BEYOND SEQUENCE-ONLY MODELS: LEVERAGING PROTEIN STRUCTURE FOR ANTIBIOTIC RESISTANCE PREDICTION IN SPARSE GENOMIC DATASETS

Mahbuba Tasmin & Anna G. Green
Manning College of Information and Computer Sciences
University of Massachusetts Amherst
140 Governors Dr., Amherst, MA, USA
{mtasmin, annagreen}@umass.edu

ABSTRACT

To combat the rise of antibiotic-resistant *Mycobacterium tuberculosis*, genotype-based diagnosis of resistance is critical, as it could substantially speed time to treatment. However, machine learning efforts at genotype-based resistance prediction are hindered by limited sequence diversity and high redundancy in genomic datasets, complicating model generalization. Here, we use a dataset of *M. tuberculosis* sequences for nine key resistance-associated genes and corresponding resistance phenotypes, performing genotype de-duplication to mitigate the effects of data leakage. This study introduces a Fused Ridge approach that moves beyond sequence-only prediction by introducing protein structure constraints. We compare to baseline Ridge regression and zero-shot mutation effect prediction using ESM-2 embeddings.

Our results show that Fused Ridge achieves the highest mean AUC (0.766), outperforming Ridge regression (0.755) and ESM-2-based log-likelihood ratio scoring (0.603). It also exhibits improved precision and recall in identifying resistance-conferring variants, particularly for genes such as *gyrA* and *rpoB*, likely due to strong association between the 3D location of mutations and resistance. The fusion penalty enforces smoothness in regression coefficients for spatially adjacent residues, embedding biological knowledge into the predictive framework, and improves generalization in sparse and redundant datasets.

1 INTRODUCTION

Tuberculosis (TB) remains a major global health challenge, causing an estimated 1.25 million deaths in 2023 (World Health Organization, 2024). In 2023, an estimated 410,000 individuals developed drug-resistant TB, yet only 43% were diagnosed and initiated on appropriate treatment (World Health Organization, 2024). The increasing prevalence of rifampicin mono-resistance and transmission-driven resistance cases (>90%), as shown by Farhat et al. (2024), highlights the urgent need for accurate resistance prediction models to guide early intervention and surveillance.

Genotype-based resistance prediction offers a promising approach for rapid TB diagnostics but faces critical challenges. Current sequence-based models suffer from data redundancy and limited training diversity, reducing their ability to generalize to real-world multidrug-resistant TB cases (Farhat et al., 2024; Ektefaie et al., 2024). Protein language models (PLMs) such as ESM-2 leverage evolutionary sequences to predict mutation effects in a zero-shot manner, and therefore may hold promise for limited data regimes. (Meier et al., 2021). However,t Jiao et al. (2024) showed that ESM may struggle in contexts where structural organization of mutations in protein 3D structure is important, which is the case for resistance-conferring mutations that disrupt protein active sites.

Inspired by recent approaches showing success integrating protein 3D structure constraints for mutation effect prediction (Cheng et al., 2023; Gao et al., 2023) we propose Fused Ridge Regression, which extends ridge regression with a fusion penalty to integrate protein 3D structural informa-

tion into resistance prediction. We benchmark against the effectiveness of zero-shot prediction of antibiotic resistance using a PLM and a baseline ridge regression approach (Green et al., 2022).

This study demonstrates the power of integrating structural and sequence-based modeling for computational diagnostics for antibiotic resistant tuberculosis, particularly in high-burden settings where novel resistance mutations frequently arise (Cohen et al., 2015).This study makes several key contributions:

- **Prediction in Real-World Dataset:** Our approach addresses a key challenge in global TB diagnostics, and achieves high accuracy despite data scarcity.
- **Limitations of Zero-Shot prediction with ESM:** We demonstrate that ESM-2 embeddings underperform simple supervised models in distinguishing resistant from susceptible *M. tuberculosis* strains, highlighting the utility of simple approaches for data-scarce tasks.
- **Fused Ridge for Sequence-Structure Integration:** We introduce Fused Ridge regression (Figure 1), which leverages 3D structural information to enhance prediction over sequence-based models.

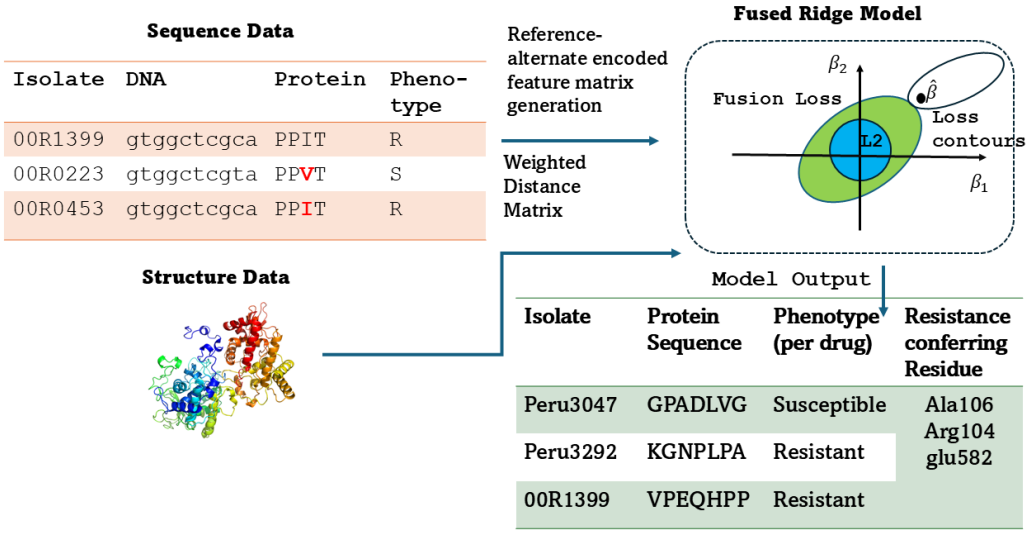

Figure 1: Training a Fused Ridge model for phenotype prediction from unique protein sequences. Schematic of the Fused Ridge model, incorporating an L2 penalty (blue) for shrinkage and a fusion penalty (green) for smoothness across adjacent coefficients. The outer ellipse represents the loss contours of the model's objective function. The model is trained using protein sequences and 3D protein structure.

