BEYOND SEQUENCE-ONLY MODELS: LEVERAGING STRUCTURAL CONSTRAINTS FOR ANTIBIOTIC RESISTANCE PREDICTION IN SPARSE GENOMIC DATASETS

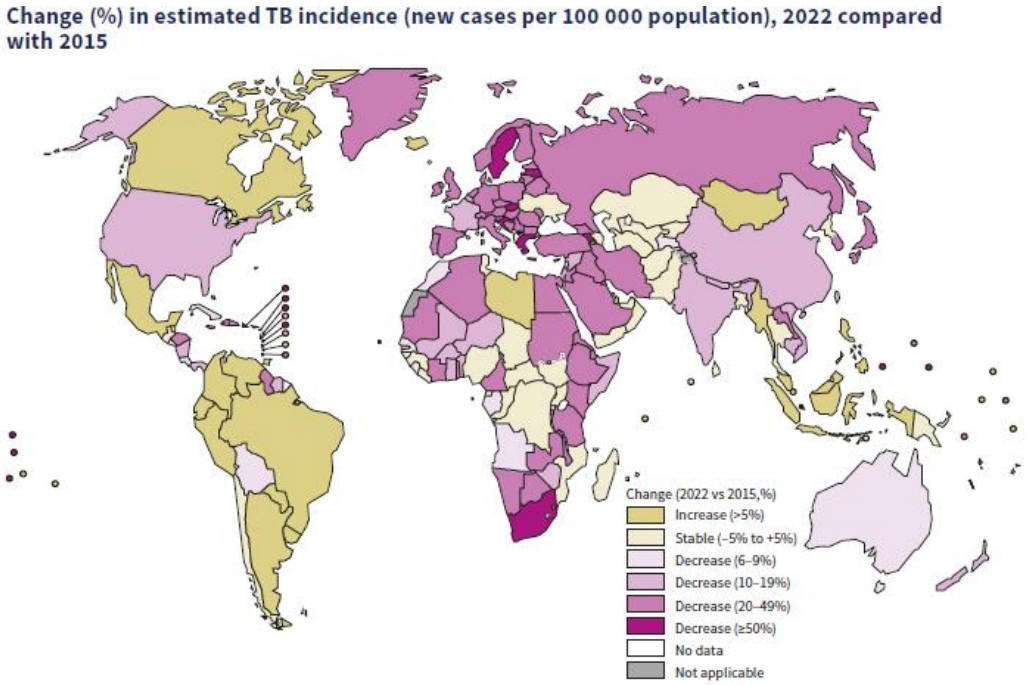
Mahbuba Tasmin , Anna G. Green
Manning College of Information & Computer Sciences, Amherst, MA

Discussion points

- Motivation
- Contribution
- Dataset
- Methodology
- Result

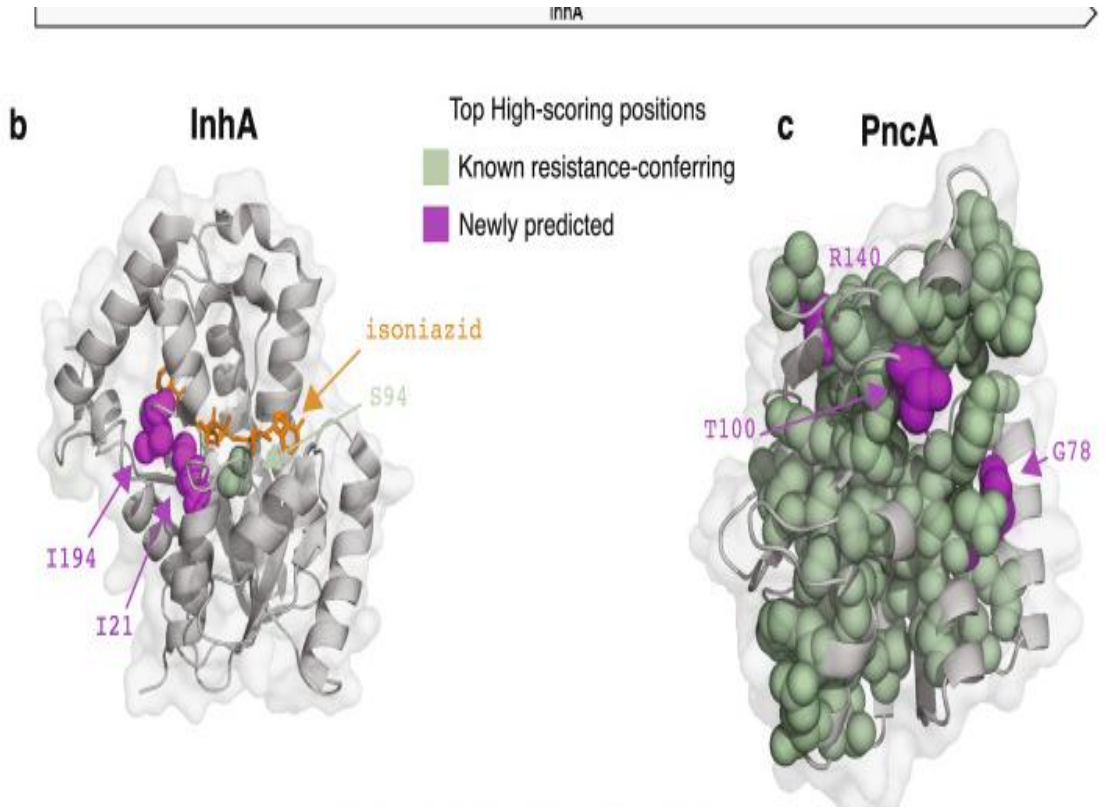
Motivation

- The rise of antibiotic-resistant *Mycobacterium tuberculosis* requires faster and more accurate **sequence-based methods** for **resistance prediction.**
- Traditional diagnostics are slow, and current sequence-only models struggle with accuracy on resistance prediction, particularly for second-line antibiotics, highlighting the need to integrate structural data.[1]



