

Where Tabular Foundation Models Falter on Genetic Data

Datasets That Expose and Provide a Path to Address the Gap

Anirban Das

Yan Cui

Department of Genetics, Genomics & Informatics
The University of Tennessee Health Science Center, Memphis, TN

The problem: genetic risk varies across ancestry

- The **same DNA variant** can strongly raise disease risk in one population and barely matter in another.
- Most genetic data come from people of **European ancestry**.
- A model that ignores this learns the majority's effects, and is **most wrong for the least represented**.

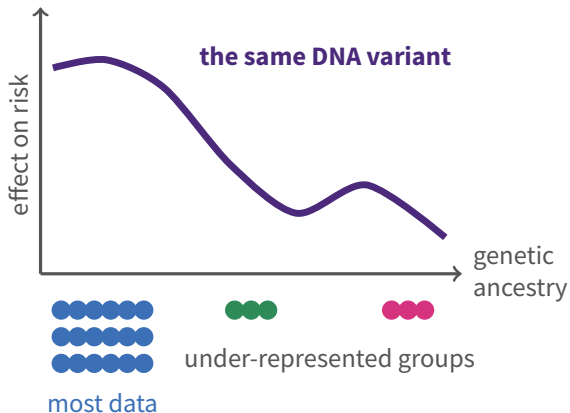


Illustration only.

Tabular foundation models were not built for this

- **TabPFN** and **TabICL**: pretrained once on millions of synthetic tables, then predict on a new table *in context*, with no training.
- Attractive as off-the-shelf predictors for genotype-to-phenotype tables.
- But their synthetic pretraining tasks do not capture **variant effects that change across ancestry**.

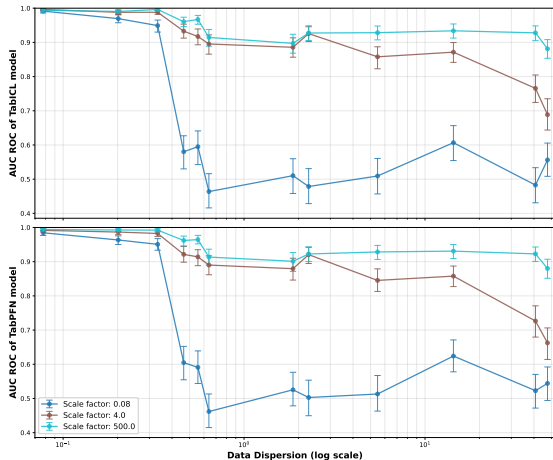
Question: what happens when they meet ancestry-dependent genetic effects?

They fail in controlled tests

Simulated genotype-to-disease tables with two dials:

- **scale factor:** how strongly variant effects change across ancestry (smaller = stronger change)
- **data dispersion:** how spread out in ancestry the examples given to the model are

**Strong ancestry effects:
near-perfect → near-random
as data become more diverse.**



Off-the-shelf TabICL (top) and TabPFN (bottom). Each point is one synthetic task.

...and on real patient data

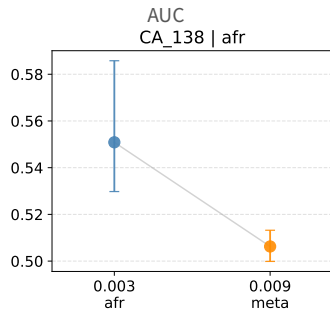
Cancer outcomes in the All of Us biobank. For each disease and ancestry group, the model gets the **same number** of examples: from the target group only, or pooled across all ancestries (more diverse). Both are tested on the same people.

13 / 16

TabICL: pooled did worse

12 / 16

TabPFN: pooled did worse



spread of the examples in ancestry (data dispersion)
TabICL, benign neoplasm of the skin: African-ancestry examples (afr) vs. pooled examples (meta).

Same amount of data: ancestral diversity of the examples is what hurts.

Our datasets: a path to address the gap

NEW: two released families of synthetic genetic tasks

Built to model what pretraining missed:

- variant effects that change across ancestry
- allele frequencies that differ between groups
- uneven group sizes

Gaussian-process family and a PRS-CSx-inspired family.

kaggle.com/datasets/lazydazz/where-tabular-fo-neurips26-sub-1839



Download on Kaggle

Continual learning of existing TFNs on these synthetic tasks only (no real data), tested on real genotype-to-phenotype tasks (All of Us):

17 / 17

cancer and respiratory diseases where TabICL improves for the people most distant in ancestry

No cost to well-represented people.

Take-home

1. Genetic effects vary across ancestry; today's tabular foundation models are not pretrained for this.
2. They **fail** on controlled synthetic tasks and on real All of Us data.
3. Our **released synthetic task families** expose the gap, and continual learning on them **improves real genotype-to-phenotype prediction** for under-represented people.

Datasets: kaggle.com/datasets/lazydazz/where-tabular-fo-neurips26-sub-1839

Contact: adas23@uthsc.edu ycui2@uthsc.edu



Funding and Acknowledgements

Funding. This work was supported by the US National Cancer Institute grant R01CA262296.

Acknowledgements. We gratefully acknowledge All of Us participants for their contributions, without whom this research would not have been possible. We also thank the National Institutes of Health's All of Us Research Program for making available the participant and sample data examined in this study. This study used data from the All of Us Research Program's Controlled Tier Dataset v7 and v8, available to authorized users. Ancestry coordinates and group labels used in this study are provided by the All of Us Research Program for research purposes and reflect genetic principal component-based groupings; they should not be interpreted as proxies for race, ethnicity, or social identity. We emphasize the importance of responsible interpretation of ancestry-related findings in genomic research and caution against conflating genetic ancestry with social constructs of race. All analyses were conducted on controlled-access data; no participant-level genetic or phenotype data are reported, and all results are aggregated over cohorts larger than 50 participants to reduce disclosure risk. This work builds on off-the-shelf models and software released by the teams behind TabICL, TabPFN, and CausalPFN; we gratefully acknowledge the researchers and developers who created and released these resources.