2 METHODS

2.1 DATASET OVERVIEW

We analyzed nine *M. tuberculosis* genes associated with antibiotic resistance using complete genome sequences aligned to the H37Rv reference genome (Green et al., 2022). Table 1 summarizes the dataset.

2.1.1 SEQUENCE PROCESSING AND DE-DUPLICATION

DNA sequences were translated into protein sequences using Biopython (Cock et al., 2009), with frameshift mutations identified, flagged, and assigned an indicator variable. Reference-alternate encoding was applied relative to the H37Rv reference genome. Each amino acid was assigned 0 (wild-type) or 1 (mutation).

Table 1: Data Summary for *M. tuberculosis* Antibiotic Resistance Genes.

Gene	No. of Unique Sequences	Gene Length (nt)	No. of Variable Positions	Protein Structure Length (aa)
gyrA	439	2516	220	766
embB	681	3296	443	1054
inhA	102	809	65	246
rpsL	13	374	17	122
katG	905	2222	498	716
gid	342	674	205	202
ethA	371	1469	342	482
pncA	257	560	182	185
rpoB	877	3518	453	1138

To mitigate redundancy, we removed duplicate sequences and excluded entries with missing or invalid labels. Stratified split of trains and tests ensured a balanced representation of resistance phenotypes. Further details on translation, frameshift handling, and deduplication are provided in the Appendix A.

2.2 Modeling Approaches

This study evaluates three computational approaches to predict antibiotic resistance phenotypes: field standard ridge regression, zero-shot embeddings from protein language models (ESM-2), and Fused Ridge regression, which integrates 3D structural constraints of proteins. Details on ridge regression, ESM model architecture, and log-likelihood ratio scoring are provided in Appendices B.1 and B.2.

Fused Ridge Regression We extend ridge regression by incorporating structural constraints via a fusion penalty, enforcing smoothness in regression coefficients for spatially adjacent mutations, inspired by previous work on fused lasso regression for adjacent data points (Tibshirani et al., 2004). The objective function extends the standard ridge loss by introducing a fusion penalty:

$$L(\beta) = \frac{1}{2}\sum_{i=1}^{N}\left(y_i - \sum_{j=1}^{p} x_{ij}\beta_j\right)^2 + \alpha\sum_{j=1}^{p}\beta_j^2 + \lambda_{\text{fuse}}\sum_{j=1}^{p-1}\sum_{k=j+1}^{p} w_{jk}(\beta_j - \beta_k)^2$$

where w_{jk} encodes 3D spatial relationships. Residue pairs beyond 5 Å were excluded ($w_{jk} = 0$).The weight matrix construction and alignment strategy are detailed in Appendix A.4. Details of objective formulation and model optimization are included in Appendix C.1 and C.2.

3 Prediction Performance of Ridge, Fused Ridge, and ESM-2

We evaluated Ridge regression, Fused Ridge, and ESM-2 embeddings across nine resistance-associated genes using AUC as the primary metric (Table 2). Fused Ridge achieved the highest mean AUC (0.766), outperforming Ridge (0.755) and ESM-2 (0.603), with notable gains in *pncA* and *inhA*.

We also analyzed the ability of the models to discover known resistance-conferring variants based on the WHO table of resistance-conferring mutations (World Health Organization, 2023a). In most cases, Fused Ridge improved precision and recall for identifying resistance-conferring variants by a small amount over baseline ridge, particularly for genes like *gyrA* and *rpoB*. However, for genes such as *ethA*, Fused Ridge traded recall for precision. Details of evaluation of precision and recall metrics are in Appendix D.2.

Table 2: AUC Comparison for Ridge, Fused Ridge, and ESM-2 Models Across Genes.

Gene	Ridge AUC	Fused Ridge AUC	ESM-2 AUC
embB	0.868	**0.871**	0.686
ethA	**0.572**	0.547	0.540
gid	**0.562**	0.529	0.470
gyrA	0.919	**0.926**	0.742
inhA	0.685	**0.722**	0.678
katG	0.719	**0.727**	0.559
pncA	0.522	**0.625**	0.604
rpoB	**0.950**	0.946	0.645
rpsL	**1.000**	**1.000**	0.500

We further evaluated ESM-2's log-likelihood ratio (LLR) scores across resistant (R) and susceptible (S) strains in *rpoB*, *katG*, and *pncA*. LLR score distributions exhibited substantial overlap (Appendix D.3), highlighting ESM-2's inability to fully capture resistance-driving structural effects.

3.1 Novel Mutation Rates Between Train and Test

To assess the model's ability to generalize to unseen resistance variants, we analyzed the distribution of novel mutations between training and test sets after genotype-level deduplication. This deduplication step produces a dataset of unique isolates and naturally introduces out-of-distribution mutation profiles, serving as an implicit test of generalization. As shown in Figure 4, a substantial fraction of test-time mutations are not observed during training for several genes—e.g., *gyrA* (36.1%), *rpsL* (33.3%), *inhA* (15.4%), and *embB* (14.7%). Even *rpoB*, a well-characterized resistance gene, exhibited 14.3% novel mutations in the test set.

In contrast, genes such as *katG*, *ethA*, and *pncA* showed no novel variants, indicating either a highly recurrent mutational landscape or saturation of common resistance mutations in the dataset. While our current model does not explicitly stratify performance by mutation novelty, the deduplicated data structure inherently enforces generalization to unseen variants.

3.2 Subsampling Analysis

To evaluate model robustness under varying training set sizes, we conducted a subsampling experiment across eight genes. For each gene, we fixed the test set and trained the Fused Ridge model on progressively larger random subsets (20% to 100%) of the training set. Performance was measured via AUC and averaged over random seeds. We excluded the *rpsL* gene due to its small sample size, which made subsampling infeasible [Table1]. Unlike our main experiments result shown in Table 2, this analysis was performed without warm-start approach (initialized with zero coefficients instead of ridge coefficients) to objectively measure the robustness of the Fused Ridge model.

