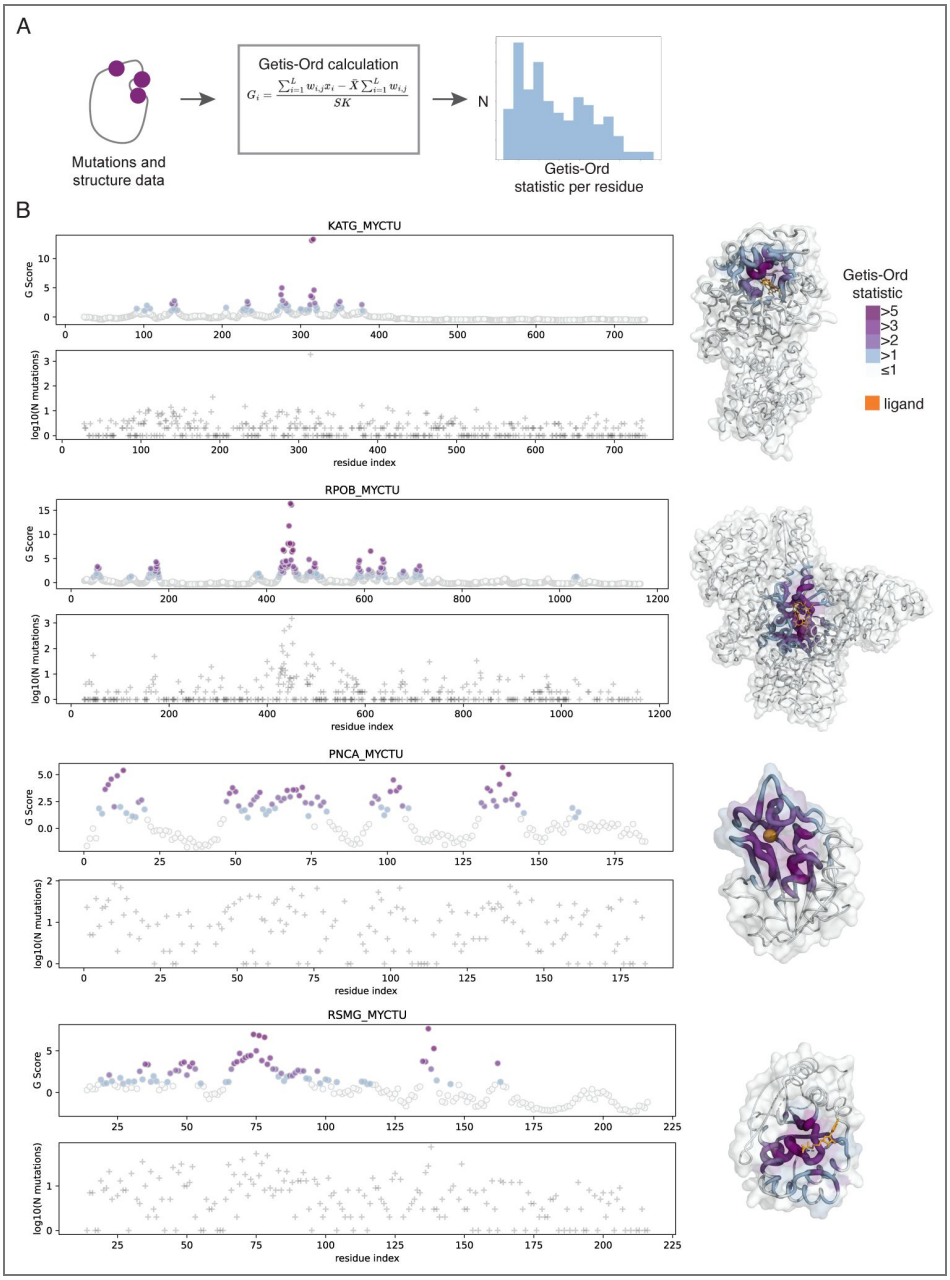

Figure 2. G-statistic reveals clustering of mutations in antibiotic-resistance conferring proteins.

(A) Workflow to compute residue-wise Getis-Ord statistic for proteins in *M. tuberculosis.* (B) Results for proteins: KatG is shown in complex with heme (orange), PDB ID = 4C51 chain A.[48] RpoB is shown in complex with rifampin (orange), PDB ID = 5UH6 chain C,[49] aligned as described in **Methods**. PncA is shown in complex with Fe2+, PDB ID = 3PL1 chain A.[50] RsmG (encoded by the *gidB* gene) structure from *Thermus thermophilus* is shown in complex with ligand adenosine monophosphate (please note the streptomycin binding site is unknown in *M. tuberculosis* and thus is not shown here), PDB ID = 3G8A chain A.[51]

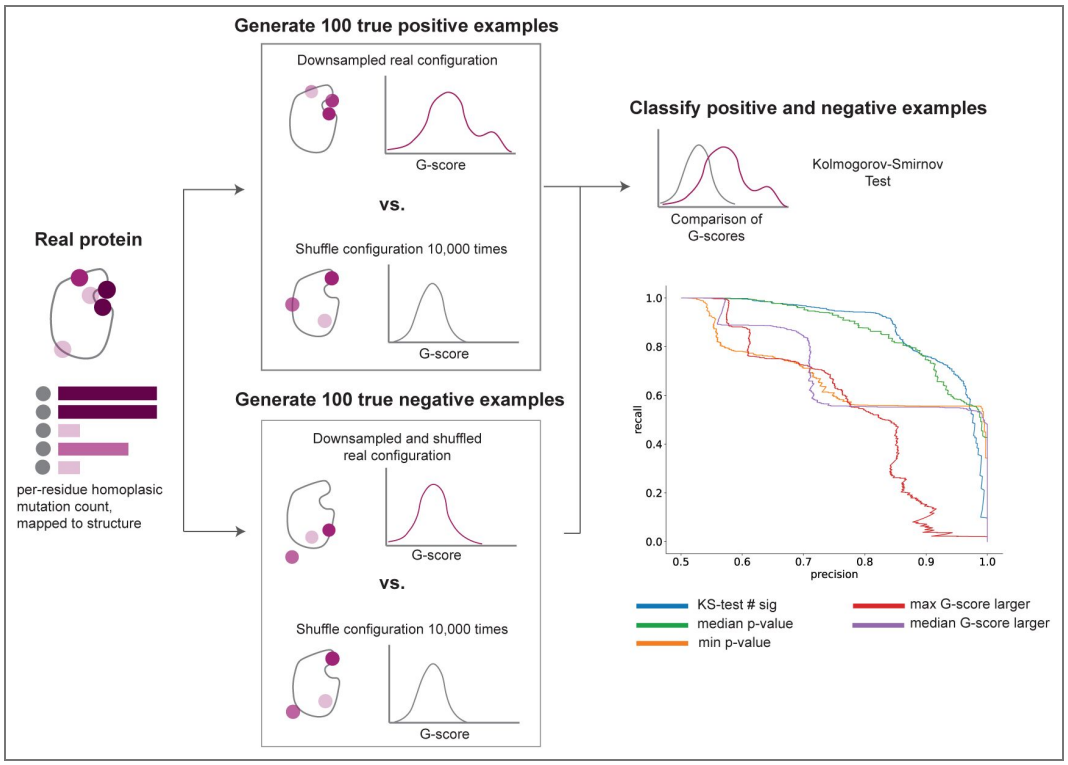

Figure 3. Benchmarking the ability of G-scores to find significant clustering in protein structures.

The procedure for generating downsampled true positive and true negative examples from real proteins. We then test five scores for their ability to distinguish true positives and negatives.

