Project 3b: Exploring Overtraining and Membership in ESM-2

Egor Rakcheev Timon Giess Ala Sleimi Alexander Nielsen Polina Yugantyseva

November 24, 2025

Team

Ala Sleimi
MS Informatics

Timon Giess
MS Informatics

Alexander Nielsen
MS Data Science

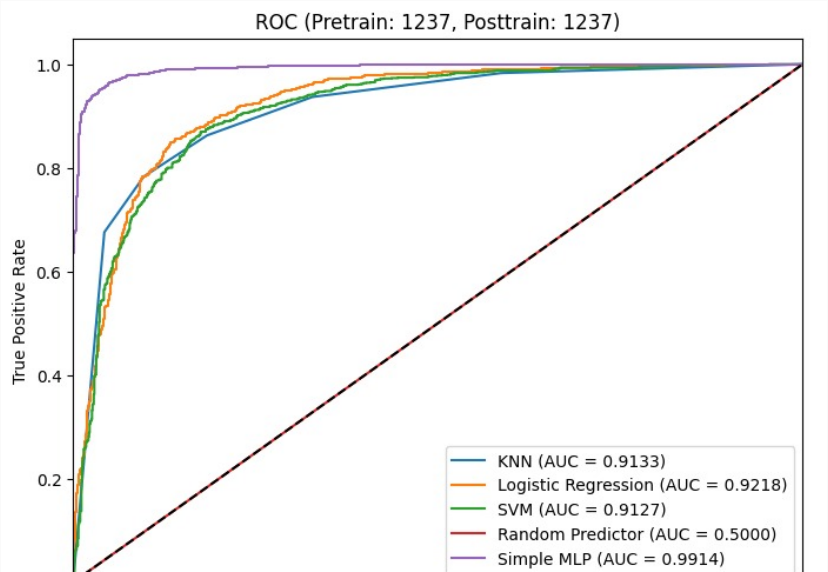
Egor Rakcheev
MS Informatics

Polina Yugantyseva
BS Bio+Info

Membership attack recipe

- Training-time proxy: sequences sampled from $\sim$ UniRef50.
- Post-train proxy: newly discovered proteins (N)
- Goal: learn a detector that flags members vs. non-members

Results snapshot



Great job! 0.99 AUC!

Meta UR50D process (ESM-2 original)

- Meta samples 43M UniRef50 clusters, then pulls UniRef90 sequences within clusters; ∼65M unique sequences seen from 138M [1].
- The exact UR50D training file was never released; only the sampling recipe is public.

[1] Zeming Lin et al., *Evolutionary-scale prediction of atomic-level protein structure with a language model*. Science 379, 1123–1130 (2023). DOI:10.1126/science.ade2574.

What does that mean for MI?

- For any candidate protein, only $\approx 65/138$ chance it was actually in the ESM-2 train pool.
- Our MI probes risk targeting "maybe-seen" sequences $\Rightarrow$ unclear ground truth.
- What were we even measuring??

NVIDIA to the rescue

- Member proxy: NVIDIA UR50D train shards (T) — mirrors Meta recipe but reproducible/public.
- Non-member proxies:
  - New proteins (N) outside the train stream,
  - Validation shards (V) from NVIDIA split,
  - hard homologs from UniRef90 $\setminus$ T for .

Hardness ladder

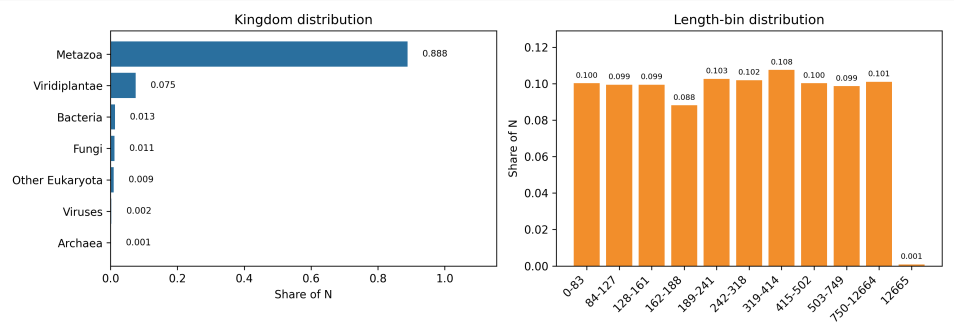
- Easy: $(S1, N)$ — $S1$ from T (train shards), N are new proteins; $|S1| = |N| = 1{,}237$.
- Medium: $(S2, S3)$ — $S2$ from T vs. $S3$ from V (val shards); equal size $K \approx 12\text{k}$.
- Hard: $(S4, U4)$ — $S4$ from T vs. $U4$ hard homologs from UniRef90 $\setminus$ T;

Each of the 3 pairs is then split 80/20 into train/test for the MI detector.