Hypothesis

- Incorporating **protein structure information enhances** the accuracy of antibiotic resistance prediction
- Unique sequences will prevent data leakage and subsequent bias in the model
- Mutations closer in 3D space exhibit similar phenotypic behavior



High-importance variants in the InhA protein mapped to its crystal structure[3]

Discussion points

- Motivation
- Contribution
- Dataset
- Methodology
- Result

Considered Approaches		
Baseline	Ridge Regression (Field Standard)	Sequence only
		Standard scikit-learn implementation
State of the art	ESM2-8M (Protein Language Model)	Sequence Only
		Trained on uniref database and has proven powerful zero-shot prediction tasks.
Our contribution	Fused Ridge Model (Explainable Model)	Sequence + Structure
		Based on prior work of fused lasso (tibshirani et al [2004])

Novel contribution

- New Dataset and Problem Space:
  - We introduce a novel dataset where traditional PLMs like ESM struggle, and
  - show that incorporating structural information improves prediction accuracy while maintaining comparable identification of true resistance-conferring mutations with baseline ridge.
- Effective with Limited Data:

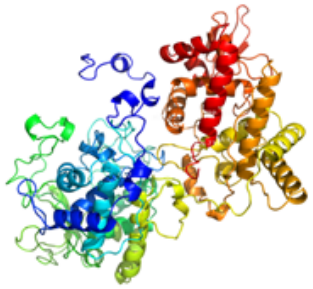
  Our approach demonstrates strong performance even with limited labeled sequence data.
- Zero-Shot ESM Performance:
  - Zero-shot ESM embeddings underperform compared to simple supervised models in distinguishing resistant from susceptible M.tuberculosis strains.

Training a fused ridge model for phenotype prediction from unique protein sequences

Sequence Data

Isolate	DNA	Protein	Pheno-type
00R1399	gtggctcgca	PPIT	R
00R0223	gtggctcgta	PPVT	S
00R0453	gtggctcgca	PPIT	R

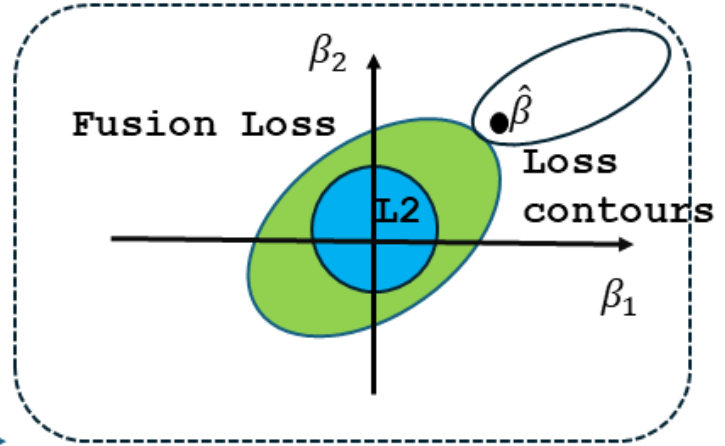
Structure Data



Reference-alternate encoded feature matrix generation

Weighted Distance Matrix

Fused Ridge Model



Model Output

Isolate	Protein Sequence	Phenotype (per drug)	Resistance conferring Residue
Peru3047	GPADLVG	Susceptible	Ala106 Arg104 glu582
Peru3292	KGNPLPA	Resistant	
00R1399	VPEQHPP	Resistant	

Discussion points

- Motivation
- Contribution
- Dataset
- Methodology
- Result

Dataset

- Data Leakage: Addressed by retaining only unique protein sequences, a critical issue often overlooked in the literature, to ensure reliable predictions

- Nine *M.tuberculosis* protein-coding genes

Gene	# of unique sequences	Gene Length (nt)	# variable positions	Protein Structure Length
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

Table 1: Data Summary of *M.tuberculosis* Antibiotic Resistance Genes. Each gene had 31452 sequences.

Discussion points

- Motivation
- Contribution
- Dataset
- Methodology
- Result

Methodology

Dataset Preparation

DNA to Protein Translation

- Nine *M.tuberculosis* protein-coding genes
- Translational Alignment:
  - Aligned to the reference H37Rv genome sequence.
  - DNA sequence cleaning (non-nucleotide characters and gap) and translation
  - Frameshift flagging
  - Insertion/Deletion handling to maintain alignment
- Feature matrix:
  - One-hot encoding of protein sequences
  - Only unique sequences retained
- Distance map:
  - Represents the minimum atomic pairwise Euclidean distance between residues
  - Protein sequence positions mapped to structures

Step 1: Translational Alignment & Feature Preparation

Genotype		Translational Alignment with Frameshift	Protein | Frameshift	Pheno-type	One-hot Encoding	Feature Matrix
Strain 1	GTTACTGTATTC		VTVF | 0	S		00000 |0
Strain 2 (Frameshift)	GTTACGTATTC		VTY- | 1	R		00001 |1
Strain 3 (Inframe Indel)	GTTACT---TTC		VT-F | 0	S		00100 |0
Strain 4	GTTACTGTAATC		VTVI | 0	R		00010 |1
Strain 5	GTTACGTTAATC		VTLI | 0	R		00110 |1

Methodology

Development of Fused ridge model

Fused Lasso Model

- The lasso (Tibshirani 1996) penalizes a least squares regression by the sum of the absolute values (L1 norm) of the coefficients.
  - The form of this penalty encourages sparse solutions
- “fused lasso", a generalization of the lasso designed for problems with features that can be ordered in some meaningful way
  - The fused lasso penalizes both the L1 norm of the coefficients and their successive differences
  - It encourages both sparsity of the coefficients and sparsity of their differences, that is, local constancy of the coefficient profile
  - Useful in the dataset of our case because the number of features p is much greater than N, the sample size

Fused Ridge Model

- Custom Adaptation:
  - Modifying the fusion penalty to use squared differences and an L2 norm
- Fusion penalty:
  - 3D Euclidean distances between protein mutations
- Enforces similarity in the coefficients of structurally adjacent mutations
  - based on our hypothesis that mutations closer in 3D space exhibit similar phenotypic effects

Optimizing the fused ridge model

- Developed a customized sub-gradient descent algorithm to optimize the fused ridge model
  - Explored four variants of the subgradient descent: vanilla, gradient clipping (enhanced), momentum and nesterov
- Warm-up coefficients:
  - used the final coefficient values from the baseline Ridge model as the initial coefficients for the fused ridge model
  - leveraged the stability and performance of the Ridge model to provide a good starting point for the fused ridge optimization.