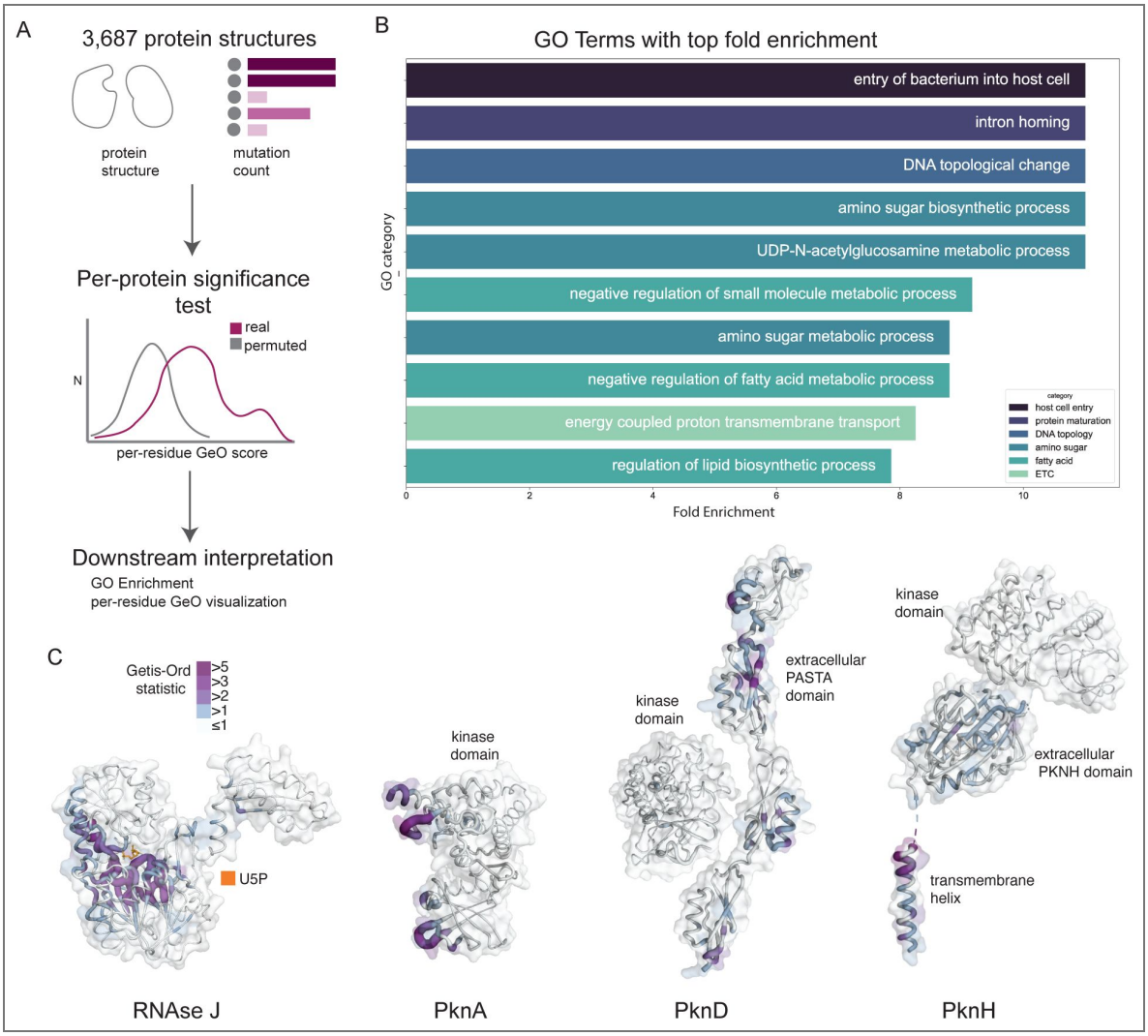

Figure 4. Hits of proteome-wide screen for clustering of mutations.

(A) pipeline for detecting hits in 3,687 proteins. (B) GO terms with top fold enrichment (all significant at FDR < 0.05). (C) Examples of proteins with significant clustering. All structures shown are from AlphaFold with low-confidence residues filtered out (note that relative domain orientation for PknB and PknH is low confidence).

We perform a Gene Ontology Enrichment analysis on the 452 hits with a close proximity top pair to understand their possible functional implications (**Methods**). After removing the less specific gene categories (with >20 genes), 54 GO categories are significantly enriched (FDR < 0.05) (Figure 4). The most enriched categories by fold change (Supplementary Table 1, Supplementary Data 6) are regulation of fatty acid metabolism and biosynthesis; UDP-N-acetylglucosamine/amino sugar metabolic processes; entry of bacterium into host; protein maturation; DNA topological change; and energy-coupled proton transport. We find clustering in the DNA topological change proteins GyrA and GyrB, both known to confer resistance to fluoroquinolones, and Top1, DNA topoisomerase I, which is not known to be associated with resistance.

A second notable GO category is composed primarily of the Pkn family of protein kinases – PknA, PknB, PknD, PknE, and PknH – thought to negatively regulate fatty acid biosynthesis. Data supports that PknA, PknB, and PknH have a role in regulating cellular permeability and thus intrinsic drug resistance, potentially through post-translational modification of proteins involved in cell wall synthesis.[53–55] For all of the Pkn proteins found by our clustering score, the region of significant clustering is not the protein kinase domain but instead in other functional domains of the protein. For example, the significant clustering in PknB is in the extracellular domains, and in PknH is in the putative transmembrane helix and sensor domain[56] (Figure 4). Therefore, we suspect that the mutations may relate to regulation of activity rather than the kinase reactions themselves.

A third notable group of proteins is involved in amino sugar metabolic processes: GlmM, GlmS, GlmU, and MurA. The related protein MurD was previously identified in a screen for targets of convergent positive selection in the *M. tuberculosis* proteome.[57] Amino sugars are a key component of the cell wall, and thus these proteins could be under positive selection for phenotypes including intrinsic drug resistance and adaptation to the host environment.

We find significant clustering in RnJ, Ribonuclease J, which has previously been identified in GWAS for antibiotic resistance,[5] and whose loss is known to lead to multi-drug tolerance.[58] which has been shown to have both ribonuclease and beta lactamase activity. We find significant clustering around the AlphaFill-inferred binding site of ribonucleotides (**Methods**), which corresponds to the known active site for RNase activity (Figure 4).[59]

Structure supports the prediction of resistance-conferring variants

Given that most drug resistance genes demonstrate 3-D mutational clustering, as detailed above, we tested if structural proximity to known mutations carries information on the functional impact of new mutations on phenotype. We use a previously compiled catalogue of resistance-associated mutations from *M. tuberculosis*,[9] consisting of 583 unique missense protein-coding variants annotated as associated with resistance (R-assoc) and 58 as not associated and isolates carrying only these mutation are antibiotic susceptible (S-assoc) (**Methods**, Supplementary Data 7)[44].

We ask whether protein structure information is useful at distinguishing resistance-conferring variants. In addition to the G-score computed using the homoplasy data from the ~31,000 isolate dataset, we compute the distance in 3D between any variant and the nearest resistance-conferring variant in that protein from the catalogue. We benchmark both metrics against distance in 1D (on the protein sequence). We trained a logistic regression classifier to predict whether a mutation is resistance associated (R-assoc, *vs.* S-assoc, susceptibility associated), using a 70-30 train-test split of the cataloged mutations. 3D structural proximity alone best predicted resistance association (F1 score of 94.6%, Table 1, **Methods**). Prediction from 1D proximity alone is less accurate at F1 of 92.8%, and prediction from G-score had the lowest F1 at 80.8%. The relatively strong performance of 1D proximity prediction is due to the fact that most, but not all, mutations in our dataset that have close 3D proximity to a mutated residue also have close 1D proximity.

While the G-score prediction has the lowest F1 score and sensitivity, it has the highest precision (98.3%). This may be because the G-score is a combination of structural information and mutational information, as a residue will only have a high G-score if it is proximal to at least one

Feature Set	F1	Precision/PPV	Sensitivity
3D-Proximity	94.6	97.5	91.9
1D-Proximity	92.8	96.3	89.5
G-score	80.8	98.3	68.6

Table 1. Performance of classification models on predicting whether mutations are R-conferring from the mutation catalog.

Proximity-1D was trained using just the distance in primary sequence to the nearest known R mutation, Proximity-3D was trained using the distance in 3D to the nearest known R mutation, and G-score was trained using the G-score calculated in this manuscript.

residue with a high number of mutations (presumably due to being causal of resistance). This means that the G-score model is more conservative at calling R-assoc variants as truly R-assoc.

Discussion

In this paper, we demonstrate that mutations cluster in three dimensions in the structure of proteins known to confer antibiotic resistance, through a novel application of the Getis-Ord statistic to evolutionary sequence data from clinical tuberculosis isolates. We apply our method to all proteins in the *M. tuberculosis* H37Rv proteome, using homoplasy data generated from a dataset of over 30,000 *M. tuberculosis* complex isolates. We find significant clustering of mutations in over 450 proteins, including eight known resistance-conferring proteins, and in pathways thought to be important for pathogenesis and antibiotic resistance.

We observe spatial clustering of mutations in known resistance-conferring proteins. For proteins like RpoB, where mutations are known to cluster in the rifampicin binding site, this serves as an internal control for our method. However, for proteins where mutations that lead to loss of function are known to cause resistance, such as PncA and RsmG (GidB), it is not necessarily expected to find clustering of mutations. We suspect that the observed clustering is due to mutations in a certain region of the protein being more likely to cause loss of function.

When running our analysis on the entire proteome, we find a significant enrichment for hits in proteins involved in regulation of fatty acid biosynthetic processes, and in amino sugar metabolism. Both of these pathways are related to cell envelope synthesis, which may be a response to antibiotic pressure or an adaptation to host environments. We also find hits to the proteins MycP1, P2, and P3, which are secreted from the cell and thought to be involved in host cell entry. We chose to apply our approach to homoplasy data, not allele frequency data, to ensure that we were not biased to overweight more ancient mutations in our analysis. Thus, we believe significant Getis-Ord score clustering indicates protein hotspots of positive and/or diversifying selection, where *M. tuberculosis* is adapting to antibiotic pressure and host environments.

Our variant effect prediction results support that structural information is a meaningful feature for predicting whether specific genetic variants confer antibiotic resistance. Distance to ligand binding sites have previously been shown to be a meaningful feature for classification of resistance in the proteins RpoB,[26,27] PncA,[28] and AtpE,[31] and our work extends this to all proteins with known resistance-conferring mutations. These results on resistance variant classification how potential methods for expanding the catalog of known resistance-conferring mutations. In addition, recent efforts to predict *M. tuberculosis* antibiotic resistance from sequence data have suggested that our ability to predict antibiotic resistance from sequence alone may be reaching a plateau, and that more data and different data modalities, not more innovative model architectures, are needed in order to further improve predictions.[12] While we do not suggest that our current model be used in clinical settings, we envision structure being added as an additional feature in future work to predict organismal resistance phenotypes from sequences, as it has already been successfully used in resistance variant classification.[34,35]

One drawback of our approach is that our random shuffling procedure implicitly assumes that all sites in a protein are equally likely to mutate. This may pose a problem for proteins with conserved hydrophobic cores, or multidomain proteins where one domain is more conserved than the other, which could manifest in apparent higher or lower rates of mutation in one region of the protein. We account for this bias when we compute the distance between the top-scoring pair of residues in each protein, finding that for over 90% of the proteomic hits, those two residues are within 15 Ångstroms of one another, indicating that the patch of high scoring residues are in a single location on the protein, not spread throughout a surface or domain.

