A DIAGNOSTIC BENCHMARK FOR GENOMIC ML

GENEB

Why Genomic Models Are Hard to Compare

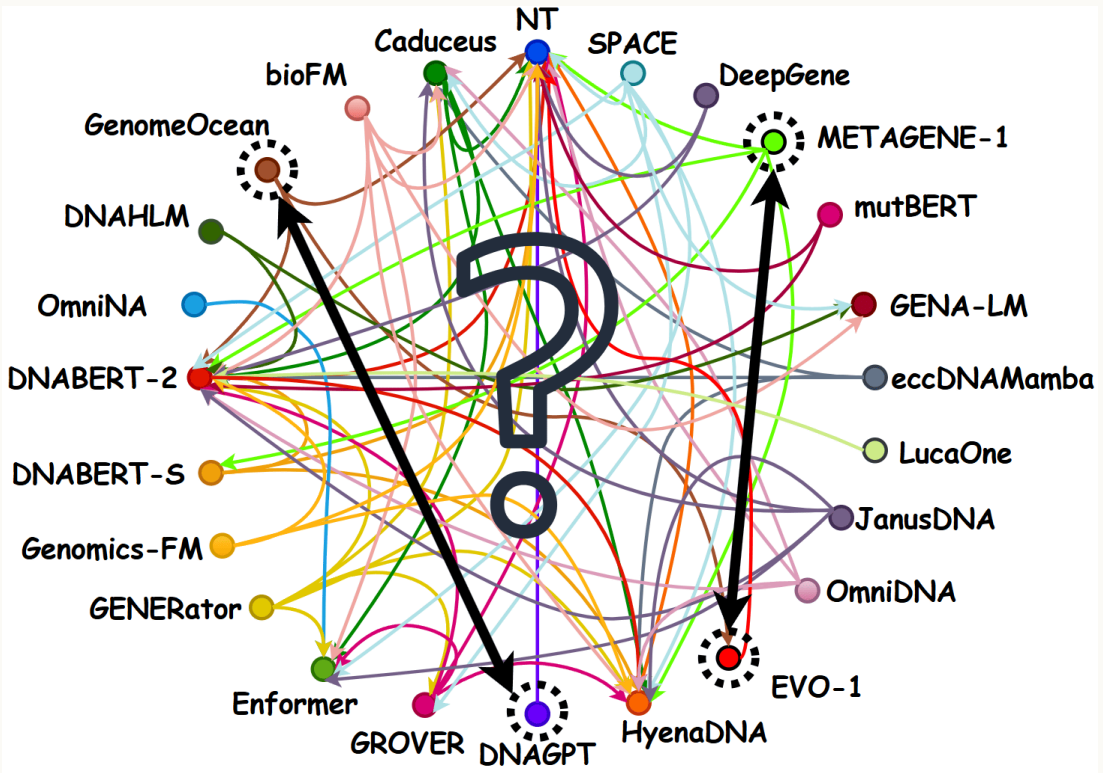
Daria Ledneva · Mikhail Nuridinov · Denis Kuznetsov

Moscow Independent Research Institute of AI · Moscow State Institute of Steel and Alloys

40 models, no common ground.

Genomic foundation models are evaluated on **disjoint benchmarks** under **incompatible protocols**.

The same model is a breakthrough in one paper and underwhelming in another — so claims of superiority simply aren't comparable.



Nodes = published models; an arrow A → B means model A's paper compared against B as a baseline. The graph is sparse and disconnected.

ONE UNIFIED PROTOCOL

Everything, finally comparable.

40	100	13	1·10·∞
foundation models	DNA classification tasks	functional categories	1-shot, 10-shot & full-data regimes

Frozen embeddings → a lightweight logistic probe → **MCC**. Representation quality is isolated and matched across architecture, tokenization and pretraining data — in spirit, an MTEB for genomics.

> Why frozen, not fine-tuning? By design: probing isolates representation quality for a controlled comparison across all 40 models - and rankings stay stable under non-linear probes.

