

Robustness of transferability estimation metrics for medical imaging

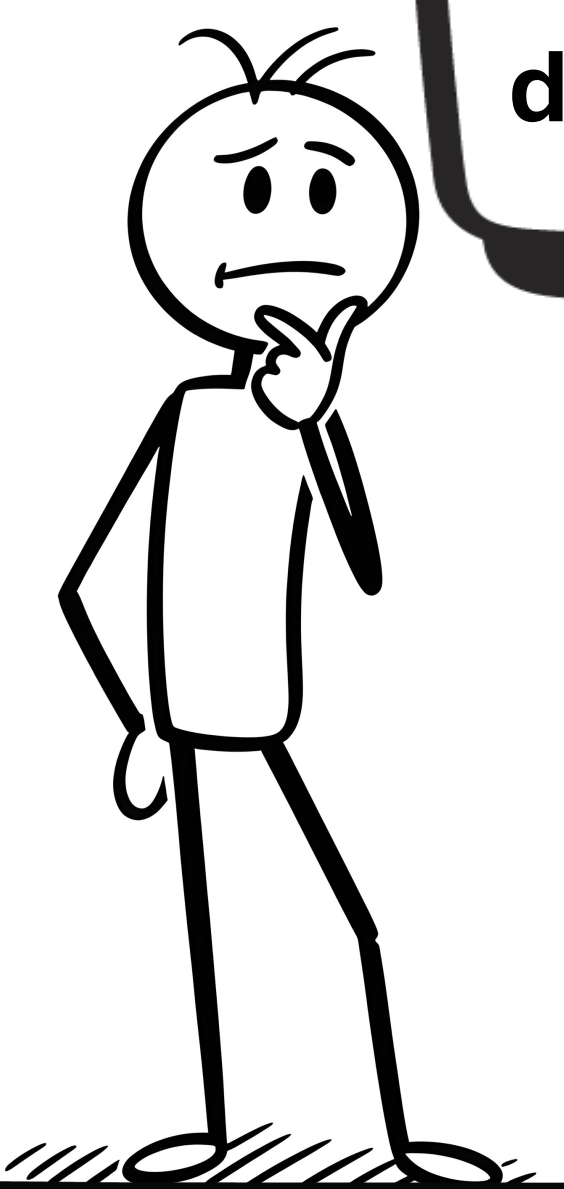
Niclas Claßen, Théo Sourget, Dovile Juodelyte, Rob van der Goot, Veronika Cheplygina

niclc@itu.dk



Motivation

Hmm, what source **model** should I use to **fine-tune** on this **target dataset** of brain tumors?



Method

Questions

1. How **robust** are TE metrics to changes of the target dataset?
2. What is the **impact** of the evaluation metric used for the reference ranking?

Setup

Data:

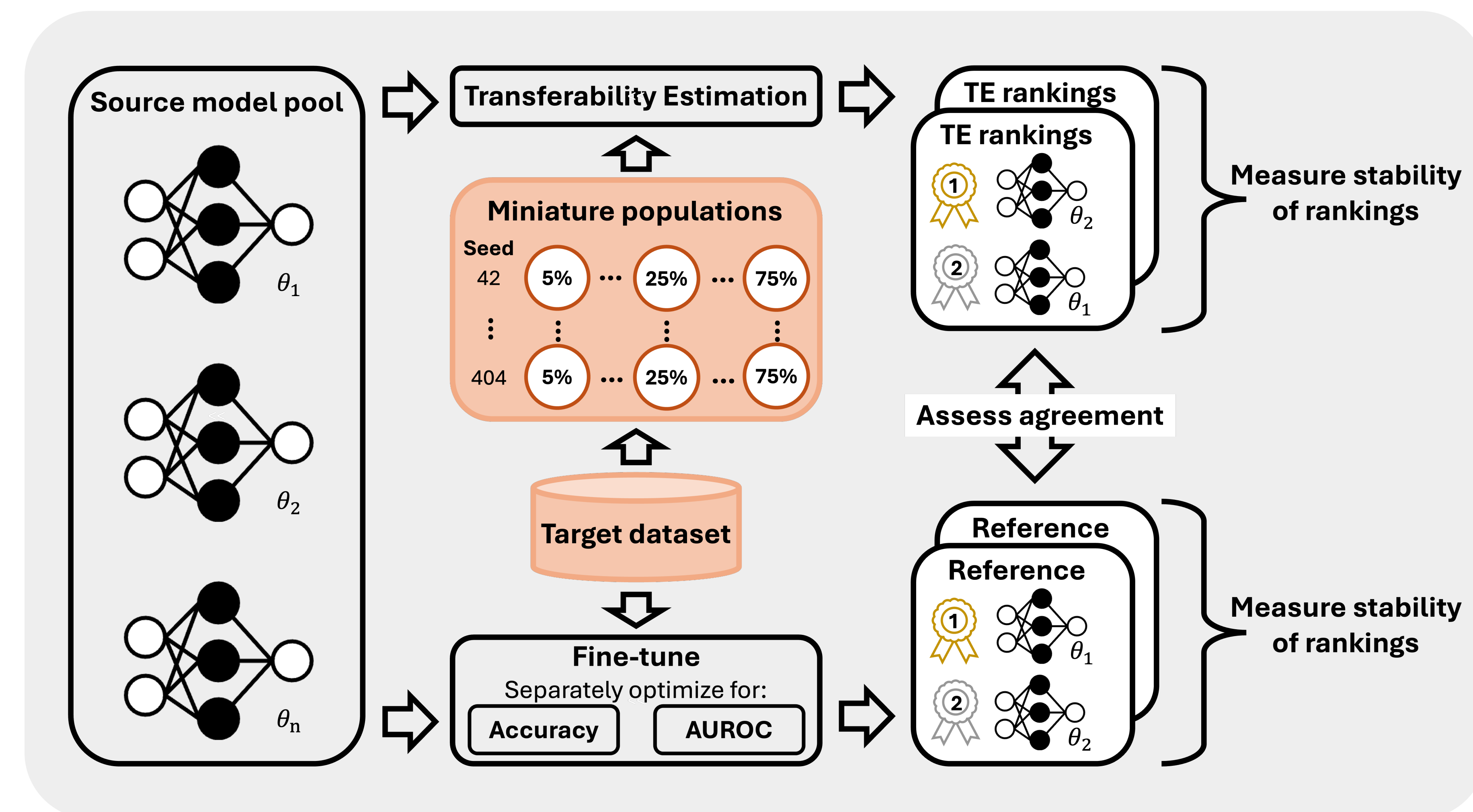
- MedMNIST v2 (2D) & ImageNet
- 8 different target datasets
- 12 source models per target

Model architecture:

- ResNet-18

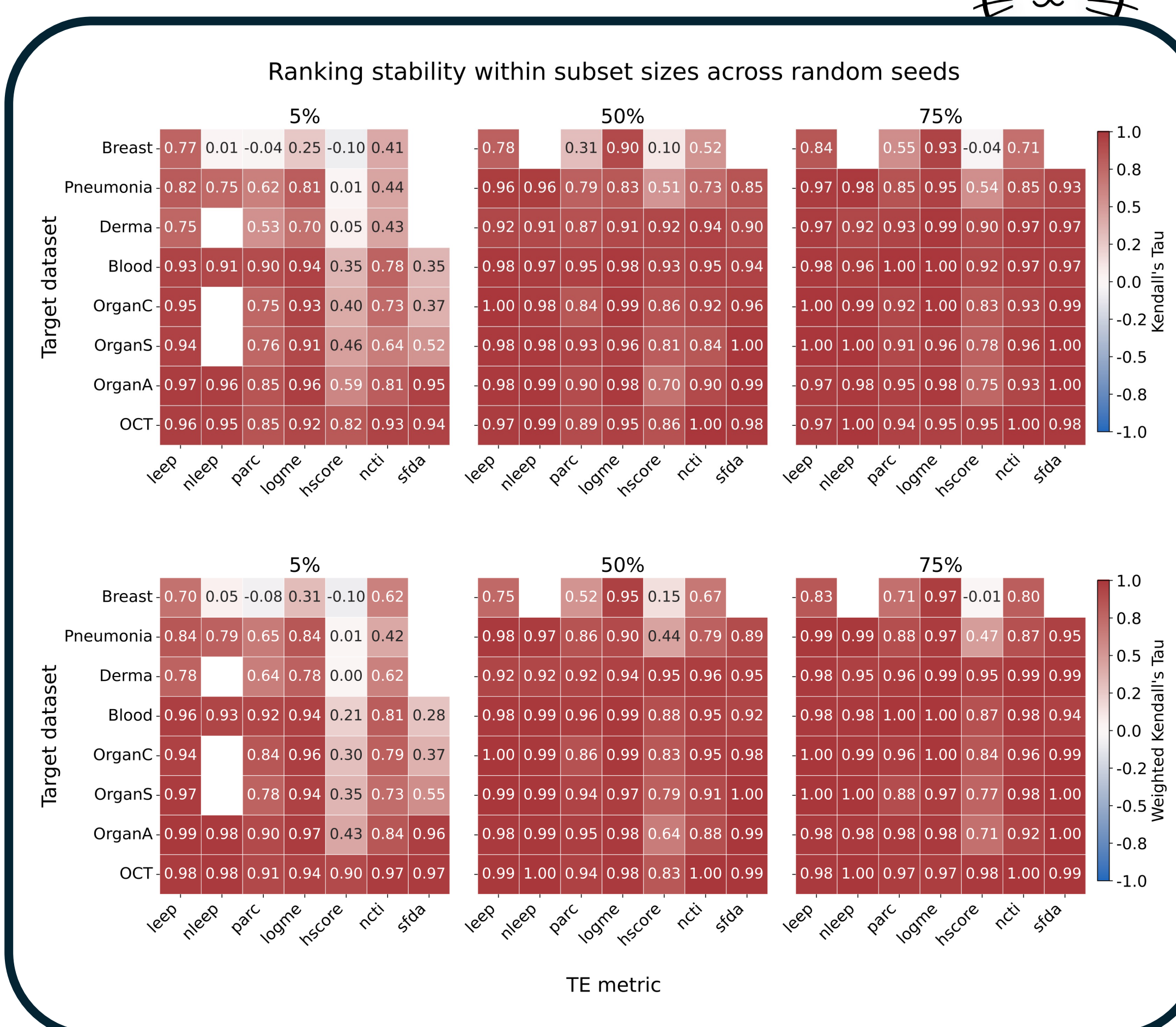
Evaluation of rank agreement:

- Kendall's tau
- Weighted Kendall's tau
- Spearman's rho



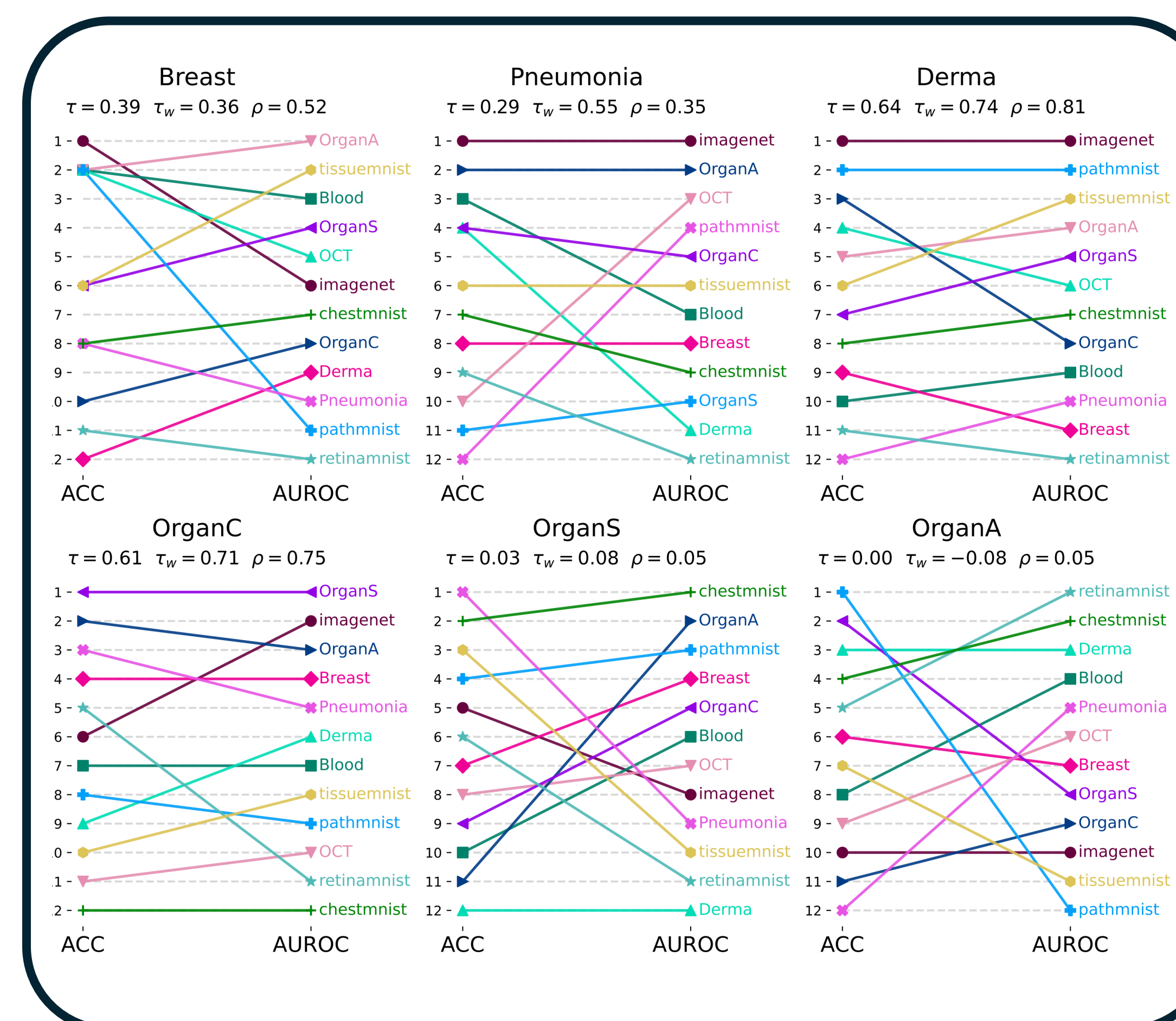
Results

TE metrics are sensitive to target dataset variations



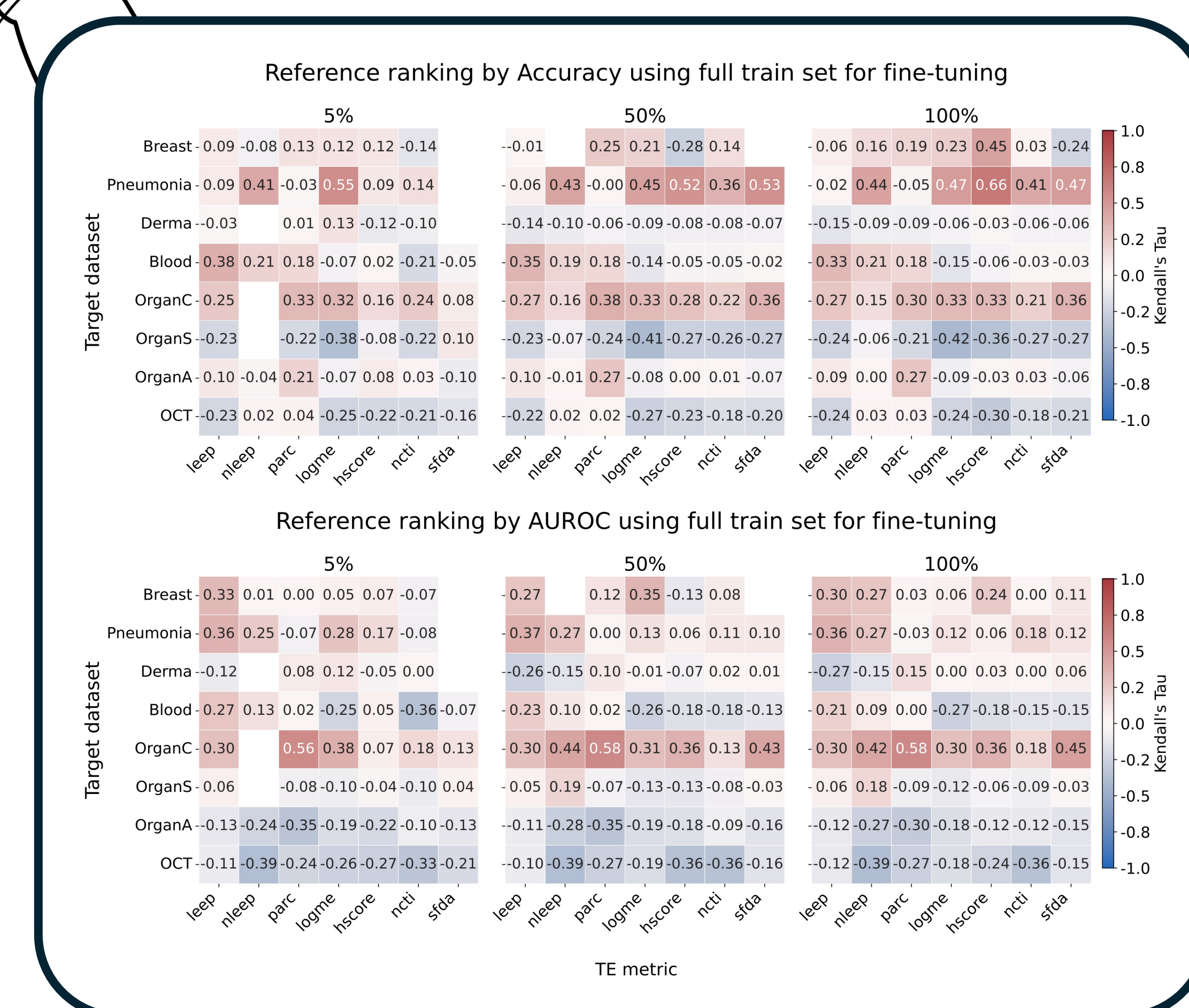
Changing the random seed for sampling a subset can result in a **different** ranking.