However, we cannot entirely rule out the effects of differing mutation rates throughout a protein. Future work could consider simulating the accumulation of substitutions according to their predicted effects on protein stability, to better reflect a real-world evolutionary scenario.

The research community has long sought to determine the genetic basis of antibiotic resistance in *M. tuberculosis*, using both supervised methods that rely on labelled data, and unsupervised methods that seek to find patterns of positive selection indicative of antibiotic resistance. The availability of predicted protein structures has opened a new avenue for understanding the effects of mutations in microbial genomes, and in *M. tuberculosis* specifically. We hypothesize that protein structural proximity is a useful feature because it captures how likely any given pair of mutations are to have a similar phenotypic effect – mutations in the same functional regions of proteins are more likely to cause the same effect. We show that protein 3D structure can be used to classify resistance-conferring variants across all proteins in a supervised framework, and used in an unsupervised fashion to discover targets of positive selection in the *M. tuberculosis* genome.

Methods

Mycobacterium tuberculosis mutation dataset

We use a dataset of SNPs and indels found in *M. tuberculosis* isolates, mapped to their occurrences on a phylogenetic tree, published by previous work.[18,37] This dataset was constructed using a previously validated pipeline for calling variants in *M. tuberculosis*, by mapping reads to the H37Rv reference genome using BWA-MEM v0.7.17[60,61] after trimming and filtering with PRINSEQ v0.20.4,[62] contaminant removal with Kraken v0.10.6,[63] and duplicate read removal with Picard v2.9.2. All isolates were required to have at least 95% of bases in the reference genome with at least 10x coverage. Variant calling was performed with Pilon,[64] and additional quality control filters were applied to ensure accurate allele calls. The resulting dataset was two matrices: a list of all positions found to have a SNP in any isolate, relative to the H37Rv reference (N=782,565), and a list of all positions found to have an insertion or deletion in any isolate, relative to the H37Rv reference (N=47,425).[18]

Phylogeny and ancestral sequence reconstruction were performed by Vargas et al as described in a previous publication.[18] In their procedure, phylogenies were constructed separately for sub-lineages (L1, L2, L3, L4A, L4B, L4C, L5, L6) for computational feasibility. Phylogenies were constructed from concatenated SNP data using IQTree.[65] SNPPar[66] was used to reconstruct SNVs, and the method previously published by Vargas *et al.* was used to reconstruct INDELs.[41]

Searching PDB structure database

We search the RCSB Protein Data Bank structure database (download date: Feb 5, 2021) for experimentally determined structures with sequence similarity to the Uniprot Proteome of Mycobacterium tuberculosis (ID = UP000001584). In order to execute a sensitive search for homologous structures, we built a modified version of the EVcouplings pipeline,[67] which operates in two stages: first, we construct sequence alignments against the Uniprot sequence database (download date Feb 5, 2021) at bitscores of 0.1, 0.2, and 0.3 times the query sequence length, using jackhmmer from hmm-suite v3.1 with five iterations[68]. Second, we use the hmmbuild and hmmsearch tools of hmm-suite to search the RCSB PDB sequence database using the constructed alignments.

Selecting structure hits

We then aim to select high-coverage protein structures from the experimental structure database to represent the *M. tuberculosis* proteins. We define coverage as whether an amino acid residue in the *M. tuberculosis* protein is represented by a resolved amino acid in the protein structure. Of the 3993 proteins in the proteome, 2984 (75%) have any hits to the structure database. 1509 (38%) of these hits have at least 90% coverage. In cases where more than one hit had 90% coverage, we select the representative hit in the following way: if a protein structure is from *M. tuberculosis*, we select that structure, otherwise we select the hit with the lowest e-value.

Using AlphaFold database

Predicted structures for the Mycobacterium tuberculosis reference proteome (ID = UP000001584) were downloaded from the AlphaFold Protein Structure Database on Feb 15, 2023. Residues with pLDDT < 70 were removed from each structure.

Preparing protein structure data for analysis

We exclude from consideration proteins with documented difficulties in mutation calling using short-read sequencing. Using data from Marin *et al*[38], proteins with <90% mean empirical base pair recall (across all residues in the protein) or <90% mean residue mappability were removed from consideration. Proteins with fewer than 30 residues with resolved structure coordinates were removed, for a final total of 3687 proteins, 1350 of which are experimentally determined structures and the remaining 2337 of which are from AlphaFold.

Preparing mutations for analysis

For SNP mutations, we consider all missense SNPs that occur at least once in our *M. tuberculosis* dataset, for a total of 476,607 unique polymorphisms. For insertion and deletion (indel) mutations, we exclude frameshift mutations as these are likely to ablate protein function and thus are not analogous to missense mutations. We observe 5,678 unique in-frame indels. Combining the SNP and inframe indel mutations, we observe 483,117 unique mutations. To account for possible artifacts from short-read sequencing base calling errors, we exclude mutations in positions designated as blindspots by Modlin *et al* or with empirical base-pair recall <90%,[38,39] as well as all mutations occurring in proteins with <90% mean empirical base-pair recall or <90% mean mappability.[22] After filtering, we have 475,046 unique polymorphisms for downstream analysis (469,942 total unique missense mutations and 5,104 INDELs) (Supplementary Data 2).

Merging mutation and structure data

Upon merging our structure dataset with our mutation dataset, 335,224 mutations can be mapped to a residue in a protein with high-quality structure information. We exclude proteins with fewer than two mutations mapped to their structure from the clustering calculation. Indel mutations were treated as occurring at the first position in the sequence. We proceed with clustering analysis on 3657 proteins with mutations mapped.

Computing inter-residue distances

The EVcouplings Python package was used to compute the distance between wild-type amino acid residues in all protein structures.[67] The package calculates the distance between all heavy (non-hydrogen) atoms in residue *i* and residue *j*, then returns the minimum of those distances.

Computing the Getis-Ord score for clustering of homoplastic mutations

Our implementation of the Getis-Ord score for protein three-dimensional structure was inspired by PIVOTAL, which applied the Getis-Ord score to clustering of mutations in human disease.[45–47] The two values input to the Getis-Ord statistic computation are a per-residue score $\mathbf{x}$, here the per-amino acid homoplasy score, and a weight matrix W that contains the inverse of the inter-residue distances computed from the wild-type amino acids. $\mathbf{x}$ is an L x 1 vector where the entry x_i contains the homoplasy score for the i-th residue in the protein, and $\mathbf{W}$ is an L x L matrix with entries defined as:

$$w_{i,j} = \begin{cases} \frac{1}{d_{i,j}} \text{ if } i \neq j \\ 0 \text{ if } i = j \end{cases}$$

Where $d_{i,j}$ is the minimum inter-residue atomic distance, in Å, between i and j. Thus, residues that are closer in three dimensions will have higher weights.

The Getis-Ord statistic for residue i is calculated as:

$$G_i = \frac{\sum_{j=1}^{L} w_{i,j}x_j - \bar{X}\sum_{j=1}^{L} w_{i,j}}{SK}$$

Where $\bar{X}$ is the mean of **x**, and:

$$S = \sqrt{\frac{\sum_{j=1}^{L} x_j^2}{L} - \bar{X}^2}$$

$$K = \sqrt{\frac{L\sum_{j=1}^{L} w_{i,j}^2 - \left(\sum_{j=1}^{L} w_{i,j}\right)^2}{L-1}}$$

Preparing GeO score calibration data

We sought to generate a dataset of positive controls – proteins known to have clustering of mutations – with a frequency of mutation similar to the average protein in our dataset (mean of 1.4 mutations per site). For this, we used the nine proteins known to be involved in antibiotic resistance with demonstrated clustering of mutations (RpoB, EmbB, KatG, InhA, PncA, GyrA, RsmG (GidB), EthA, and Rs12). For each of these nine proteins, we generate 100 positive control examples.

Because the total number of mutations in a site, as well as the 3D configuration of sites, contributes to the G-score, we chose to downsample these proteins to generate positive controls that are more similar to other proteins in the total number of mutations. To generate positive control examples, we build an empirical distribution based on the observed number of mutations per site in the protein, divided by the total mutations. We add a pseudocount of one to all residues with zero mutations. We then sample M mutations from the empirical mutation distribution, where M = 0.42 times the number of residues in the protein with structure data, because each site is mutated on average 0.42 times. This sampling strategy preserves the relative frequencies of each mutation, and their 3D location, while greatly reducing the number of total mutations observed in a protein. We generate the negative controls by sampling from a uniform distribution over all residues in a protein with structure data. Note that we do not re-compute inter-residue distances when simulating mutations in an amino acid, as the distances used as input to GeO score are the wild-type inter-residue distances.