Sampling

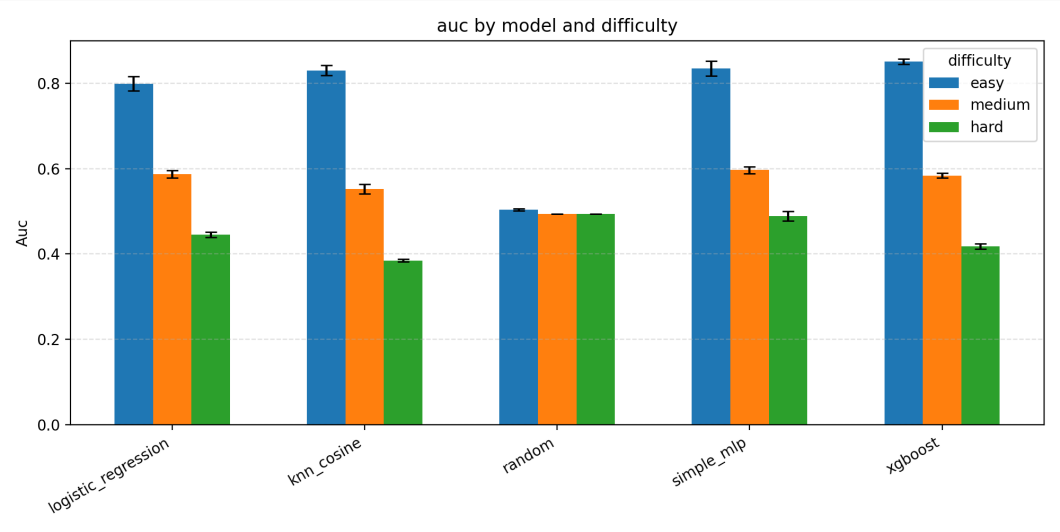
- Bin by (Kingdom, Length bin) on N.
- All sampled sets mirror the empirical N distribution to avoid easy shortcuts.
- Cluster-aware sampling (UR50) keeps representation balanced

N distribution (kingdom + length)

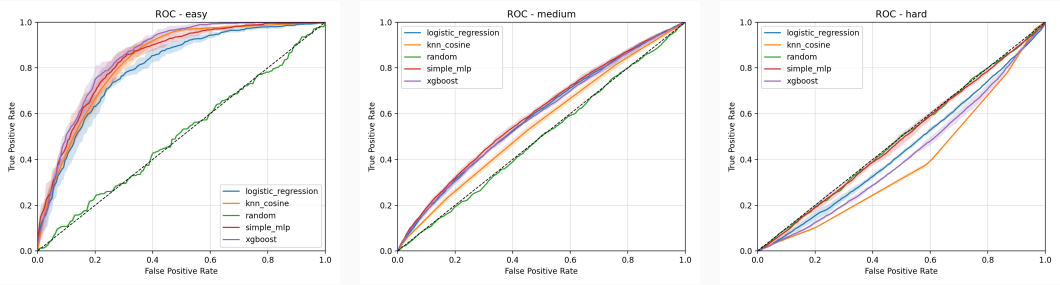


We use this empirical N distribution to set bin targets across T/V and UniRef90 sampling.

Accuracy (AUC)



ROC curves



Left: N vs S1 (easy). Middle: S2 vs S3 (medium). Right: S4 vs U4 (hard).

Idea from last week

- Results were not promising even on flawed data.
- 650M embeddings are not ready yet (NVIDIA released 8M, 650M, 3B) so we still didn't try them.
- Likely a waste of time.

sorry

Biggest success & next steps

Biggest success since last presentation

- Improved the data curation process.
- Identified the issue with using the Meta model.

Next steps (why/how they help)

- Finish the 650M embeddings to test larger capacity models.
- Try other MI methods and ideas from the first presentation.
- Improve the MLP model and explore additional approaches.