Reference ranking depends on evaluation metric



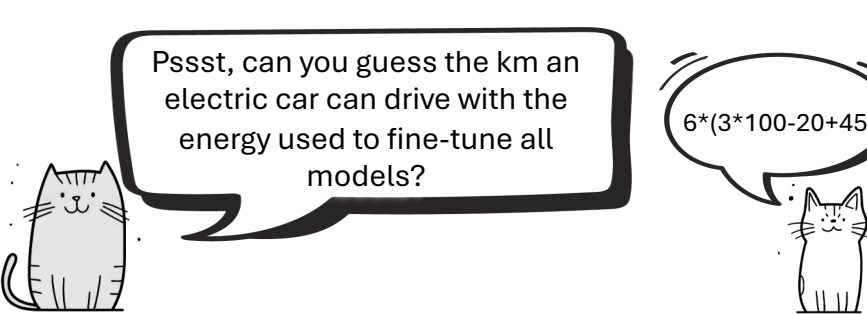
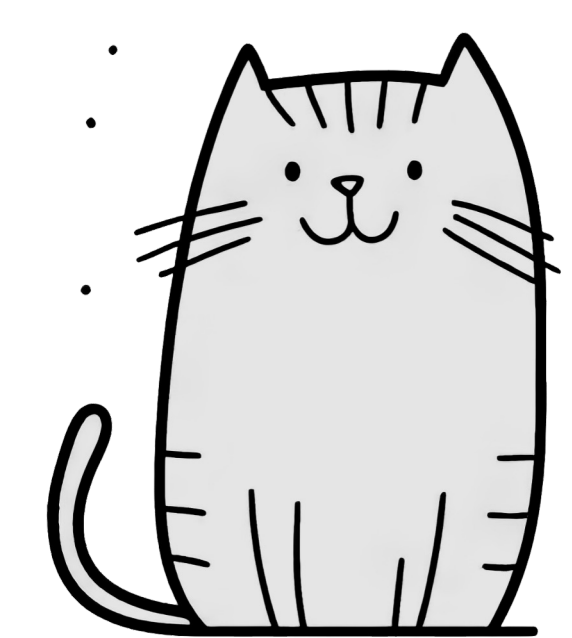
Hmm, none of the considered TE metrics agreed reliably with the reference rankings.

Low agreement between TE metrics and reference rankings



Not so fast!

1. Most TE metrics are **optimized for accuracy**.
2. It's **unclear** whether TE metrics are **applicable** to your dataset.



The "ground truth" changes when we optimize for AUROC instead of accuracy.