Muslims in ML Workshop @ NeurIPS 2026

Multi-Drug Generalization of Genomic Tuberculosis Resistance Prediction

Does a Permutation-Invariant Encoder Transfer to Underrepresented Cohorts?

Aicha Boutorh · **Kamar Hibatallah Baghdadi**
National School of Artificial Intelligence (ENSIA) · Sidi Abdallah Campus, Algiers, Algeria
aicha.boutorh@ensia.edu.dz
kamar.baghdadi@ensia.edu.dz

6.9% — **Cohort share** 710 of 10,310 isolates from Muslim-majority, Global-South countries

0.91–0.97 — **AUROC in-distribution** 4 drugs, 3-fold CV

≤ 0.036 — **Zero-shot AUROC drop** for RIF, RFB and INH

−0.142 — **EMB AUROC collapse** recall −0.305 under geographic shift

1 Motivation

- TB is the world's deadliest bacterial infection; **MDR-TB** is a persistent global health emergency.
- Whole-genome sequencing + ML promises **rapid, culture-free** resistance prediction, most valuable where phenotypic testing is slowest.
- Public WGS compendia are dominated by **high-income and East-Asian** settings; South & Central Asia and North & Sub-Saharan Africa are sparse or absent.
- Random or stratified CV on pooled data **never holds out a population**, so it cannot reveal this failure.

Does a modern genomic resistance model actually transfer to the populations it is meant to serve?

2 Contributions

- **One shared permutation-invariant Set-Transformer encoder** feeding four drug heads (RIF RFB INH EMB), trained with a masked multi-task loss.
- **Transferability audit**: representation gap (710/10,310 = 6.9%), strict Leave-Country-Out zero-shot, Leave-Lineage-Out control, few-shot adaptation with bootstrap CIs.
- **Transfer is drug-specific**: three of four drugs transfer well, ethambutol does not.

3 Method: single-path Set Transformer

Isolate = unordered set of called genomic variants

`chr1:12345:A>G` `chr1:761155:C>T` `chr1:2155168:A>C` `chr1:4247429:G>A` … ≤ 1,200 variants, no order

Token embedding (d = 128) → Set Attention Blocks ×2 (epistatic variant–variant interactions) → PMA pooling (permutation-invariant set embedding) → Shared MLP trunk

RIF head R / S · RFB head R / S · INH head R / S · EMB head R / S

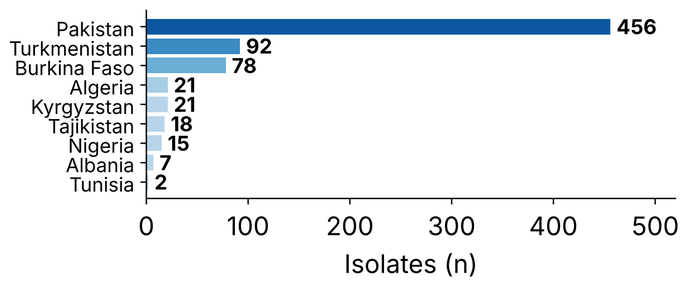
masked multi-task binary cross-entropy (4 drugs)

- Variants are tokens *chrom:pos:ref>alt*; a **BERT WordPiece** tokenizer reduces out-of-vocabulary issues.
- SAB is permutation-equivariant and PMA invariant, so predictions **do not depend on variant order**.
- Masked loss: isolates with **missing labels** still train the drugs they are labelled for.
- Hidden dim 32, 2 layers, dropout 0.33, AdamW (lr 1.41×10^{-3}, wd 10^{-5}); sets capped at 1,200 tokens (51% of isolates truncated) to fit a 14.6 GB GPU.