Then, we build a heuristic based on the Getis-Ord score to determine if we observe significant clustering in a protein. For each control example, we perform 10,000 random permutations of the mutations, shuffling their configuration in space but keeping the number of mutation events the same. For positive controls, we expect the real configuration of mutations to produce a significantly different configuration of G scores than random permutations. For negative controls, we do not expect the real configuration of mutations to produce a significantly different configuration of G scores than random permutations.

Next, for each generated control, we test the ability of various scores to distinguish between the real example and a random reshuffling. By comparing the true distribution with the empirical random distribution, we compute a p-value for whether the overall distribution of Getis-Ord statistics for residues in the protein is different than the one expected by chance. We compare six scores on the basis of their precision recall curves for recovering the positive controls. The first three scores are based on comparing the p-value of the Kolmogorov-Smirnov test when comparing the real distribution to 10,000 random shuffles: the number of significant p-values ($p < 0.01$), the smallest p-value, and the mean p-value. The other two features are based on the raw G-score for the real distribution vs. the 10,000 random shuffles: how often the max G-score is greater for the real distribution, and how often the median G-score is greater for the real distribution.

Running on whole proteome

We then run the G-score calculation on the whole proteome using the above pipeline. We apply our 95% precision threshold to find 451 initial hits.

GO Enrichment Analysis

We test whether our gene hits are enriched in particular GO functional categories.[69,70] We use the online GO enrichment tool (*https://geneontology.org/*) to search for enriched biological processes among our 452 hit proteins. We downloaded the .json file from the GO enrichment tool, and filtered for GO categories with fewer than 20 members in the *M. tuberculosis* H37Rv refence genome, to remove overly broad categories like "biological process," and terms with identical constituent genes. We retained categories with FDR < 0.05, for a total of 37 enriched terms.

AlphaFill

We downloaded hits from the AlphaFill v1 database (access date: 7 October 2024) to find potential ligand-binding locations in our proteins. For RnJ, two ligands are in proximity of the high GeO score regions: "U5P" and "C5P", uridine-5'-monophosphate and cytidine-5'-monophosphate.

Parsing the WHO mutation catalog

The World Health Organization (WHO) Mutation Catalogue maintains a list of over 30,000 unique variants observed in *M. tuberculosis* genomes and whether those mutations are diagnostic of antibiotic resistance against 13 antibiotics[71]. Mutations are graded with five confidence categories: (1) Associated with Resistance, (2) Associated with Resistance - Interim, (3) Uncertain Significance, (4) Not Associated with Resistance - Interim, and (5) Not Associated with Resistance. Kulkarni *et al.* provided an update to this catalog which increased the number of labelled variants using regression-based grading based on frequency of mutations in resistant and susceptible isolates.[9]

We limit our analysis to the variants observed in the following 15 protein-coding genes: *Rv0678*, *atpE*, *ddn*, *embB*, *ethA*, *gid*, *gyrA*, *gyrB*, *inhA*, *katG*, *pncA*, *rplC*, *rpoB*, *rpsL*, *tlyA*. After removing duplicates – because multiple genomic variants can cause the same missense mutation – we were left with 9365 variants, composed mostly of uncertain variants. After filtering for variants with structure information, we have 583 R-assoc (category 1 or 2), and 58 as S-assoc (category 4 or 5).

Fitting a classifier on the WHO mutation catalog

For each of 9365 unique missense variants in the catalogue,[9] we extracted the following features: The minimum coordinate difference to the nearest non-self R variant along the amino-acid sequence ("1D proximity"). The inter-atomic distance to the nearest non-self R variant in Ångstroms ("3D proximity"), the G-score of the residue as computed in our previous analysis. 3D proximity was calculated using the EVcouplings Python package.[67]

We used a 70-30 train-test split to construct a dataset. We employed weighted sampling to address the class imbalance between resistance-conferring and non-resistance-conferring variants. We employed a logistic regression classifier in scikit-learn, and hyperparameters were tuned by grid search.[72] Models were evaluated for F1-Score, precision, and recall. To select an optimal model threshold, we choose the threshold which maximizes the sum of sensitivity and specificity.[73]

Data availability

Code is available on GitHub at *https://github.com/aggreen/MTB_Mut_Clust*, and the modified EVcouplings pipeline for structure search is found at *https://github.com/aggreen/EVcouplings/tree/feature/structure_finder*. All strains used in our analyses are publicly available, and the raw read data are available for download from the NCBI using accession codes found in the isolate annotation table. Data and code for ancestral sequence reconstruction data from Vargas et al, PNAS, 2023 is available here: *https://github.com/farhat-lab/phase-variation-in-Mtbc*. WHO mutation classification data 2nd edition is publicly available here: *https://www.who.int/publications/i/item/9789240082410*

Acknowledgements

We thank members of the Farhat lab at Harvard Medical School and the SAGE lab at University of Massachusetts Amherst for valuable discussion about the project. Computational resources and support were provided by the Orchestra High Performance Compute Cluster at Harvard Medical School, which is funded by the NIH (NCRR 1S10RR028832-01). A.G.G. was supported by NIH/NIAID F32AI161793. R.V.J. was supported by the National Science Foundation Graduate Research Fellowship under Grant No. DGE1745303.

Additional information

Author Contributions

A.G.G. and M.R.F. conceived the study and designed the analyses. A.G.G. implemented the mutation clustering code. M.T. implemented the resistance mutation classification code. A.G.G., M.T., and M.R.F. interpreted the data and results. R.V.J contributed data and provided important discussions of results. M.R.F. and A.G.G. supervised the research. A.G.G., M.T., and M.R.F. wrote the manuscript with input from all authors.

Funding

Funder	Grant reference number	Author
HHS | National Institutes of Health (NIH)	F32AI161793	Anna G Green
NSF | National Science Foundation Graduate Research Fellowship Program (GRFP)	DGE1745303	Roger Vargas

Author ORCID iDs

Anna G Green: https://orcid.org/0000-0001-7548-3682
Mahbuba Tasmin: https://orcid.org/0000-0003-1884-8838
Maha R Farhat: https://orcid.org/0000-0002-3871-5760

Additional files

Supplementary Figures and Tables

Supplementary Data 1

Supplementary Data 2

Supplementary Data 3

Supplementary Data 4

Supplementary Data 5

Supplementary Data 6

Supplementary Data 7

References

1. **World health Organization** (2021) Global Tuberculosis Report.
2. **Global Tuberculosis Programme, W. H. (hq)** (2021) WHO Consolidated Guidelines on Tuberculosis. Module 3: Diagnosis - Rapid Diagnostics for Tuberculosis Detection 2021 Update.
3. **Lange C,** et al. (2018) Drug-resistant tuberculosis: An update on disease burden, diagnosis and treatment. *Respirology* **23**:656-673 https://doi.org/10.1111/resp.13304 | PubMed
4. **Bush WS,** Moore JH (2012) Chapter 11: Genome-wide association studies. *PLoS Comput Biol* **8**:e1002822 https://doi.org/10.1371/journal.pcbi.1002822 | PubMed

5. **Farhat MR,** et al. (2019) GWAS for quantitative resistance phenotypes in Mycobacterium tuberculosis reveals resistance genes and regulatory regions. *Nat Commun* **10**:2128 https://doi.org/10.1038/s41467-019-10110-6 | PubMed
6. **Gröschel MI,** et al. (2021) GenTB: A user-friendly genome-based predictor for tuberculosis resistance powered by machine learning. *Genome Med* **13**:138 https://doi.org/10.1186/s13073-021-00953-4 | PubMed
7. **Conkle-Gutierrez D,** et al. (2022) Distribution of Common and Rare Genetic Markers of Second-Line-Injectable-Drug Resistance in Mycobacterium tuberculosis Revealed by a Genome-Wide Association Study. *Antimicrob Agents Chemother* **66**:e02075-21 https://doi.org/10.1128/aac.02075-21 | PubMed
8. **Coll F,** et al. (2018) Genome-wide analysis of multi- and extensively drug-resistant Mycobacterium tuberculosis. *Nat Genet* **50**:307-316 https://doi.org/10.1038/s41588-017-0029-0 | PubMed
9. **Kulkarni SG,** et al. (2024) Regression for accurate and sensitive grading of mutations diagnostic of antibiotic resistance in Mycobacterium tuberculosis. *medRxiv* https://doi.org/10.1101/2024.07.01.24309598
10. **Green AG,** et al. (2022) A convolutional neural network highlights mutations relevant to antimicrobial resistance in Mycobacterium tuberculosis. *Nat Commun* **13**:3817 https://doi.org/10.1038/s41467-022-31236-0 | PubMed
11. **Chen ML,** et al. (2019) Beyond multidrug resistance: Leveraging rare variants with machine and statistical learning models in Mycobacterium tuberculosis resistance prediction. *EBioMedicine* **43**:356-369 https://doi.org/10.1016/j.ebiom.2019.04.016 | PubMed
12. **Wang Y,** et al. (2024) TB-DROP: deep learning-based drug resistance prediction of Mycobacterium tuberculosis utilizing whole genome mutations. *BMC Genomics* **25**:167 https://doi.org/10.1186/s12864-024-10066-y | PubMed
13. **Pruthi SS,** et al. (2024) Leveraging large-scale Mycobacterium tuberculosis whole genome sequence data to characterise drug-resistant mutations using machine learning and statistical approaches. *Sci. Rep* **14**:27091 https://doi.org/10.1038/s41598-024-77947-w | PubMed
14. **Serajian M,** et al. (2024) Scalable de novo classification of antibiotic resistance of Mycobacterium tuberculosis. *Bioinformatics* **40**:i39-i47 https://doi.org/10.1093/bioinformatics/btae243 | PubMed
15. **Kulkarni SG,** et al. (2025) Convolutional neural networks quantify antibiotic resistance in Mycobacterium tuberculosis with diagnostic grade accuracy and predict treatment response. *medRxiv* https://doi.org/10.1101/2025.08.05.25333066
16. **Pal A,** Mohanty D (2025) Machine learning-based approach for identification of new resistance associated mutations from whole genome sequences of Mycobacterium tuberculosis. *Bioinforma Adv* **5**:vbaf050 https://doi.org/10.1093/bioadv/vbaf050 | PubMed
17. **Desjardins CA,** et al. (2016) Genomic and functional analyses of Mycobacterium tuberculosis strains implicate ald in D-cycloserine resistance. *Nat Genet* **48**:544-551 https://doi.org/10.1038/ng.3548 | PubMed
18. **Vargas R,** et al. (2023) Phase variation as a major mechanism of adaptation in Mycobacterium tuberculosis complex. *Proc. Natl. Acad. Sci* **120**:e2301394120 https://doi.org/10.1073/pnas.2301394120 | PubMed
19. **Green AG,** et al. (2023) Analysis of Genome-Wide Mutational Dependence in Naturally Evolving Mycobacterium tuberculosis Populations. *Mol. Biol. Evol* **40**:msad131 https://doi.org/10.1093/molbev/msad131 | PubMed
20. **Miller ML,** et al. (2015) Pan-Cancer Analysis of Mutation Hotspots in Protein Domains. *Cell Syst* **1**:197-209 https://doi.org/10.1016/j.cels.2015.08.014 | PubMed
21. **Gao J,** et al. (2017) 3D clusters of somatic mutations in cancer reveal numerous rare mutations as functional targets. *Genome Med* **9**:4 https://doi.org/10.1186/s13073-016-0393-x | PubMed