Subgradient Descent

- Combines subgradients from MSE, L2 penalty, and fusion penalty.
- Utilizes gradient clipping and learning rate decay for stability and convergence
- Runs quicker because of its analytic gradient property (~8 minutes for all the 9 proteins)
- $\mathcal{O}(T.n)$ – T (number of iterations), n (number of parameters)

Subgradient descent

Working Process

- Returns local best minima point
- Adaptive learning rate through harmonic decay rule
  - Harmonic rule: learning_rate_i= learning_rate / (1+ beta * i) where i is the iteration
- Allows the algorithm to take larger steps initially and fine-tune the convergence as it approaches the optimal solution
- Gradient clipping prevents instability caused by excessively large gradients.

Objective Function

Convex Optimization Problem

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

Mean Sq. Error + L2 Regularization + Fusion Penalty

Table 2: Components of the Objective Function

Term	Description
$L(\beta)$	Total loss function
N	Number of observations
p	Number of features
y_i	Observed value for the i-th observation
x_{ij}	Value of the j-th feature for the i-th observation
β_j	Coefficient associated with the j-th mutation
α	Regularization parameter for the L2 penalty
λ_{fuse}	Regularization parameter for the fusion penalty
w_{jk}	Weights derived from the distance matrix, penalizing the difference between coefficients β_j and β_k

Step 3: Phenotype Prediction and True Variant Discovery		
Test Genomic Data	Predicted Phenotype	Variant Discovery
CAGGAGCT...	Resistant	True Variants discovered from model coefficients: [450, 445, 431,170, 428]
CCAACTCG...	Susceptible	
CTCCTCCA...	Susceptible	
CGCTGTCA...	Resistant	

Methodology

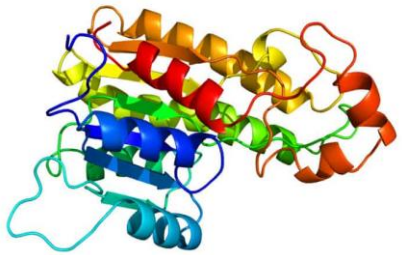
Zero shot phenotype prediction using ESM

Zero-shot Approach

- Literature:
  - ESM has demonstrated performance in mutation effect prediction with zero-shot approach.
  - LLR can identify beneficial vs deleterious mutations for protein functions
- Our steps:
  - Derive embeddings of the protein sequences from ESM
  - Compute Log-likelihood ratio of each mutated sequences embeddings compared to the wild-type H37Rv sequence following “masked-marginal” scoring function.
  - Train a logistic regression on computed LLR score to predict phenotype.

Zero-shot Phenotype Prediction using ESM2-150M Model [2]

Step 1:Tokenize Input Sequence



Tokenization →

```
Tokens: ['<cls>', 'M', 'R',
'A', 'L', 'I', 'I', 'V',
'D', 'V', 'Q', 'N', 'D',
'<eos>']
```

Step 2:Mutation Parsing

Mutation	Output: wildtype residue, position, mutated residue
Ala152Val	Ala , 152, Val
Ala285Thr	Ala, 285, Thr

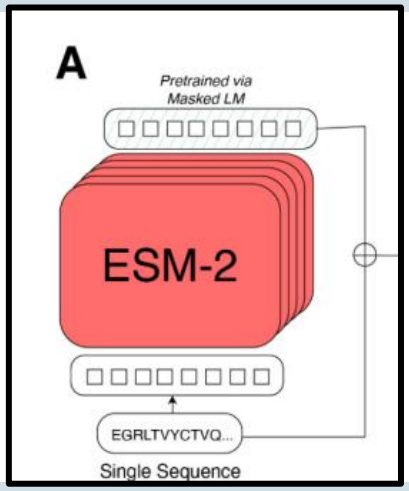
Step 3: Masking

1. Take the Tokenized Input Sequence
2. Locate each mutation position in the tokenized Sequence
3. Replace the wild-type residue at the specified position with mask token

Zero-shot Phenotype Prediction using ESM2-150M Model [2]

Step 4: Forward Pass through Model

Masked Input Sequence



Output Logit

Tensor of shape (1, sequence length, 33)

Step 5: Compute Log Likelihood Ratio

Scoring Function: $\log p\left(x_i = x_i^{mt} \middle| x_{-M}\right) - \log p\left(x_i = x_i^{wt} \middle| x_{-M}\right)$