As shown in Table 7, AUCs for genes such as *embB, inhA, ethA* improve with more training data and indicate the effective utilization of additional labels. Conversely, genes like *rpoB, katG* show minimal variance, suggesting that either the available data is already sufficient or further optimization improvements are needed (Figure 5).

3.3 Comparison of Approaches

- **Ridge Regression:** Simple, interpretable, and effective for linear relationships but lacks structural context.
- **ESM-2 Embeddings:** Leverages pre-trained embeddings to infer mutation effects but struggles with domain-specific antibiotic resistance phenotypes.
- **Fused Ridge:** Integrates sequence and structural information, improving generalization and robustness in sparse datasets.

4 Discussion and Conclusion

This study demonstrates that integrating protein structural constraints into predictive models enhances antibiotic resistance classification in *M. tuberculosis*, particularly in data-limited settings dominated by redundant sequences. Our Fused Ridge framework incorporates protein structure priors via fusion penalty, enforcing coefficient smoothness for spatially adjacent residues. Our biologically informed regularization improves generalization despite limited training samples, addressing a key limitation of standard ridge regression and ESM-2 embeddings. This performance improvement is likely because resistance-conferring mutations often cluster in functionally critical protein regions (Green et al., 2022).

ESM-2 embeddings rely on patterns in amino acid sequences and struggle to capture functional complexities influenced by structural context as shown in Jiao et al. (2024), which is critical for distinguishing resistance mechanisms. The limited performance of ESM-2 in distinguishing resistant and susceptible strains for certain genes, such as *katG* and *gid*, suggests that evolutionary embeddings alone may not fully capture mutational impacts driven by drug pressure (Farhat et al., 2024).

Model performance is gene-specific. Longer genes with more genetic variation, such as *rpoB* and *embB* (see Table 1), consistently achieved strong predictive performance across all models. In contrast, genes like *ethA* and *gid*, which have fewer observed variants, showed poorer performance. These observations suggest that both sequence diversity and structural factors may play a role in predictive accuracy, though further analysis is needed to confirm these relationships. Notably, *pncA* demonstrated improved performance with the Fused Ridge model despite being one of the smallest genes with no clear hotspot for mutations (Miotto et al., 2014).

While our analysis focuses on nine key resistance-associated genes, it is important to note that these genes represent the vast majority of clinically significant antibiotic resistance mutations currently characterized in Mycobacterium tuberculosis (Green et al., 2022; World Health Organization, 2023b). The primary data restriction in our study is the need for sequences with labeled phenotype data. Future studies could expand the scope of genes considered.

Future directions include hybrid models that integrate pre-trained evolutionary embeddings (e.g., ESM-2, MSA-Transformer) with structural constraints for improved generalization. Integrating structural insights into computational models can strengthen MDR-TB surveillance and early detection of antibiotic resistance, enabling more effective treatment strategies in this inherently data-limited setting.

Acknowledgments

We thank Professor Benjamin R. Marlin for his invaluable guidance in the development and optimization of the Fused Ridge model. This research was supported in part by NIH/NIAID F32AI161793. This work utilized resources from Unity, a collaborative, multi-institutional high-performance computing cluster managed by UMass Amherst Research Computing and Data.

References

Jun Cheng, Guido Novati, Joshua Pan, Clare Bycroft, Akvilė Žemgulytė, Taylor Applebaum, Alexander Pritzel, Lai Hong Wong, Michal Zielinski, Tobias Sargeant, Rosalia G Schneider, Andrew W Senior, John Jumper, Demis Hassabis, Pushmeet Kohli, and Žiga Avsec. Accurate proteome-wide missense variant effect prediction with AlphaMissense. *Science*, 381(6664): eadg7492, September 2023.

Peter J. A. Cock, Tiago Antao, Jeffrey T. Chang, Brad A. Chapman, Cymon J. Cox, Andrew Dalke, Iddo Friedberg, Thomas Hamelryck, Frank Kauff, Bartek Wilczynski, and Michiel J. L. de Hoon. Biopython: freely available python tools for computational molecular biology and bioinformatics. *Bioinformatics*, 25(11):1422–1423, 03 2009. ISSN 1367-4803. doi: 10.1093/bioinformatics/btp163. URL https://doi.org/10.1093/bioinformatics/btp163.

Keira A Cohen, Thomas Abeel, Abigail Manson McGuire, Christopher A Desjardins, Vanisha Munsamy, Terrance P Shea, Bruce J Walker, Nonkqubela Bantubani, Deepak V Almeida, Lucia Alvarado, Sinéad B Chapman, Nomonde R Mvelase, Eamon Y Duffy, Michael G Fitzgerald, Pamla

Govender, Sharvari Gujja, Susanna Hamilton, Clinton Howarth, Jeffrey D Larimer, Kashmeel Maharaj, Matthew D Pearson, Margaret E Priest, Qiandong Zeng, Nesri Padayatchi, Jacques Grosset, Sarah K Young, Jennifer Wortman, Koleka P Mlisana, Max R O'Donnell, Bruce W Birren, William R Bishai, Alexander S Pym, and Ashlee M Earl. Evolution of extensively drug-resistant tuberculosis over four decades: Whole genome sequencing and dating analysis of mycobacterium tuberculosis isolates from KwaZulu-Natal. *PLoS Med.*, 12(9):e1001880, September 2015.

Yasha Ektefaie, Andrew Shen, Daria Bykova, Maximillian G Marin, Marinka Zitnik, and Maha Farhat. Evaluating generalizability of artificial intelligence models for molecular datasets. *Nat. Mach. Intell.*, 6(12):1512–1524, December 2024.

Maha Farhat, Helen Cox, Marwan Ghanem, Claudia M. Denkinger, Camilla Rodrigues, Mirna S. Abd El Aziz, Handaa Enkh-Amgalan, Debrah Vambe, Cesar Ugarte-Gil, Jennifer Furin, and Madhukar Pai. Drug-resistant tuberculosis: a persistent global health concern. *Nature Reviews Microbiology*, 22:617–635, 2024. URL https://doi.org/10.1038/s41579-024-01025-1.