22. **Kamburov A,** et al. (2015) Comprehensive assessment of cancer missense mutation clustering in protein structures. *Proc. Natl. Acad. Sci* **112**:E5486-E5495 https://doi.org/10.1073/pnas.1516373112 | PubMed

23. **Meyer MJ,** et al. (2016) mutation3D: Cancer Gene Prediction Through Atomic Clustering of Coding Variants in the Structural Proteome. *Hum. Mutat* **37**:447-456 https://doi.org/10.1002/humu.22963 | PubMed

24. **Niu B,** et al. (2016) Protein-structure-guided discovery of functional mutations across 19 cancer types. *Nat. Genet* **48**:827-837 https://doi.org/10.1038/ng.3586 | PubMed

25. **Phelan J,** et al. (2016) Mycobacterium tuberculosis whole genome sequencing and protein structure modelling provides insights into anti-tuberculosis drug resistance. *BMC Med* **14** https://doi.org/10.1186/s12916-016-0575-9 | PubMed

26. **Lynch CI,** Adlard D, Fowler PW (2025) Predicting rifampicin resistance in Mycobacterium tuberculosis using machine learning informed by protein structural and chemical features. *ERJ Open Res* **11**:00952-02024 https://doi.org/10.1183/23120541.00952-2024 | PubMed

27. **Portelli S,** et al. (2020) Prediction of rifampicin resistance beyond the RRDR using structure-based machine learning approaches. *Sci. Rep* **10**:18120 https://doi.org/10.1038/s41598-020-74648-y | PubMed

28. **Karmakar M,** Rodrigues CHM, Horan K, Denholm JT, Ascher DB (2020) Structure guided prediction of Pyrazinamide resistance mutations in pncA. *Sci. Rep* **10**:1875 https://doi.org/10.1038/s41598-020-58635-x | PubMed

29. **Carter JJ,** et al. (2024) Prediction of pyrazinamide resistance in Mycobacterium tuberculosis using structure-based machine-learning approaches. *JAC-Antimicrob Resist* **6**:dlae037 https://doi.org/10.1093/jacamr/dlae037 | PubMed

30. **Dissanayake D,** Brunner V, Adlard D, Morrone JA, Fowler PW (2026) Predicting pyrazinamide resistance in Mycobacterium tuberculosis using a graph convolutional network. *bioRxiv* https://doi.org/10.1101/2025.10.28.685176

31. **Karmakar M,** et al. (2019) Empirical ways to identify novel Bedaquiline resistance mutations in AtpE. *PLOS One* **14**:e0217169 https://doi.org/10.1371/journal.pone.0217169 | PubMed

32. **Modlin SJ,** et al. (2021) Structure-Aware Mycobacterium tuberculosis Functional Annotation Uncloaks Resistance, Metabolic, and Virulence Genes. *mSystems* **6**:e00673-21 https://doi.org/10.1128/msystems.00673-21 | PubMed

33. **Jumper J,** et al. (2021) Highly accurate protein structure prediction with AlphaFold. *Nature* **596**:583-589 https://doi.org/10.1038/s41586-021-03819-2 | PubMed

34. **Pal A,** Pal S, Mahapatra S, Pandey A, Mohanty D (2025) *In silico* analysis of the functional implications of drug resistance associated mutations in *Mycobacterium tuberculosis*. *Comput. Struct. Biotechnol. J* **27**:5425-5440 https://doi.org/10.1016/j.csbj.2025.11.054 | PubMed

35. **Wood JJ,** Portelli S, Ascher DB, Furnham N (2026) Risk-Based Prediction of Novel AMR Variants Using Protein Language Models. *bioRxiv* https://doi.org/10.1101/2025.09.12.672331

36. **Farhat MR,** et al. (2016) Genetic Determinants of Drug Resistance in Mycobacterium tuberculosis and Their Diagnostic Value. *Am J Respir Crit Care Med* **194**:621-630 https://doi.org/10.1164/rccm.201510-2091oc | PubMed

37. **Vargas R,** et al. (2021) Role of Epistasis in Amikacin, Kanamycin, Bedaquiline, and Clofazimine Resistance in Mycobacterium tuberculosis Complex. *Antimicrob Agents Chemother* **65** https://doi.org/10.1128/aac.01164-21 | PubMed

38. **Marin M,** et al. (2022) Benchmarking the empirical accuracy of short-read sequencing across the M. tuberculosis genome. *Bioinformatics* https://doi.org/10.1093/bioinformatics/btac023 | PubMed

39. **Modlin SJ,** et al. (2021) Exact mapping of Illumina blind spots in the Mycobacterium tuberculosis genome reveals platform-wide and workflow-specific biases. *Microb Genom* **7** https://doi.org/10.1099/mgen.0.000465 | PubMed

40. **Comas I,** et al. (2011) Whole-genome sequencing of rifampicin-resistant Mycobacterium tuberculosis strains identifies compensatory mutations in RNA polymerase genes. *Nat Genet* **44**:106-110 https://doi.org/10.1038/ng.1038 | PubMed

41. **Vargas R,** et al. (2022) Phase variation as a major mechanism of adaptation in Mycobacterium tuberculosis complex. *bioRxiv* https://doi.org/10.1101/2022.06.10.495637

42. **Berman HM,** et al. (2000) The protein data bank. *Nucleic Acids Res* **28**:235-242 https://doi.org/10.1093/nar/28.1.235 | PubMed

43. **Varadi M,** et al. (2022) AlphaFold Protein Structure Database: massively expanding the structural coverage of protein-sequence space with high-accuracy models. *Nucleic Acids Res* **50**:D439-D444 https://doi.org/10.1093/nar/gkab1061 | PubMed

44. **Walker TM,** et al. (2022) The 2021 WHO catalogue of Mycobacterium tuberculosis complex mutations associated with drug resistance: A genotypic analysis. *Lancet Microbe* **3**:e265-e273 https://doi.org/10.1016/s2666-5247(21)00301-3 | PubMed

45. **Getis A,** Ord JK (1992) The Analysis of Spatial Association by Use of Distance Statistics. *Geogr Anal* **24**:189-206 https://doi.org/10.1111/j.1538-4632.1992.tb00261.x

46. **Ord JK,** Getis A (1995) Local Spatial Autocorrelation Statistics: Distributional Issues and an Application. *Geogr Anal* **27**:286-306 https://doi.org/10.1111/j.1538-4632.1995.tb00912.x

47. **Liang S,** Mort M, Stenson PD, Cooper DN, Yu H (2021) PIVOTAL: Prioritizing variants of uncertain significance with spatial genomic patterns in the 3D proteome. *bioRxiv* https://doi.org/10.1101/2020.06.04.135103

48. **Zhao X,** Hersleth H.-P, Zhu J, Andersson KK, Magliozzo RS (2013) Access channel residues Ser315 and Asp137 in Mycobacterium tuberculosis catalase-peroxidase (KatG) control peroxidatic activation of the pro-drug isoniazid. *Chem. Commun* **49**:11650-11652 https://doi.org/10.1039/c3cc47022a | PubMed

49. **Lin W,** et al. (2017) Structural Basis of *Mycobacterium tuberculosis* Transcription and Transcription Inhibition. *Mol. Cell* **66**:169-179.e8 https://doi.org/10.1016/j.molcel.2017.03.001 | PubMed