M – set of mutated position

x_{-M} – sequences with masked mutations

x_i^{mt} – mutated aa at position i

x_i^{wt} – wildtype aa at position i

Step 6: Fit a Logistic Regression

Input: LLR for each sequence

Training

Prediction

Target: Resistant or susceptibility

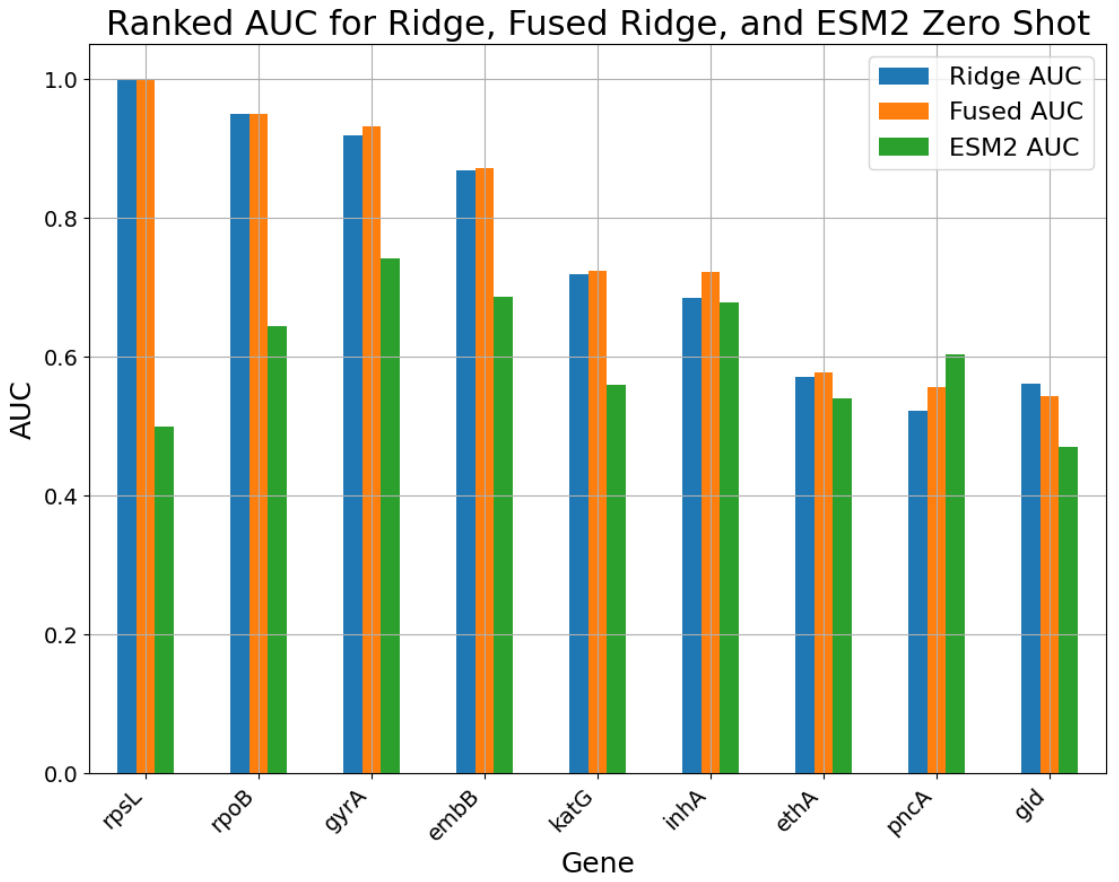
Discussion points

- Motivation
- Contribution
- Dataset
- Methodology
- Result

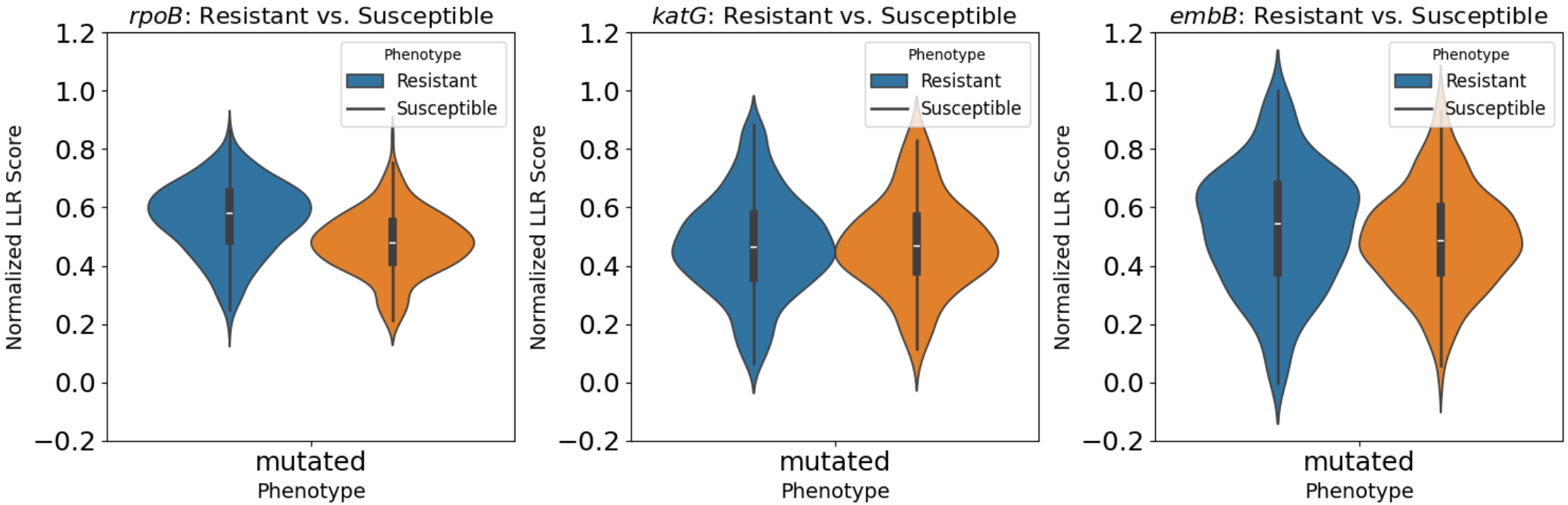
Fused Ridge model leverages 3D structural input to yield higher prediction scores on small dataset.

Best Performing Model: Fused Ridge	
Metric	**Value**
Fused Ridge outperformed ESM2	88.89%
Fused Ridge outperformed baseline Ridge	66.67%
Outperformed both baseline Ridge and ESM	55.59%
Baseline Ridge outperformed both Fused Ridge and ESM	22.22%
ESM outperformed both baseline Ridge and Fused Ridge	11.11%

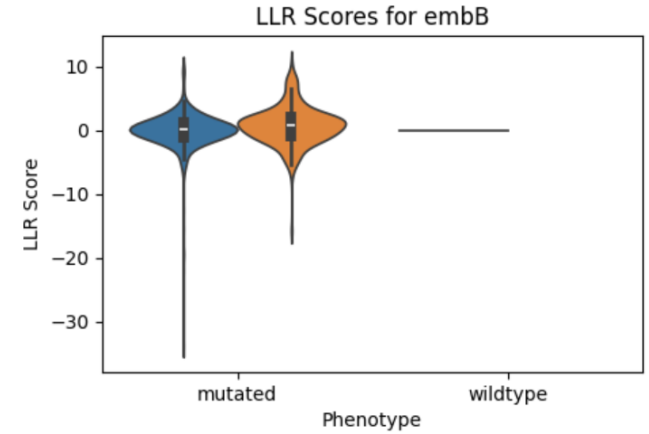
Average AUC	Value
Baseline Ridge	0.755
Fused Ridge	**0.764**
ESM	0.603