4 Data & cohort definition

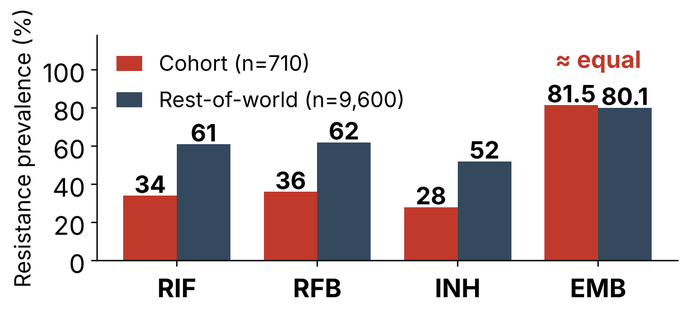
CRyPTIC compendium: 4 per-drug tables merged → **10,310 isolates** (RIF, RFB, INH, EMB phenotypes).

Cohort (built here; not published by CRyPTIC): Muslim-majority countries present in the data → **9 countries, 710 isolates (6.9%)**. The other **9,600** form the rest-of-world (RoW) pool.



Cohort composition. Pakistan 64.2%, Turkmenistan 13.0%, Burkina Faso 11.0%. Albania and Tunisia are too small for per-country claims.

5 Prevalence audit



Resistance prevalence (%), cohort vs rest-of-world.

- RIF/RFB/INH: cohort prevalence is **lower** (28–36% vs 52–62%).
- **EMB is nearly identical** (81.5% vs 80.1%): any EMB gap **cannot be a class-balance artifact**.

6 Four evaluation protocols

1. **In-distribution** — 3-fold stratified CV on all 10,310 isolates.
2. **Zero-shot Leave-Country-Out (LCO)** — Train on the 9,600 RoW isolates only; test on a held-out half of the cohort (n=355). The other half is the adaptation pool.
3. **Leave-Lineage-Out (LLO)** — Hold out each well-represented lineage (barcode SNPs) in turn; 3 seeds.
4. **Few-shot adaptation** — Fine-tune on RoW + 10/20/50% of the local pool; 20 bootstrap draws → 95% CIs; permutation test.

7 In-distribution performance

Drug	F1	Recall	Precision	AUROC	MCC
RIF	0.946	0.935	0.957	0.970	0.871
RFB	0.940	0.934	0.947	0.967	0.851
INH	0.933	0.930	0.936	0.965	0.866
EMB	**0.865**	**0.783**	**0.966**	**0.910**	**0.560**

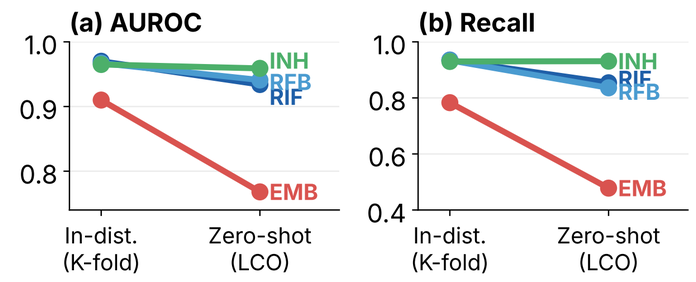
3-fold CV mean over 10,310 isolates (std ≤ 0.030).

- RIF, RFB, INH: **F1 ≈ 0.93–0.95**, **AUROC ≈ 0.965–0.970**.
- EMB is weakest **even in-distribution** (F1 0.86, recall 0.78): weaker genotype–phenotype link.

8 Zero-shot geographic transfer

Drug	Prev.	F1	Recall	AUROC	ΔAUROC	ΔRecall
RIF	0.35	0.833	0.854	0.934	**−0.036**	**−0.081**
RFB	0.36	0.817	0.836	0.941	**−0.026**	**−0.098**
INH	0.29	0.860	0.931	0.959	−0.006	+0.001
EMB	**0.83**	**0.639**	**0.478**	**0.768**	**−0.142**	**−0.305**

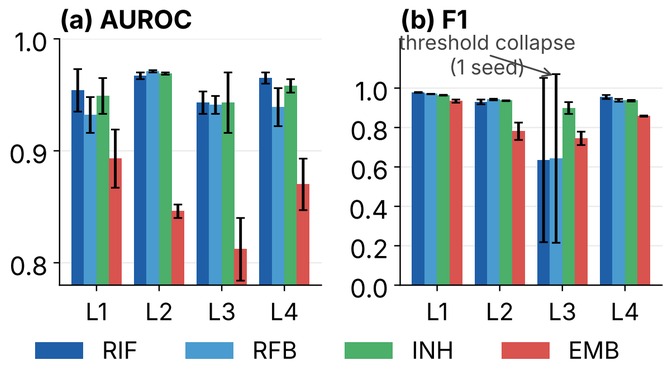
Held-out cohort test set (n=355); Δ relative to in-distribution.