50. **Petrella S,** et al. (2011) Crystal structure of the pyrazinamidase of Mycobacterium tuberculosis: insights into natural and acquired resistance to pyrazinamide. *PLoS One* **6**:e15785 https://doi.org/10.1371/journal.pone.0015785 | PubMed

51. **Gregory ST,** et al. (2009) Structural and functional studies of the Thermus thermophilus 16S rRNA methyltransferase RsmG. *RNA* **15**:1693-1704 https://doi.org/10.1261/rna.1652709 | PubMed

52. **Echave J,** Spielman SJ, Wilke CO (2016) Causes of evolutionary rate variation among protein sites. *Nat Rev Genet* **17**:109-121 https://doi.org/10.1038/nrg.2015.18 | PubMed

53. **Sun M,** Ge S, Li Z (2022) The Role of Phosphorylation and Acylation in the Regulation of Drug Resistance in Mycobacterium tuberculosis. *Biomedicines* **10**:2592 https://doi.org/10.3390/biomedicines10102592 | PubMed

54. **Zeng J,** et al. (2020) Protein kinases PknA and PknB independently and coordinately regulate essential Mycobacterium tuberculosis physiologies and antimicrobial susceptibility. *PLoS Pathog* **16**:e1008452 https://doi.org/10.1371/journal.ppat.1008452 | PubMed

55. **Sharma K,** et al. (2006) Transcriptional Control of the Mycobacterial embCAB Operon by PknH through a Regulatory Protein, EmbR, In Vivo. *J. Bacteriol* **188**:2936-2944 https://doi.org/10.1128/jb.188.8.2936-2944.2006 | PubMed

56. **Cavazos A,** Prigozhin DM, Alber T (2012) Structure of the Sensor Domain of *Mycobacterium tuberculosis* PknH Receptor Kinase Reveals a Conserved Binding Cleft. *J. Mol. Biol* **422**:488-494 https://doi.org/10.1016/j.jmb.2012.06.011 | PubMed

57. **Farhat MR,** et al. (2013) Genomic analysis identifies targets of convergent positive selection in drug-resistant Mycobacterium tuberculosis. *Nat Genet* **45**:1183-1189 https://doi.org/10.1038/ng.2747 | PubMed

58. **Martini MC,** et al. (2022) Loss of RNase J leads to multi-drug tolerance and accumulation of highly structured mRNA fragments in Mycobacterium tuberculosis. *PLoS Pathog* **18**:e1010705 https://doi.org/10.1371/journal.ppat.1010705 | PubMed

59. **Bao L,** et al. (2023) Structural insights into RNase J that plays an essential role in Mycobacterium tuberculosis RNA metabolism. *Nat. Commun* **14**:2280 https://doi.org/10.1038/s41467-023-38045-z | PubMed

60. **Li H** (2013) Aligning sequence reads, clone sequences and assembly contigs with BWA-MEM. *arXiv* https://doi.org/10.48550/arxiv.1303.3997

61. **Li H,** Durbin R (2009) Fast and accurate short read alignment with Burrows–Wheeler transform. *Bioinformatics* **25**:1754-1760 https://doi.org/10.1093/bioinformatics/btp324 | PubMed

62. **Schmieder R,** Edwards R (2011) Quality control and preprocessing of metagenomic datasets. *Bioinformatics* **27**:863-864 https://doi.org/10.1093/bioinformatics/btr026 | PubMed

63. **Wood DE,** Salzberg SL (2014) Kraken: ultrafast metagenomic sequence classification using exact alignments. *Genome Biol* **15**:R46 https://doi.org/10.1186/gb-2014-15-3-r46 | PubMed

64. **Walker BJ,** et al. (2014) Pilon: an integrated tool for comprehensive microbial variant detection and genome assembly improvement. *PLoS One* **9**:e112963 https://doi.org/10.1371/journal.pone.0112963 | PubMed

65. **Nguyen L.-T,** Schmidt HA, von Haeseler A, Minh BQ (2015) IQ-TREE: A Fast and Effective Stochastic Algorithm for Estimating Maximum-Likelihood Phylogenies. *Molecular biology and evolution* https://doi.org/10.1093/molbev/msu300 | PubMed

66. **Edwards DJ,** Duchene S, Pope B, Holt KE (2021) SNPPar: identifying convergent evolution and other homoplasies from microbial whole-genome alignments. *Microb*. *Genomics* **7**:000694 https://doi.org/10.1099/mgen.0.000694 | PubMed

67. **Hopf TA,** et al. (2019) The EVcouplings Python framework for coevolutionary sequence analysis. *Bioinformatics* **35**:1582-1584 https://doi.org/10.1093/bioinformatics/bty862 | PubMed

68. (no date) HMMER. http://hmmer.org/

69. **Ashburner M,** et al. (2000) Gene ontology: tool for the unification of biology. The Gene Ontology Consortium. *Nat Genet* **25**:25-29 https://doi.org/10.1038/75556 | PubMed

70. **The Gene Ontology Consortium,** et al. (2023) The Gene Ontology knowledgebase in 2023. *Genetics* **224**:iyad031 https://doi.org/10.1093/genetics/iyad031 | PubMed

71. (2023) Catalogue of mutations in Mycobacterium tuberculosis complex and their association with drug resistance, 2nd ed. https://www.who.int/publications/i/item/9789240082410

72. **Pedregosa F,** et al. (2011) Scikit-learn: Machine Learning in Python. *J Mach Learn Res* **12**:2825-2830

73. **Ruopp MD,** Perkins NJ, Whitcomb BW, Schisterman EF (2008) Youden Index and Optimal Cut-Point Estimated from Observations Affected by a Lower Limit of Detection. *Biom J Biom Z* **50**:419-430 https://doi.org/10.1002/bimj.200710415 | PubMed

Peer reviews

Reviewer #1 (Public review):

Summary:

In this manuscript, Green et al. attempt to use large-scale protein structure analysis to find signals of selection and clustering related to antibiotic resistance. This was applied to the whole proteome of Mycobacterium tuberculosis, with a specific focus on the smaller set of known antibiotic-resistance-related proteins.

Strengths:

The use of geospatial analysis to detect signals of selection and clustering on the structural level is really intriguing. This could have a wider use beyond the AMR-focussed work here and could be applied to a more general evolutionary analysis context. Much of the strength of this work lies in breaking ground into this structural evolution space, something rarely seen in such pathogen data. Additional further research can be done to build on this foundation, and the work presented here will be important for the field.

The size of the dataset and use of protein structure prediction via AlphaFold, giving such a consistent signal within the dataset, is also of great interest and shows the power of these approaches to allow us to integrate protein structure more confidently into evolution and selection analyses.

Comments on revised version.

All my comments from the previous round of reviews have been addressed.

https://doi.org/10.7554/eLife.109450.2.sa1

Author response:

The following is the authors' response to the original reviews.

> ***eLife Assessment***
>
> *This valuable study leverages a large global dataset of tens of thousands of tuberculosis samples to place recurrent protein-coding mutations into their three-dimensional structural context, offering an expanded view of how antibiotic resistance emerges compared to traditional genetic analyses alone. The strength of evidence is convincing, supported by the scale and breadth of the dataset and the systematic structural analysis, although some of the assumptions made in the the modeling approach are only partially supported. Overall, the work will be of broad interest to researchers studying microbial evolution, antibiotic resistance, and structure-function relationships in pathogens.*

We thank the reviewers and editors for their careful critique of our work. We believe the work has been strengthened by addressing the comments and are delighted to submit a revised version. This version has a detailed discussion of prior literature on structure analysis of antibiotic resistance variants in *Mycobacterium tuberculosis*, more details on dataset origins and data processing to improve reproducibility, better explanations of the evolutionary assumptions underlying the scoring method we developed, and more discussion of proteins that show homoplasic and clustered mutational signals that are not known to confer antibiotic resistance. We include detailed responses to the comments below.

> ***Public Reviews:***
>
> ***Reviewer #1 (Public review):***
>
> *Summary:*
>
> *In this manuscript, Green et al. attempt to use large-scale protein structure analysis to find signals of selection and clustering related to antibiotic resistance. This was applied to the whole proteome of Mycobacterium tuberculosis, with a specific focus on the smaller set of known antibiotic-resistance-related proteins.*
>
> *Strengths:*
>
> *The use of geospatial analysis to detect signals of selection and clustering on the structural level is really intriguing. This could have a wider use beyond the AMR focussed work here and could be applied to a more general evolutionary analysis context. Much of*

> *the strength of this work lies in breaking ground into this structural evolution space, something rarely seen in such pathogen data. Additional further research can be done to build on this foundation, and the work presented here will be important for the field.*
>
> *The size of the dataset and use of protein structure prediction via AlphaFold, giving such a consistent signal within the dataset, is also of great interest and shows the power of these approaches to allow us to integrate protein structure more confidently into evolution and selection analyses.*
>
> *Weaknesses:*
>
> *There are several issues with the evolutionary analysis and assumptions made in the paper, which perhaps overstate the findings, or require refining to take into account other factors that may be at play.*
>
> *(1) The focus on antimicrobial resistance (AMR) throughout the paper contains the findings within that lens. This results in a few different weaknesses:*
>
> *(a) While the large size of the analysis is highlighted in the abstract and elsewhere, in reality, only a few proteins are studied in depth. These are proteins already associated with AMR by many other studies, somewhat retreading old ground and reducing the novelty.*
>
> *(b) Beyond the AMR-associated proteins, the proteome work is of great interest, but only casually interrogated and only in the context of AMR. There appears to be an assumption that all signals of positive selection detected are related to AMR, whereas something like cas10 is part of the CRISPR machinery, a set of proteins often under positive selection, and thus unlikely to be AMR-related.*

We agree that environmental pressures beyond AMR may impose positive selection. In response to the reviewer's comment we have now included more results and a supplementary figure of the findings about Cas10. We have expanded our results about the proteins found with significant clustering that are not known AMR-associated proteins, and clarified in the discussion that we don't believe positive selection is caused only by AMR.