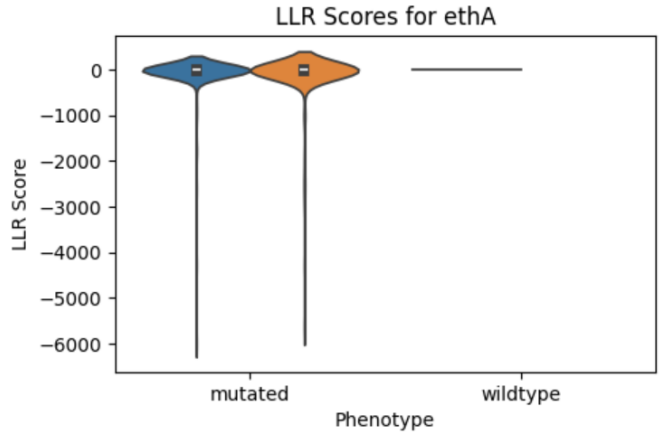


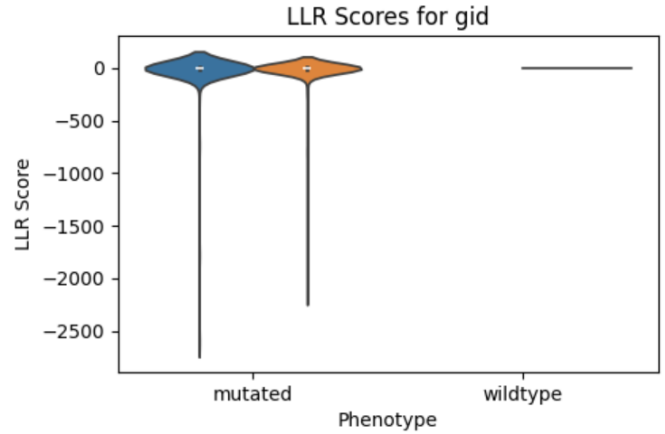
ESM2 struggles to distinguish resistant and susceptible phenotypes

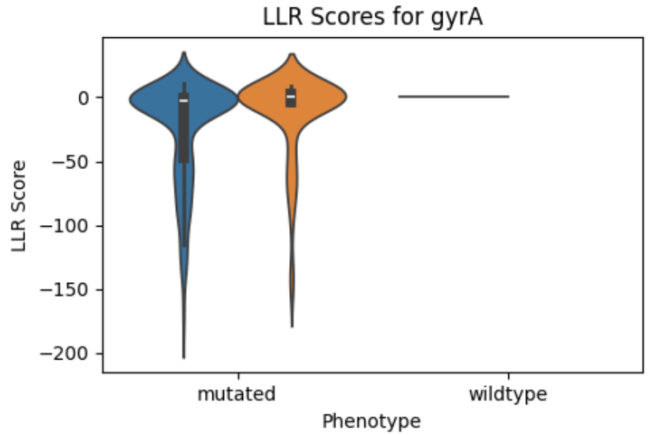


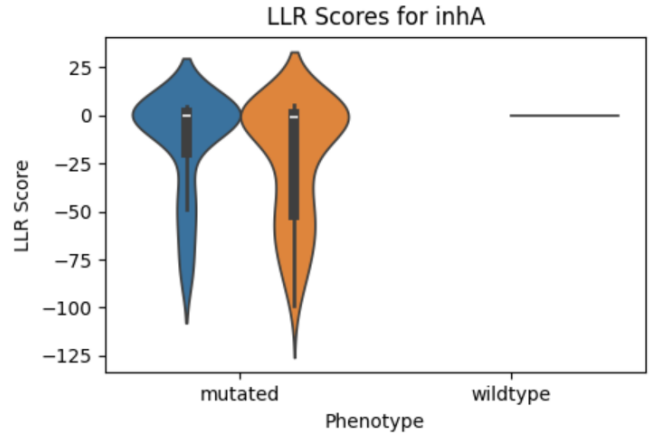
LLR score distribution for distinguishing resistant (R) and susceptible (S) phenotypes based on embeddings derived from the ESM-2 model.

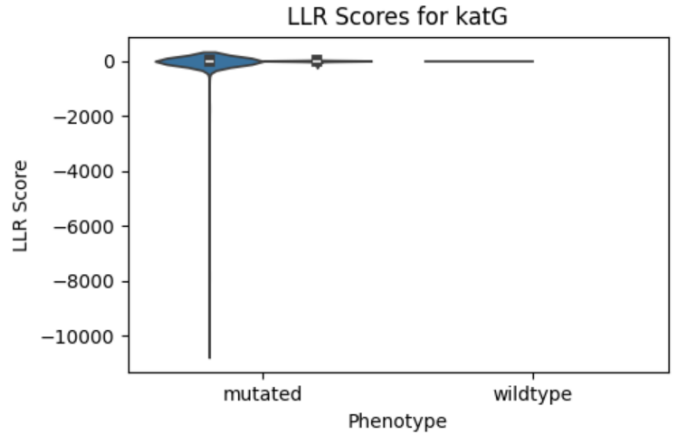

LLR Scores for embB
LLR Score
10
0
−10
−20
−30
mutated
wildtype
Phenotype

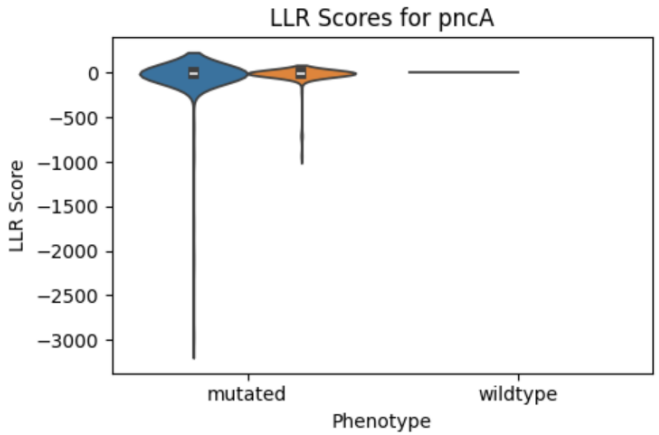

LLR Scores for ethA
LLR Score
0
−1000
−2000
−3000
−4000
−5000
−6000
mutated
wildtype
Phenotype