Hong Gao, Tobias Hamp, Jeffrey Ede, Joshua G Schraiber, Jeremy McRae, Moriel Singer-Berk, Yanshen Yang, Anastasia S D Dietrich, Petko P Fiziev, Lukas F K Kuderna, Laksshman Sundaram, Yibing Wu, Aashish Adhikari, Yair Field, Chen Chen, Serafim Batzoglou, Francois Aguet, Gabrielle Lemire, Rebecca Reimers, Daniel Balick, Mareike C Janiak, Martin Kuhlwilm, Joseph D Orkin, Shivakumara Manu, Alejandro Valenzuela, Juraj Bergman, Marjolaine Rousselle, Felipe Ennes Silva, Lidia Agueda, Julie Blanc, Marta Gut, Dorien de Vries, Ian Goodhead, R Alan Harris, Muthuswamy Raveendran, Axel Jensen, Idriss S Chuma, Julie E Horvath, Christina Hvilsom, David Juan, Peter Frandsen, Fabiano R de Melo, Fabrício Bertuol, Hazel Byrne, Iracilda Sampaio, Izeni Farias, João Valsecchi do Amaral, Mariluce Messias, Maria N F da Silva, Mihir Trivedi, Rogerio Rossi, Tomas Hrbek, Nicole Andriaholinirina, Clément J Rabarivola, Alphonse Zaramody, Clifford J Jolly, Jane Phillips-Conroy, Gregory Wilkerson, Christian Abee, Joe H Simmons, Eduardo Fernandez-Duque, Sree Kanthaswamy, Fekadu Shiferaw, Dongdong Wu, Long Zhou, Yong Shao, Guojie Zhang, Julius D Keyyu, Sascha Knauf, Minh D Le, Esther Lizano, Stefan Merker, Arcadi Navarro, Thomas Bataillon, Tilo Nadler, Chiea Chuen Khor, Jessica Lee, Patrick Tan, Weng Khong Lim, Andrew C Kitchener, Dietmar Zinner, Ivo Gut, Amanda Melin, Katerina Guschanski, Mikkel Heide Schierup, Robin M D Beck, Govindhaswamy Umapathy, Christian Roos, Jean P Boubli, Monkol Lek, Shamil Sunyaev, Anne O'Donnell-Luria, Heidi L Rehm, Jinbo Xu, Jeffrey Rogers, Tomas Marques-Bonet, and Kyle Kai-How Farh. The landscape of tolerated genetic variation in humans and primates. *Science*, 380 (6648):eabn8153, June 2023.

Laura C. Gomes, Susana Campino, Cláudio R. F. Marinho, Taane G. Clark, and Jody E. Phelan. Whole genome sequencing reveals large deletions and other loss of function mutations in mycobacterium tuberculosis drug resistance genes. *Microbial Genomics*, 7(12):000724, December 2021. ISSN 2057-5858. doi: 10.1099/mgen.0.000724. URL https://doi.org/10.1099/mgen.0.000724. Research Support, Non-U.S. Gov't.

Anna G Green, Chang Ho Yoon, Michael L Chen, Yasha Ektefaie, Mack Fina, Luca Freschi, Matthias I Gröschel, Isaac Kohane, Andrew Beam, and Maha Farhat. A convolutional neural network highlights mutations relevant to antimicrobial resistance in mycobacterium tuberculosis. *Nat. Commun.*, 13(1):3817, July 2022.

Thomas A Hopf, Anna G Green, Benjamin Schubert, Sophia Mersmann, Charlotta P I Schärfe, John B Ingraham, Agnes Toth-Petroczy, Kelly Brock, Adam J Riesselman, Perry Palmedo, Chan Kang, Robert Sheridan, Eli J Draizen, Christian Dallago, Chris Sander, and Debora S Marks. The evcouplings python framework for coevolutionary sequence analysis. *Bioinformatics*, 35(9): 1582–1584, 10 2018. ISSN 1367-4803. doi: 10.1093/bioinformatics/bty862. URL https://doi.org/10.1093/bioinformatics/bty862.

Shujian Jiao, Bingxuan Li, Lei Wang, Xiaojin Zhang, Wei Chen, Jiajie Peng, and Zhongyu Wei. Beyond esm2: Graph-enhanced protein sequence modeling with efficient clustering. 2024. URL https://arxiv.org/abs/2404.15805.

Zeming Lin, Halil Akin, Roshan Rao, Brian Hie, Zhongkai Zhu, Wenting Lu, Nikita Smetanin, Robert Verkuil, Ori Kabeli, Yaniv Shmueli, Allan Dos Santos Costa, Maryam Fazel-Zarandi,

Tom Sercu, Salvatore Candido, and Alexander Rives. Evolutionary-scale prediction of atomic-level protein structure with a language model. *Science*, 379(6637):1123–1130, March 2023. doi: 10.1126/science.ade2574. URL https://www.science.org/doi/abs/10.1126/science.ade2574.

Joshua Meier, Roshan Rao, Robert Verkuil, Jason Liu, Tom Sercu, and Alexander Rives. Language models enable zero-shot prediction of the effects of mutations on protein function. In *Proceedings of the 35th International Conference on Neural Information Processing Systems*, NIPS '21, Red Hook, NY, USA, 2021. Curran Associates Inc. ISBN 9781713845393.

Paolo Miotto, Andrea M Cabibbe, Silke Feuerriegel, Nicola Casali, Francis Drobniewski, Yulia Rodionova, Daiva Bakonyte, Petras Stakenas, Edita Pimkina, Ewa Augustynowicz-Kopeć, Massimo Degano, Alessandro Ambrosi, Sven Hoffner, Mikael Mansjö, Jim Werngren, Sabine Rüsch-Gerdes, Stefan Niemann, and Daniela M Cirillo. Mycobacterium tuberculosis pyrazinamide resistance determinants: a multicenter study. *MBio*, 5(5):e01819–14, October 2014.

Fabian Pedregosa, Gaël Varoquaux, Alexandre Gramfort, Vincent Michel, Bertrand Thirion, Olivier Grisel, Mathieu Blondel, Peter Prettenhofer, Ron Weiss, Vincent Dubourg, Jake Vanderplas, Alexandre Passos, David Cournapeau, Matthieu Brucher, Matthieu Perrot, and Édouard Duchesnay. Scikit-learn: Machine learning in python. *J. Mach. Learn. Res.*, 12(null):2825–2830, November 2011. ISSN 1532-4435.