We note that data from homoplasic substitutions in proteins (Figure 1) does indeed support that AMR is the strongest driver of positive selection in Mtb. A challenge is that knowledge about protein function varies in depth by protein, and in Mtb the AMR proteins are among the most well-studied in the proteome. Thus, explanations from literature are most readily available for AMR proteins. Moreover, proteins may exhibit positive selection for more than one reason – for example, we find significant clustering in the proteins GlmM, GlmS, GlmU, and MurA, all of which are involved in amino sugar metabolism, a key component of the cell wall. These proteins could plausibly have a role in AMR via cell wall permeability mechanisms, and could plausibly have a role in adaptation to host environment via the same (or other) mechanisms.

> *(2) The strength of the signal from the structural information and the novelty of the structural incorporation into prediction are perhaps overstated.*
>
> *(a) A drop of 13% in F1 for a gain of 2% in PPV is quite the trade-off. This is not as indicative of a strong predictor that could be used as the abstract claims. While the approach is novel and this is a good finding for a first attempt at such complex analysis, this is perhaps not as significant as the authors claim.*
>
> *(b) In relation to this, there is a lack of situating these findings within the wider research landscape. For instance, the use of structure for predicting resistance has been done, for example, in PncA (https://academic.oup.com/jacamr/article/6/2/dlae037/7630603,*

> *https://www.sciencedirect.com/science/article/pii/S1476927125003664, https://www.nature.com/articles/s41598-020-58635-x) and in RpoB (https://www.nature.com/articles/s41598-020-74648-y). These, and other such works, should be acknowledged as the novelty of this work is perhaps not as stark as the authors present it to be.*

We appreciate this comment and have made efforts to better situate our work in the wider context of structure-based prediction of antibiotic resistance. We have included description of and citation to these works and others in the introduction, results, and discussion. A differentiator between our work and previous is that we have trained a predictor across all proteins in the WHO catalogue of known resistance variants, not limited ourselves to a single protein at a time. We feel this makes the case that structure (and specifically proximity to known resistance-conferring variants) is a universally useful feature for resistance mutation prediction. We have also updated our abstract to reflect the exact performance of our method.

Introduction: “Protein three-dimensional structure has shown utility as an input feature for identifying resistance-conferring variants in known resistance-conferring proteins such as RpoB,[26,27] PncA [28–30], and AtpE.[31]While past work has sought to reannotate parts of the *M. tuberculosis* proteome with computationally predicted protein structures using older structure prediction methods,[32] we can now infer a protein structure for nearly every protein in the proteome using AlphaFold,[33] leading to new works examining the 3D location of mutations in known and suspected resistance-conferring proteins.[34,35]”

> *(3) The authors postulate that neutral AA substitutions would be randomly distributed in the protein structure and thus use random mutations as a negative control to simulate this neutral evolution. However, I am unsure if this is a true negative control for neutral evolution. The vast majority of residues would be under purifying selection, not neutral selection, especially in core proteins like rpoB and gyrA. Therefore, most of these residues would never be mutated in a real-world dataset. Therefore, you are not testing positive selection against neutral selection; you are testing positive against purifying, which will have a much stronger signal. This is likely to, in turn, overestimate the signal of positive selection. This would be better accounted for using a model of neutral evolution, although this is complex and perhaps outside the scope. Still, it needs to be made clear that these negative controls are not representative of neutral evolution.*

The goal of our negative control was to simulate the random accumulation of amino acid substitutions without the effects of selection, which we had originally referred to as “neutral evolution” but is better described as “randomly accumulating substitutions.” We agree that in the absence of antibiotics, essential proteins like RpoB and GyrA are probably under purifying selection, and thus will have depletion of mutations in their hydrophobic cores. We have revised our wording in the results section “A protein-level statistic to test for mutational clustering” to make it clear that our negative control is that of randomly accumulating substitutions. We have included possible extension to more realistic evolutionary scenarios in the discussion.

As a side note, if we were to compare the observed mutation 3-D pattern to a purifying selection model, that may overestimate the signal compared with the randomly accumulating substitution model that we currently present in the manuscript. Purifying selection would tend to result in slower evolutionary rates than positive selection or randomly accumulation substitutions. So, a control based on purifying selection would have to have fewer mutations to account for the same evolutionary time, and this could lead to underestimating clustering.

> *(4) In a similar vein, the use of 15 Å as a cut-off for stating co-localisation feels quite arbitrary. The average radius of a globular protein is about 20 Å, so this could be quite a*

large patch of a protein. I think it may be good to situate the cut-off for a 'single location' within a size estimator of the entire protein, as 15 Å could be a neighbourhood in a large protein, but be the whole protein for smaller ones.

We interrogated the use of 15 Å as a cutoff and found that it is indeed not very stringent, and functions more as a filter to remove the most egregious examples of proteins lacking single-location clustering. We include a new supplementary figure showing the number of significant hits as the cutoff is varied from 2 Å to 40 Å, and summarize these results in the main text. We note that we in fact find a weak negative relationship between protein length and the distance between the top two residues with highest G-score (R^2 = 0.008, b = -3.1806, *p*-value = 0.048), the opposite of what would be expected under a scenario the distance between residues is simply driven by protein size and not a signal for clustering.

> ***Reviewer #2 (Public review):***
>
> *Summary:*
>
> *This is an important study that, for the first time, systematically places the homoplastic genetic variation observed in the coding regions in a large collection of >31,000 M. tuberculosis samples into the protein structural context. This should be much more informative when, e.g. predicting antimicrobial resistance. The authors imaginatively apply the Getis-Ord score, which originated in geographical spatial analysis but has also been used in human disease to demonstrate that missense mutations in M. tuberculosis known to be associated with antimicrobial resistance are clustered in space. That they are able to consider almost all of the proteome using a large dataset of 31,000 M. tuberculosis complex clinical samples, which makes the evidence convincing.*
>
> *Strengths:*
>
> *To my knowledge, this is the first study to place the homoplastic missense mutations from a large clinical dataset into their protein structural context and attempt to look for clustering in space, which could be indicative of a recent evolutionary pressure, such as the use of antibiotics. The field usually only views resistance through the genetic paradigm, so it is delightful to see a structural paradigm being brought to bear, as this should, in theory, be much more informative, as protein structure is much closer to function. In addition, the dataset used is large (>31,000 clinical M. tuberculosis samples), and the authors are able to consider almost all of the ORFs (3,687/3,996) in the M. tuberculosis reference, and hence the analysis is comprehensive.*
>
> *Weaknesses:*
>
> *It is not apparent at the time of this review if the study could be reproduced by other researchers as e.g. whilst the authors state that the raw sequencing files (FASTQ) underpinning the dataset of 31,428 M. tuberculosis isolates can be downloaded the table in the Supplement containing the sample and accession identifiers contains rows that do not contain NCBI accessions e.g. '01R0685' or 'IDR 1600023875' or '1479144813357T181715lib5022nextseqn0035151bp' instead of the expected form e.g. 'SAMEA1016138'. I have searched the NCBI SRA using these terms and got no results, so they cannot be used to download any FASTQ files. There is also no information in the preprint on how the reads were processed (which is a complex process) and the dataset of SNPs subsequently built. One can trace back through the references, but I cannot find anywhere where one can download the SNP dataset, which would permit researchers to reproduce at least the latter stages of the work -- one obvious option would be to make the SNP dataset available. Likewise, the authors have constructed a "M. tuberculosis structureome", which would be very useful for the community but does not appear to be*

> *publicly available. At the time of the review, not all the GitHub repositories were public, so these points may have been rectified when that was corrected.*

We have made a number of changes to improve reproducibility of the manuscript.

First, we have updated Table 1 with additional information to reflect the dataset of origin. While most of the isolates used are available from NCBI (94.5%), the remainder are from other sources. An additional 4.8% are exclusively from PATRIC (now the BV-BRC) and 0.4% are from ENA. Some of the isolates were originally named by their internal identifiers, not their NCBI BioSamples, which has been rectified. Of the three identifiers the reviewer cites, two were originally from Reseq-TB and are now listed with their NCBI BioSamples, and the third, 01R0685, corresponds to one isolate deposited with others in a single BioProject (*https://www.ncbi.nlm.nih.gov/bioproject/?term=PRJEB26000;* individual isolate available here: *https://www.ncbi.nlm.nih.gov/biosample/10125872*). We have clarified in the table metadata that the relevant search identifier.

Second, we have added a new section to our methods about the origins of the Mtb genomic data used in our manuscript, and describing the process of variant calling and SNP dataset construction.

Third, we have provided the data in a zenodo repository along with instructions for using the protein distance map files provided: 10.5281/zenodo.20766453

Lastly, we have ensured that the github is publicly available.