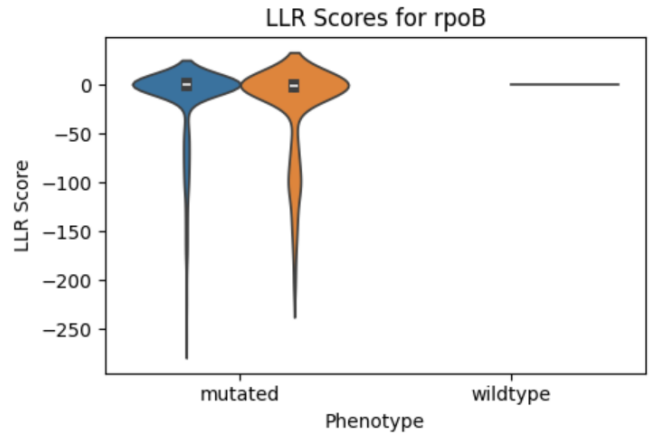

LLR Scores for gid
LLR Score
0
−500
−1000
−1500
−2000
−2500
mutated
wildtype
Phenotype

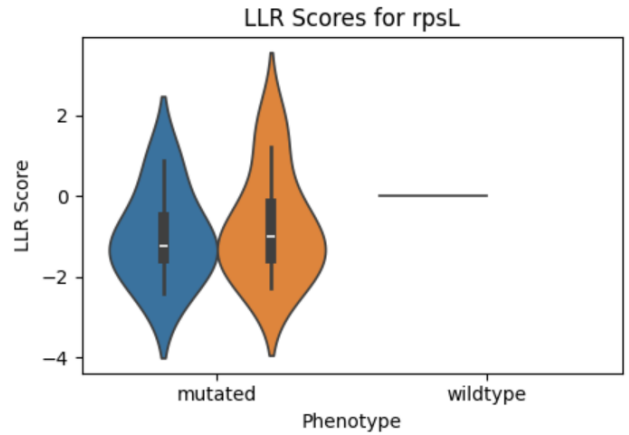

LLR Scores for gyrA
LLR Score
0
−50
−100
−150
−200
mutated
wildtype
Phenotype


LLR Scores for inhA
LLR Score
25
0
−25
−50
−75
−100
−125
mutated
wildtype
Phenotype


LLR Scores for katG
LLR Score
0
−2000
−4000
−6000
−8000
−10000
mutated
wildtype
Phenotype