Zero-shot degradation. INH is essentially flat; EMB drops sharply on AUROC and recall (below 0.5).

- **INH transfers almost perfectly**; RIF and RFB lose a clinically tolerable amount.
- **EMB collapses**: the model misses **more than half** of truly EMB-resistant cohort isolates.

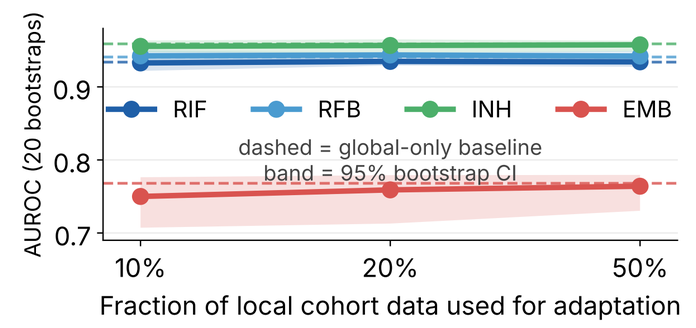
9 Leave-Lineage-Out: geographic or phylogenetic?



Held-out lineage (mean ± std, 3 seeds). L1–L4: 586 / 3,816 / 885 / 4,694 isolates.

- **EMB is weakest under every lineage** (macro AUROC 0.855, MCC 0.467) vs INH 0.955 / 0.857 → a **phylogenetic rather than purely geographic** failure.
- **INH transfers robustly** (AUROC 0.94–0.97), echoing the portable *katG* S315T variant.
- Lineage 3 (cohort's South-Asian lineage): AUROC ≈ 0.94 but F1 0.64 ± 0.42 after **one seed's threshold collapsed**: calibration, not ranking.

10 Few-shot domain adaptation



- RIF, RFB, INH: **negligible gains**; already near-optimal.
- EMB AUROC **0.750 → 0.759 → 0.764** (10/20/50%): **within bootstrap noise**, far below the other drugs.

11 Why does EMB fail when INH does not?

INH resistance is dominated by a few **high-penetrance variants** (notably *katG* S315T) shared across lineages: a portable rule.

EMB depends on *embB* variants with weaker, heterogeneous genotype–phenotype coupling whose frequency differs between lineages, so a representation tuned on some lineages does not generalise. The LLO control reads the geographic collapse as a **phylogenetic generalisation failure** exposed by the cohort's lineage mix.

12 Limitations & future work

- **Cohort composition:** several countries have single-digit counts.
- **Variant truncation:** 51% of isolates truncated at 1,200 tokens.
- **Geography–lineage entanglement** remains; **3-seed** LLO variance.
- **Next:** Integrated Gradients of global vs locally-adapted models against the WHO mutation catalogue to look for missed *embB* determinants.

13 Conclusion

The representation transfers well for **rifampicin, rifabutin and isoniazid** but fails for **ethambutol** under strict geographic hold-out; a lineage hold-out localises the failure as phylogenetic.

Per-drug transferability audits on held-out populations and lineages should be a standard reporting requirement for clinical genomic models.

Data & ethics: publicly available, de-identified isolate data (CRyPTIC Consortium, PLOS Biology 20(8):e3001721, 2022). **Key references:** [1] WHO, Global tuberculosis report 2024 · [6] Boutorh, Baghdadi & Daoud, VAMP-Net, Applied Intelligence 56(16):508, 2026 · [7] Lee et al., Set Transformer, ICML 2019 · [8] Zaheer et al., Deep Sets, NeurIPS 2017 · [16] Coll et al., SNP barcode for MTBC typing, Nat. Commun. 5:4812, 2014 · [18] WHO, Catalogue of mutations in MTBC, 2nd ed., 2023.