Dawei Shi, Qiulong Zhou, Sihong Xu, Yumei Zhu, Hui Li, and Ye Xu. Pyrazinamide resistance and pnca mutation profiles in multidrug resistant mycobacterium tuberculosis. *Infection and Drug Resistance*, 15:4985–4994, 2022. doi: 10.2147/IDR.S368444. URL https://doi.org/10.2147/IDR.S368444. Epub 2022 Aug 30.

Robert Tibshirani, Michael Saunders, Saharon Rosset, Ji Zhu, and Keith Knight. Sparsity and smoothness via the fused lasso. *Journal of the Royal Statistical Society Series B: Statistical Methodology*, 67(1):91–108, 12 2004. ISSN 1369-7412. doi: 10.1111/j.1467-9868.2005.00490.x. URL https://doi.org/10.1111/j.1467-9868.2005.00490.x.

World Health Organization. Catalogue of mutations in mycobacterium tuberculosis complex and their association with drug resistance, second edition. 2023a. URL https://www.who.int/publications/i/item/9789240082410. Licence: CC BY-NC-SA 3.0 IGO.

World Health Organization. *Global Tuberculosis Report 2023*. World Health Organization, Geneva, 2023b. URL https://www.who.int/publications/i/item/9789240076374. Licence: CC BY-NC-SA 3.0 IGO.

World Health Organization. Global tuberculosis report 2024. 2024. URL https://www.who.int/publications/i/item/9789240101531. Licence: CC BY-NC-SA 3.0 IGO.

A Data Preprocessing and Structural Integration

A.1 Data Sources

Proteins sequences were matched with homologous 3D structures using the default sequence alignemnt pipeline from EVcouplings (Hopf et al., 2018). For proteins that lacked a structural hit with at least 90% coverage of the query sequence, we use the AlphaFold structure, as indicated in Table 3.

A.2 Translation and Frameshift Handling

DNA sequences were translated into protein sequences using Biopython (Cock et al., 2009). Frameshift mutations were identified by detecting indels that disrupted codon alignment. Translation proceeded according to the following rules:

- **Gap Handling**: '-' symbols, representing gaps, were retained in codons.

Table 3: Input Protein, Structure, and Phenotype Label Data

Gene	Uniprot ID	Locus ID	Associated Antibiotic	PDB/AlphaFold ID (Chain)
embB	P9WNL7	Rv3795	Ethambutol	7BVF (B)
ethA	P9WNF9	Rv3854c	Ethionamide	AF-P9WNF9-F1
gid	P9WGW9	Rv3919c	Streptomycin	3G8A (A)
gyrA	P9WG47	Rv0006	Levofloxacin	AF-P9WG47-F1
inhA	P9WGR1	Rv1484	Isoniazid	2B36 (C)
katG	P9WIE5	Rv1908c	Isoniazid	4C51 (A)
pncA	I6XD65	Rv2043c	Pyrazinamide	3PL1 (A)
rpoB	P9WGY9	Rv0667	Rifampicin	5UH6 (C)
rpsL	P9WH63	Rv0682	Streptomycin	5MMJ (I)

- **Ambiguous Bases**: Ambiguous nucleotides represented by 'N' were replaced with 'X' to indicate uncertainty in translated amino acid.
- **Frameshift Detection**: Frameshifted sequences were identified and assigned an all-zero encoding with a separate frameshift flag. For example:
  - **Non-frameshift**: `GTTACTGTATTC` → `VTVF` (frameshift flag = 0).
  - **Frameshift**: `GTTACGTATTC` → `VTY-` (frameshift flag = 1).

A.3 De-duplication and Data Cleaning

To prevent bias from redundant sequences, we removed identical sequences. Entries with missing or invalid resistance labels were excluded. Stratified data splitting ensured balanced distribution of resistant and susceptible strains.

A.4 Weight Matrix for Structural Constraints

To integrate 3D structural information, we constructed a weight matrix based on pairwise Euclidean distances between residues from the protein data bank (PDB) structures. We aligned the H37Rv reference sequence to residue indices from the PDB. We computed pairwise minimum atom inter-residue atomic distances using EVcouplings (Hopf et al., 2018). We then applied a Gaussian decay function, with residue pairs beyond 5 Å excluded ($w_{jk} = 0$):

$$w_{jk} = \exp\left(-\frac{D_{jk}}{\text{scale\_param}}\right).$$

B Modeling Approaches

B.1 Ridge Regression

Ridge regression was implemented using the scikit-learn library (Pedregosa et al., 2011), with a binary-encoded feature matrix comparing mutations to the H37Rv reference genome.

B.2 ESM Model Details

The ESM2-150M model was used for zero-shot resistance prediction using the following steps. First, protein sequences were tokenized and mutations were identified by aligning sequences to H37Rv. To calculate the Log-Likelihood Ratio (LLR), wild-type residues were masked in the input sequence, and probability scores were computed as per (Meier et al., 2021):

$$\log p(x_i = x_i^{\text{mt}} \mid x_{-M}) - \log p(x_i = x_i^{\text{wt}} \mid x_{-M}).$$

We then trained logistic regression model was trained on log-likelihood ratio (LLR) scores to classify mutations as resistance-associated or not.

C Fused Ridge Model

C.1 Objective Function

The Fused Ridge objective function extends the standard ridge loss by introducing a fusion penalty:

$$L(\beta) = \frac{1}{2}\sum_{i=1}^{N}\left(y_i - \sum_{j=1}^{p} x_{ij}\beta_j\right)^2 + \alpha\sum_{j=1}^{p}\beta_j^2 + \lambda_{\text{fuse}}\sum_{j=1}^{p-1}\sum_{k=j+1}^{p} w_{jk}(\beta_j - \beta_k)^2$$

Where:

- λ_{fuse}: Regularization parameter for the fusion penalty.
- w_{jk}: Weights derived from the 3D distance matrix, penalizing the squared difference between coefficients β_j and β_k.
- $w_{jk} = \exp\left(-\frac{D_{jk}}{\text{scale\_param}}\right)$: Weight decay based on the 3D distance D_{jk} between residues j and k.