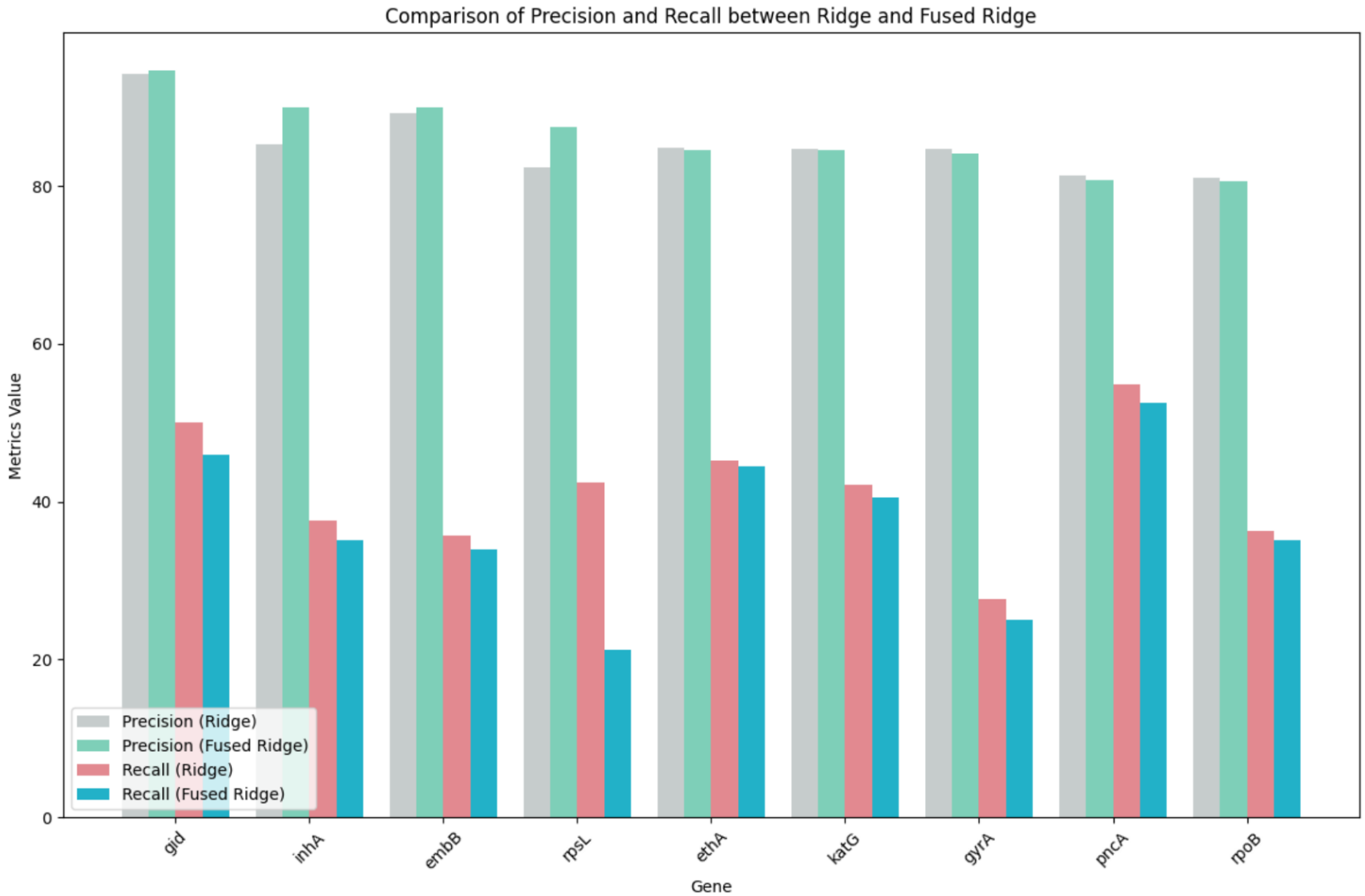

LLR Scores for pncA
LLR Score
0
−500
−1000
−1500
−2000
−2500
−3000
mutated
wildtype
Phenotype


LLR Scores for rpoB
LLR Score
0
−50
−100
−150
−200
−250
mutated
wildtype
Phenotype


LLR Scores for rpsL
LLR Score
2
0
−2
−4
mutated
wildtype
Phenotype

True Variants Discovery

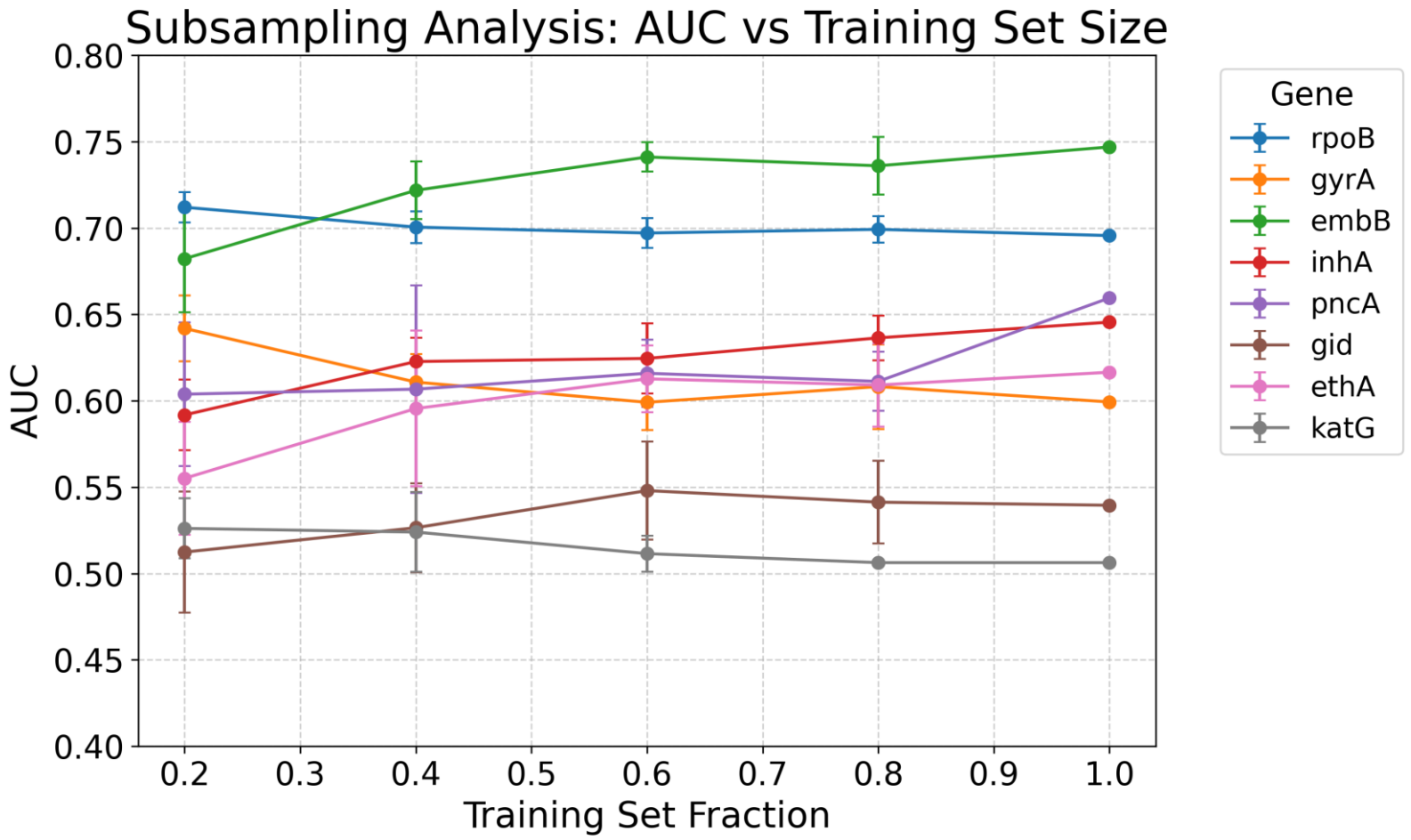
- Important step for interpretability of the models (baseline ridge and fused ridge)
- Feature Importance
  - Compute feature importance based on model coefficients.
  - Identify top features using a cutoff value (e.g., 95th percentile).
  - Sort and rank the features based on their importance
- Precision and Recall Calculation
  - Precision: Proportion of true positive variants among the predicted positive variants.
  - Recall: Proportion of true positive variants identified out of all actual positive variants.


Comparison of Precision and Recall between Ridge and Fused Ridge
Metrics Value
0
20
40
60
80
Precision (Ridge)
Precision (Fused Ridge)
Recall (Ridge)
Recall (Fused Ridge)
gid
inhA
embB
rpsL
ethA
katG
gyrA
pncA
rpoB
Gene

True Variant Discovery of Fused Ridge	
Accuracy of Prediction (precision)	• Consistently achieved high precision, with values ranging from 80.54% to 94.68%. • This indicates that most predicted variants were correct.
Ability to capture relevant variants (recall)	• Recall varied across genes, ranging from 21.21% to 52.59%. • The model sometimes missed a significant portion of true variants.