> *The authors correctly point out in the Introduction that supervised methods like GWAS or ML need datasets with matching genetic and phenotypic drug susceptibility data, which are much difficult/expensive to obtain, but don't then close the loop by comparing their results back to such supervised methods. They pick out RnJ as having previously been identified by a GWAS, but it would have provided a useful validation of their method to e.g. demonstrating that X% of the genes they identify were also identified by GWAS/ML studies, and therefore their method can achieve similar results but without having to collect pDST data.*

We agree that this is a compelling possible extension of the work but due to time constraints have chosen not to pursue it for this manuscript.

> *Whilst the authors acknowledge that assuming all sites are equally likely to mutate in their random shuffling procedure is a shortcoming, a bigger weakness is, I suspect, that one should also only consider which amino acids could arise at each codon due to a SNP. Shuffling assumes any amino acid can arise at any codon which is only possible with multiple nucleotide changes, which is possible but highly unlikely.*

Our approach is based on analyzing, in the wild-type protein structure, the 3D location at which mutations occur. In this calculation we do not consider the identity of the amino acid change per se. We have now clarified in the methods that we are not explicitly simulating biochemical change of the wild-type amino acid to any given mutant, rather we are analyzing the wild type amino acid in its structural context. We have added the following text in the methods section: In Computing inter-residue distances, “The EVcouplings Python package was used to compute the distance between wild-type amino acid residues in all protein structures”; in Computing the Getis-Ord score for clustering of homoplastic mutations, “The two values input to the Getis-Ord statistic computation are a per-residue score x, here the per-amino acid homoplasy score, and a weight matrix W that contains the inverse of the inter-residue distances computed from the wildtype amino acids”; and in Preparing GeO score calibration data, “Note that we do not recompute inter-residue distances when simulating mutations in an amino acid, as the distances used as input to GeO score are the wild-type inter-residue distances.” We hope this addresses the reviewer’s concern.

> *Finally, the authors implicitly assume that the mutations do not perturb the structure of the proteins, which is likely to be generally true for essential genes but less likely to be true for non-essential genes. This assumption underpins their entire approach and should be borne in mind when evaluating the results.*

Our approach is based on analyzing, in the wild-type protein structure, the 3D location at which missense mutations occur and are observable in a naturally evolving population. It is true that we have not undertaken an analysis of whether any given mutation does or does not perturb the protein structure in which it occurs. However, location alone is a useful piece of information to analyze, as the location of naturally occurring mutations gives a readout of what types of mutations are allowed to persist under natural selection. Among our findings is that mutations display significant clustering even in non-essential genes, which we address in our discussion, "for proteins where mutations that lead to loss of function are known to cause resistance, such as PncA and RsmG (GidB), it is not necessarily expected to find clustering of mutations. We suspect that the observed clustering is due to mutations in a certain region of the protein being more likely to cause loss of function."

> ***Recommendations for the authors:***
>
> ***Reviewer #1 (Recommendations for the authors):***
>
> *(1) It would be good if a more detailed description of the sequencing dataset origins were presented. The Supplementary Table 1, which is meant to hold these data, points to another paper, which in turn points to another paper which does not detail collection strategies for this data. This needs to be clear so that any bias in collection, which could over-inflate selection signals, can be assessed.*

[From public reviews]

First, we have updated Table 1 with additional information to reflect the dataset of origin. While most of the isolates used are available from the NCBI (94.5%), the remainder are from other sources. An additional 4.8% are exclusively from PATRIC (now the BV-BRC) and 0.4% are from ENA. Some of the isolates were originally named by their internal identifiers, not their NCBI BioSamples, which has been rectified. Of the three identifiers the reviewer cites, two were originally from Reseq-TB and are now listed with their NCBI BioSamples, and the third, 01R0685, corresponds to one isolate deposited with others in a single BioProject (*https://www.ncbi.nlm.nih.gov/bioproject/?term=PRJEB26000;* individual isolate available here: *https://www.ncbi.nlm.nih.gov/biosample/10125872*). We have clarified in the table metadata that the relevant search identifier.

> *(2) Line 351: The exact download date is missing here.*

We have added the exact download date

> *(3) Line 363: Did you mean lowest e-value? Higher would be worse.*

Thank you for catching this, it is indeed the lowest e-value (confusion stemmed from looking at highest negative log e-value).

> ***Reviewer #2 (Recommendations for the authors):***
>
> *(1) The GitHub repo* was not public at the time of review (nor was it listed under user aggreen), so I could not check how reproducible the results are -- please make it public.*

Absolutely, this has been addressed.

> *(2) The authors say "we envision structure being added as an additional feature in future work to predict resistance phenotypes from sequences" - this is not true, as some work*

> *has already been published predicting resistance in MBTC going back to 2019***

We have addressed this with wording changes and citations to the mentioned work (see public responses)

> *(3) Throughout the term 'non-synonymous' is used; that would include premature stop codons. Would 'missense' be more appropriate?*

You're correct in pointing out that the term "non-synonymous" is too general for what we mean in this paper. Our analysis included missense mutations and in-frame indels, but did not include premature stop (nonsense) mutations or frameshift mutations. So, the term missense (alone) is narrower than what we wish to convey. We have made clarifications throughout.

> *(4) The aminoglycosides are an important, albeit less used, class of antibiotics, and mutations arise in the ribosomal genes, e.g. rrs (which hence do not encode protein). They have therefore been excluded for obvious reasons, but it would help a reader from the tuberculosis field if this were acknowledged. Likewise, a reader might wonder why Rv0678 isn't in Figure 1 - I suspect it is because most of the samples were sequenced before the introduction of bedaquiline, but again, it would help if this were explained.*

See response to (5)

> *(5) On a related note, it is not surprising that rpoC appears in Figure 1 due to its role in compensating for the fitness cost that arises when a rifamipicin-resistance mutation occurs in rpoB: obviously not central but a nice "oh yeah that makes sense" point for the reader if it were briefly mentioned.*

These are both great points about the relevance of our results to the Mtb community. We have added an additional paragraph interpreting the results of Figure 1 that mentions the reason for the appearance of RpoC and Cas10, and non-appearance of non-coding genes and genes relevant to resistance to newly introduced and repurposed drugs.

> *(6) How is the "minimum coordinate difference" calculated? I assume all the structures are missing hydrogens as usual, so for two amino acids A and B, is it the smallest distance between any pair of heavy atoms from A and B? That would, I assume, introduce some bias for larger amino acids like Trp, or did you calculate from shared atoms like the backbone C_alpha atoms? That in turn will tend to make the distances a bit larger. It would be useful to know, as you explicitly mention a 1.5 nm threshold.*

We have explained this in a new section of the results, "The EVcouplings Python package was used to compute the distance between amino acid residues in all protein structures [49]. The package calculates the distance between all heavy (non-hydrogen) atoms in residue *i* and residue *j*, then returns the minimum of those distances."

> *(7) Given the reference used (H37Rv) is Lineage 4, one wonders about deeprooted/phylogenetic mutations, but then I suspect this sentence is doing a lot of that heavy-lifting: "We performed ancestral sequence reconstruction to determine the number of independent arisals of each mutation (homoplasy)". For the more general reader, it would be useful to touch on exactly what you mean and the importance of only considering homoplastic mutations.*

We have expanded our explanation in this section to better make the case for the use of homoplastic variants in our analysis, "Because analyzing the frequency of alleles in a population can be biased by oversampling of particular lineages, and by evolutionary recency, we chose to analyze the number of independent arrivals of each mutation (homoplasy) rather than their population-level frequency. This ensures that more recent

evolutionary events are not underrepresented due to lack of time to spread in the population. To accomplish this, we used a previously compiled a dataset of genomes of 31,428 isolates from the Mycobacterium tuberculosis complex (MTBC), with ancestral sequence reconstruction to determine the number of independent arrivals of each mutation (Supplementary Data 1)."

> *(8) Whilst this is true: "Evolution-based approaches are an alternative for finding variants associated with antibiotic resistance without requiring resistance phenotype data", the dataset used to, e.g. build the second edition of the WHO catalogue of resistanceassociated variants has >50,000 samples and therefore is larger than the dataset you have analysed here. The last time I looked, there were >100k M. tuberculosis samples in NCBI, and therefore, to be valid, your approach should really use more samples than are available with WGS and pDST data. I appreciate, however, that this will not be possible for this manuscript, but it is an obvious criticism.*

We appreciate this critique and acknowledge that the number of isolates with both WGS and pDST has increased rapidly in recent years. The first edition of the WHO catalogue (2021) used 38,215 isolates, which increased to over 50k in the second edition. The dataset of homoplastic mutations on which we based this paper was originally published in 2021.

To address your comment, we have softened our assertions in the introduction about the utility of evolution-based approaches, and emphasize instead the different nature of the underlying signal, "Evolution-based approaches are an alternative for finding variants associated with antibiotic resistance by analyzing their mutational frequency and phylogenetic distribution"

> *(9) Minor point, but the second sentence in the Introduction ignores that one can diagnose MDR-TB using phenotypic methods as well as genetic methods.*

We have added an additional citation and mention of laboratory phenotypic methods.

> *(10) There are a few typos: "genic" "G-sore"*

Addressed.

https://doi.org/10.7554/eLife.109450.2.sa0