Table 4 describes the components of the objective function.

Table 4: Components of the Objective Function

Term	Description
$L(\beta)$	Objective function, representing the total loss to be minimized
N	Number of observations (samples) in the dataset
p	Number of features (mutation sites) in the model
y_i	Binary phenotype (resistant or susceptible) for the i-th sample
x_{ij}	Feature encoding the presence or absence of mutation j in sample i
β_j	Regression coefficient for mutation j, representing its contribution to resistance
α	Regularization strength for the L2 penalty (ridge term) to control overfitting
λ_{fuse}	Fusion penalty parameter enforcing smoothness between coefficients of structurally adjacent mutations
w_{jk}	Weight derived from the 3D structural distance matrix, penalizing the squared difference between regression coefficients β_j and β_k
D_{jk}	Euclidean distance between residues j and k in protein 3D structure
scale_param	Normalization factor ensuring proper scaling of w_{jk}, set as $\max(1, \text{std}(D_{jk}))$

C.2 Implementation Details

Here, we detail the optimization and training procedure for the fused-ridge model. To initialize β, we adopted a **warm-start** approach by using the coefficients from a pre-trained baseline ridge regression model. This strategy not only provided a strong initialization but also accelerated convergence by leveraging the stability of ridge regression.

We explored four variants of gradient descent (GD) to identify the most effective optimization method: vanilla GD, enhanced GD with gradient clipping, momentum-based GD, and Nesterov-accelerated GD. Among these, **enhanced GD** consistently yielded the best performance. We conducted a grid search with 5-fold cross-validation to identify the optimal values of key hyperparameters: α, λ_{fuse}, learning rate, and scale_param.

D Performance Evaluation

D.1 Ranked AUC Performance

Figure 2 presents a ranked comparison of AUC scores across genes. Fused Ridge demonstrates on par or superior classification ability in most genes (six out of nine).

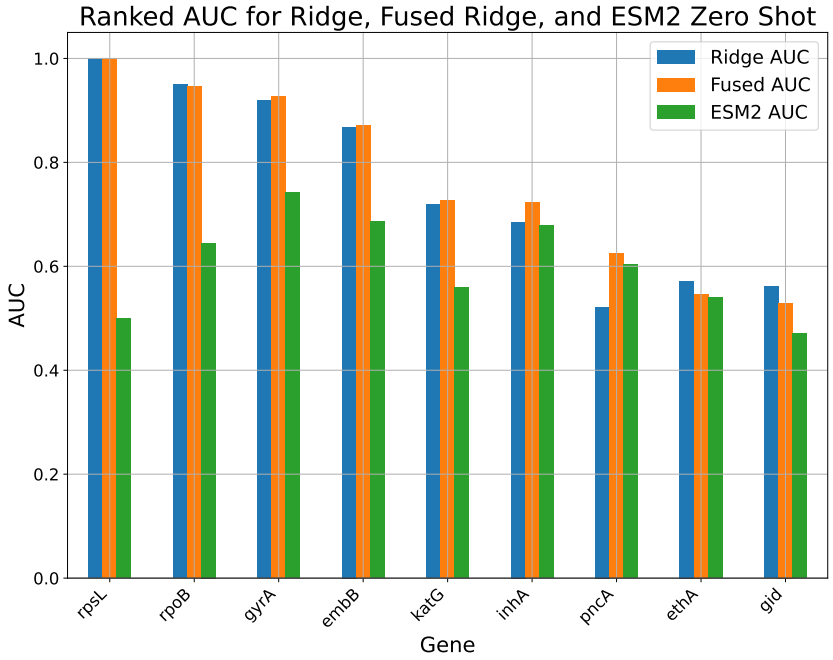

Figure 2: Ranked AUC scores for the three models across nine genes.

D.1.1 Key Observations from AUC Performance

Fused Ridge outperformed Ridge on **66.67%** of the genes. For example, in *pncA*, Fused Ridge achieved an AUC of 0.625, outperforming both Ridge (0.522) and ESM-2 (0.604). The gene *pncA* is displays resistance-conferring mutations across its entire coding region (Shi et al., 2022), because mutations that ablate the protein's function are sufficient to cause resistance. This dispersed mutational pattern makes it challenging for models to generalize from limited training data. As evident here, the sequence-only Ridge model fails to capture meaningful signal, while ESM-2—pretrained on large-scale protein datasets with implicit structural information—achieves better performance. Our Fused Ridge model, which explicitly incorporates protein structural priors, outperforms both, supporting our hypothesis that structure-aware regularization can enhance generalization in genes with diffused mutation profiles.

In *inhA*, the Fused Ridge model demonstrated a notable improvement (0.722) over Ridge (0.685) and ESM-2 (0.678). Marginal gains over the baseline Ridge were observed in genes like *gyrA* (Fused Ridge: 0.926 vs. Ridge: 0.919), though Fused Ridge still showed a substantial advantage over ESM-2 (0.742).

All three models struggled to achieve high AUCs in complex genes such as *ethA* and *gid*. For instance, in *ethA*, baseline Ridge outperformed the other two models (Ridge: 0.572, Fused Ridge: 0.547, ESM-2: 0.540). Genes like *ethA* and *gid*, which are associated with ethionamide and streptomycin resistance respectively, are frequently affected by loss-of-function (LOF) mutations, including frameshifts, premature stop codons, and large deletions (Gomes et al., 2021). These mutations can lead to resistance even in the absence of conventional SNPs, but they are poorly captured by models that rely solely on amino acid substitutions or per-residue embeddings. The presence of these cryptic resistance mechanisms likely contributes to the consistently low AUCs observed across all models and highlights the need for more comprehensive mutation modeling in genes prone to LOF-driven resistance. A second confounder for protein-based resistance prediction is that multiple genetic mechanisms, not just the two genes in question, can confer both streptomycin and ethionamide resistance.

D.2 Precision and Recall of Identifying Resistance-conferring Variants

Another important aspect of this study is to identify the true resistance-conferring mutations from significant coefficients of the trained model and evaluate how well these selected features correspond to known resistance-conferring variants in *M. tuberculosis* World Health Organization (2023a). To quantify the model's ability to retrieve true resistance-conferring mutations, we define two key evaluation metrics: **Precision** and **Recall**.