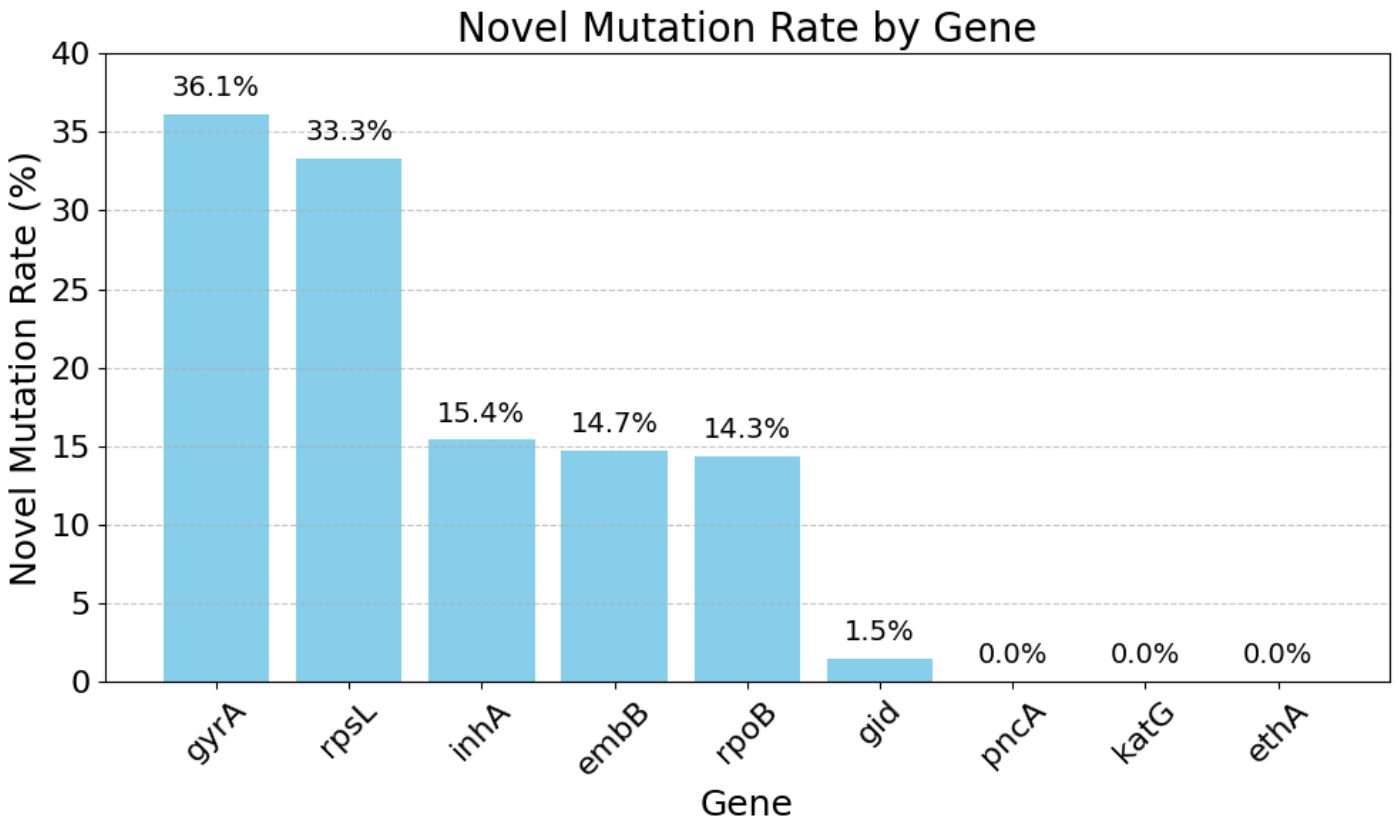
Discussion

Generalization Test: Subsampling Analysis
Model shows robustness in majority cases



NOVEL MUTATION RATES BETWEEN TRAIN AND TEST

Deduplicated data structure inherently enforces generalization to unseen variants



Comparative Model Insights

- ESM underperforms both ridge and fused ridge in majority cases.
  - Likely due to its reliance on evolutionary patterns which is not fully applicable to recent drug-resistance applications
  - Fused ridge model leverages explicit 3D structural information, yielding higher145prediction scores
- True variants discovery is as par with baseline ridge
  - Incorporation of additional priors has not impaired the ability to identify causal variants

Comparative Model Insights

- Rare variants discovery
  - Fused ridge falls behind ridge regression in rare variants discovery
  - Possibly due to our hypothesis of closer mutations in 3D space emit similar phenotypic behavior
- Higher MSE scores in fused ridge:
  - Potential overfitting due to limited data
  - Added complexity from fusion penalty can increase model variance

Future Directions

- Rare penetrant mutations confer risk of disease
  - Improve genetic risk prediction using primateAI
  - Predict disease causing genetic mutations
- Generative flow network to enhance prediction performance
  - Probabilistic model that sample from a distribution proportional to the reward function
  - Allows for diverse sampling for high-reward candidates – “r-conferring mutations”
  - Maximizes multiple objectives unlike reinforcement learning methods that maximize a single objective

References

1. World Health Organization. Global Tuberculosis Report 2023. World Health Organization, Geneva,3472023
2. Language models enable zero-shot prediction of the effects of mutations on protein function
Joshua Meier, Roshan Rao, Robert Verkuil, Jason Liu, Tom Sercu, Alexander Rives
3. Green AG, Yoon CH, Chen ML, Ektefaie Y, Fina M, Freschi L, Gröschel MI, Kohane I, Beam A, Farhat M. A convolutional neural network highlights mutations relevant to antimicrobial resistance in Mycobacterium tuberculosis. Nat Commun. 2022 Jul 2;13(1):3817. doi: 10.1038/s41467-022-31236-0. PMID: 35780211; PMCID: PMC9250494.