Precision is defined as the proportion of correctly identified resistance-conferring mutations among all mutations selected by the model. A higher precision indicates fewer false discoveries, meaning the model selects mutations that are more likely to be truly resistance-conferring.

Recall measures the model's ability to identify all known resistance-conferring mutations. A higher recall indicates that the model successfully retrieves a larger proportion of true resistance-conferring mutations.

The variant discovery analysis begins by examining the coefficients generated by the trained model to assess the importance of various features. A threshold was selected by iteratively evaluating percentiles of the absolute model coefficients and choosing the one that maximized the F1-score, i.e., the harmonic mean of precision and recall. Features with coefficients above this threshold are considered significant and are selected for further analysis.

Once the significant features are identified, they are ranked in descending order of importance based on the magnitude of their corresponding coefficients, which reflects their relative contribution to the model's predictions, and are then assessed for alignment with known resistance-conferring variants in the WHO catalog (World Health Organization, 2023b). Table 5 summarizes the precision and recall values for Ridge and Fused Ridge optimizers using the Enhanced optimizer.

Table 5: Precision and Recall Scores for Ridge and Fused Ridge using the Enhanced Optimizer.

Gene	Ridge Precision (%)	Fused Precision (%)	Ridge Recall (%)	Fused Recall (%)
embB	87.76	84.49	50.00	50.00
ethA	75.82	75.82	77.78	44.44
gid	90.27	91.15	36.36	36.36
gyrA	67.46	76.19	42.86	64.29
inhA	87.10	87.10	100.00	100.00
katG	79.01	79.88	25.00	25.00
pncA	81.32	80.68	56.12	56.12
rpoB	82.69	84.13	56.25	59.38
rpsL	70.97	81.25	50.00	100.00

Fused Ridge demonstrates higher precision compared to ridge for *gyrA* and *rpsL*. For genes like *gid*, *embB*, and *pncA*, precision is similar between the two methods. Fused Ridge also shows improvement in recall for *gyrA*, *rpsL*, and *rpoB*.

D.3 ESM Zero-Shot Performance

D.3.1 Evaluation of ESM-2 for Antibiotic Resistance Prediction

ESM-2 embeddings, derived from transformer-based protein language models, provide zero-shot mutation effect predictions based on evolutionary context (Lin et al., 2023). Here, we further explore their ability to distinguish resistant (**R**) and susceptible (**S**) *M. tuberculosis* strains across nine antibiotic resistance genes.

D.3.2 Performance Summary

The violin plots in Figure 3 illustrate the normalized log-likelihood ratio (LLR) score distributions for **R** and **S** phenotypes in three representative genes: *rpoB*, *katG*, and *pncA*.

- ***RpoB* (Rifampicin Resistance):** The LLR score distribution for *rpoB* shows a slight difference between resistant and susceptible strains (Figure 3). That leads to an AUC of **0.645**.

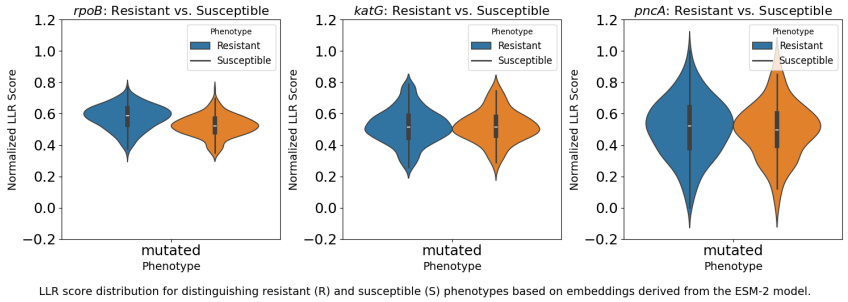

Figure 3: Normalized LLR score distributions for resistant (R) and susceptible (S) strains for *rpoB*, *katG*, and *pncA*.

Despite its critical role in rifampicin resistance, ESM-2 embeddings demonstrate limited discriminatory power in distinguishing resistant and susceptible strains.

- ***KatG* (Isoniazid Resistance):** For *katG*, responsible for isoniazid resistance, ESM-2's AUC is **0.559**, indicating limited performance in distinguishing resistant and susceptible strains. Figure 3 shows significant overlap in LLR score distributions, with close median values for resistant strains.
- ***PncA* (Pyrazinamide Resistance):** The AUC for *pncA* is **0.604**, which shows that ESM-2 is relatively performing better for distinguishing pyrazinamide resistance compared to *katG*. The violin plot in Figure 3 shows a wider range of LLR scores for resistant strains. This suggests that certain mutations in *pncA* are strongly associated with resistance.

While ESM-2 provides a baseline for zero-shot resistance prediction, it is outperformed by even simple supervised models. Since ESM-2 embeddings are trained on evolutionary sequence data, they may not effectively distinguish mutations driven by recent pressure for antibiotic resistance, rather than long-term selection. Additional fine-tuning may improve the performance of ESM-2 for this task ((Ektefaie et al., 2024)).

E Generalization to Novel Mutations

To evaluate whether our model encounters novel resistance mutations at test time, we conducted a per-gene analysis of mutation overlap between training and test sets after genotype deduplication. The results show that a substantial proportion of test-set mutations are indeed unseen during training, particularly in genes such as gyrA (36.1%), rpsL (33.3%), and inhA (15.4%).

These findings confirm that our deduplication strategy produces out-of-distribution mutations in the evaluation split. While we do not explicitly stratify performance by mutation novelty in this version, our results indicate that the model is already being tested under generalization conditions. In future work, we plan to explore performance stratified by mutation novelty, and to augment training with synthetic or rare variants for enhanced generalization.

F Subsampling AUC vs. Train Fraction

To assess the impact of training set size on model robustness, we conducted a subsampling experiment across eight genes. For each gene, we fixed the test set and trained the Fused Ridge model on increasingly larger random subsets (20% to 100%) of the training set. Performance was evaluated using AUC and averaged over five random seeds.

Figure 5 shows the subsampling analysis performance of the fused-ridge model without ridge initialization (i.e., warm start) to isolate the intrinsic data efficiency of the model. Instead, we initialized the Fused Ridge coefficients to zero for each run to avoid confounding the model's sensitivity to training size with the influence of a potentially well-informed initialization. We have excluded the *rpsL* gene from this experiment due to its small number of training samples, which made meaningful subsampling infeasible.

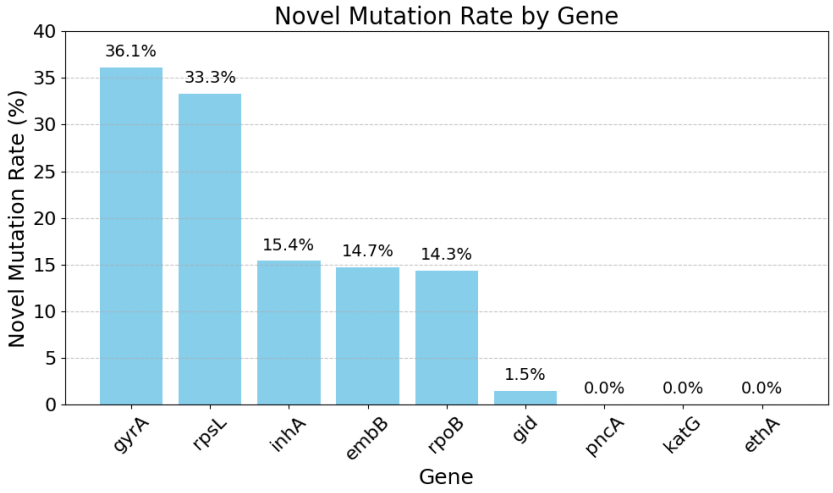

Figure 4: Novel Mutation Distribution Per Gene between Train and Test set

Table 6: Per-gene analysis of novel mutations in the test set after genotype deduplication. Novel mutations constitute a substantial portion of test mutations for several genes. Thereby, the model is inherently exposed to out-of-distribution variants.

Gene	**Train Samples**	**Test Samples**	**Test Mutations**	**Novel Mutations**	**Shared Mutations**	**Novel %**
inhA	82	21	65	10	54	15.38
katG	724	181	424	0	424	0.00
pncA	205	52	167	0	167	0.00
rpsL	10	3	3	1	2	33.33
embB	545	137	443	65	377	14.67
gid	273	69	205	3	201	1.46
ethA	296	75	325	0	325	0.00
gyrA	351	88	83	30	53	36.14
rpoB	702	176	453	65	387	14.35

As Table 7 shows, genes like *embB*, *inhA*, and *ethA* show increasing AUC trends, suggesting that the Fused Ridge model benefits from additional data. In contrast, performance for genes like *rpoB* and *katG* remains relatively stable, indicating either sufficient training coverage or limited scalability due to optimization limitations or data complexity.

To further contextualize these findings, we compare them with a warm-start variant of the model in Table 8. As expected, warm-started training improves AUCs across all genes due to better initial coefficient estimates. However, the trends with respect to training set size remain consistent, confirming that the Fused Ridge model benefits from careful initialization. These results confirm that while warm-starting can enhance performance, our main conclusions about data-driven variability in model robustness are valid even without it.

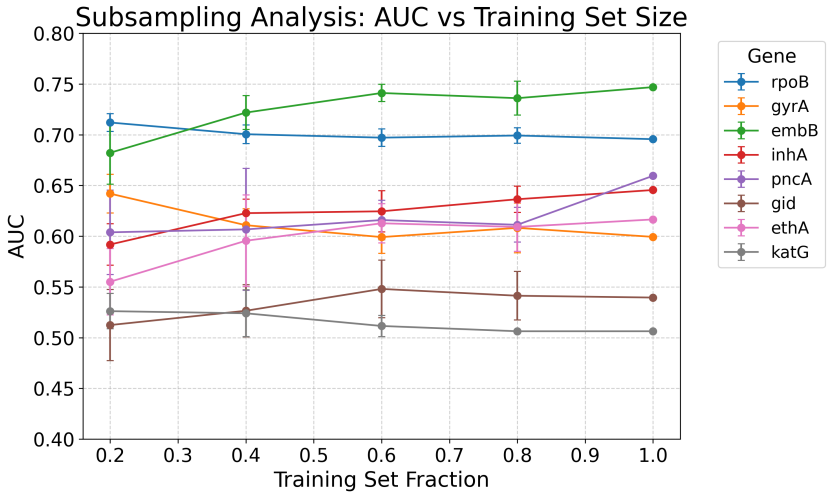

Figure 5: Subsampling Analysis:Impact of Training set size on Model Robustness

Table 7: AUC scores from subsampling experiments across training fractions (without warm start).

Gene	**0.2**	**0.4**	**0.6**	**0.8**	**1.0**	**Average AUC**
rpoB	0.712	0.701	0.697	0.699	0.696	0.701
gyrA	0.642	0.611	0.599	0.608	0.599	0.612
embB	0.682	0.722	0.741	0.736	0.747	0.726
inhA	0.592	0.623	0.625	0.636	0.646	0.624
ethA	0.555	0.596	0.613	0.609	0.616	0.598
gid	0.512	0.527	0.548	0.541	0.540	0.534
pncA	0.604	0.607	0.616	0.611	0.660	0.619
katG	0.526	0.524	0.512	0.506	0.506	0.515

Table 8: AUC scores from subsampling experiments with ridge initialization (warm start).

Gene	**0.2**	**0.4**	**0.6**	**0.8**	**1.0**	**Average AUC**
rpoB	0.931	0.932	0.929	0.934	0.934	0.932
gyrA	0.863	0.893	0.892	0.888	0.913	0.890
embB	0.786	0.829	0.844	0.843	0.852	0.831
inhA	0.552	0.601	0.719	0.694	0.659	0.645
ethA	0.563	0.575	0.561	0.579	0.571	0.570
gid	0.500	0.511	0.560	0.521	0.531	0.524
pncA	0.594	0.595	0.591	0.594	0.629	0.601
katG	0.702	0.703	0.690	0.720	0.719	0.707