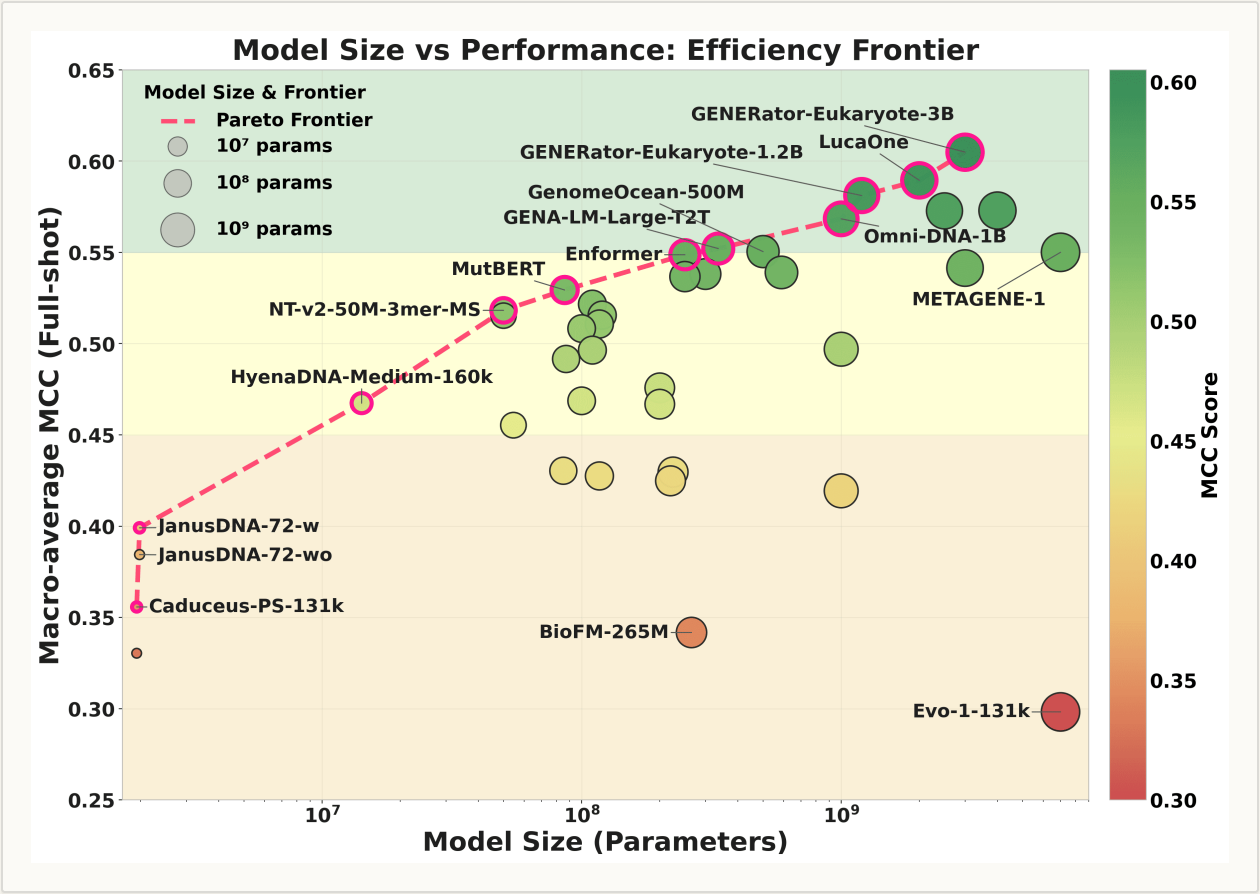
Scale ≠ model choice.

Size correlates with performance ($\rho = 0.565$), yet it doesn't decide model choice.

31 cases where a model **5× smaller** wins

MutBERT (86M) beats **eccDNAMamba** (1B) by **+0.110** MCC

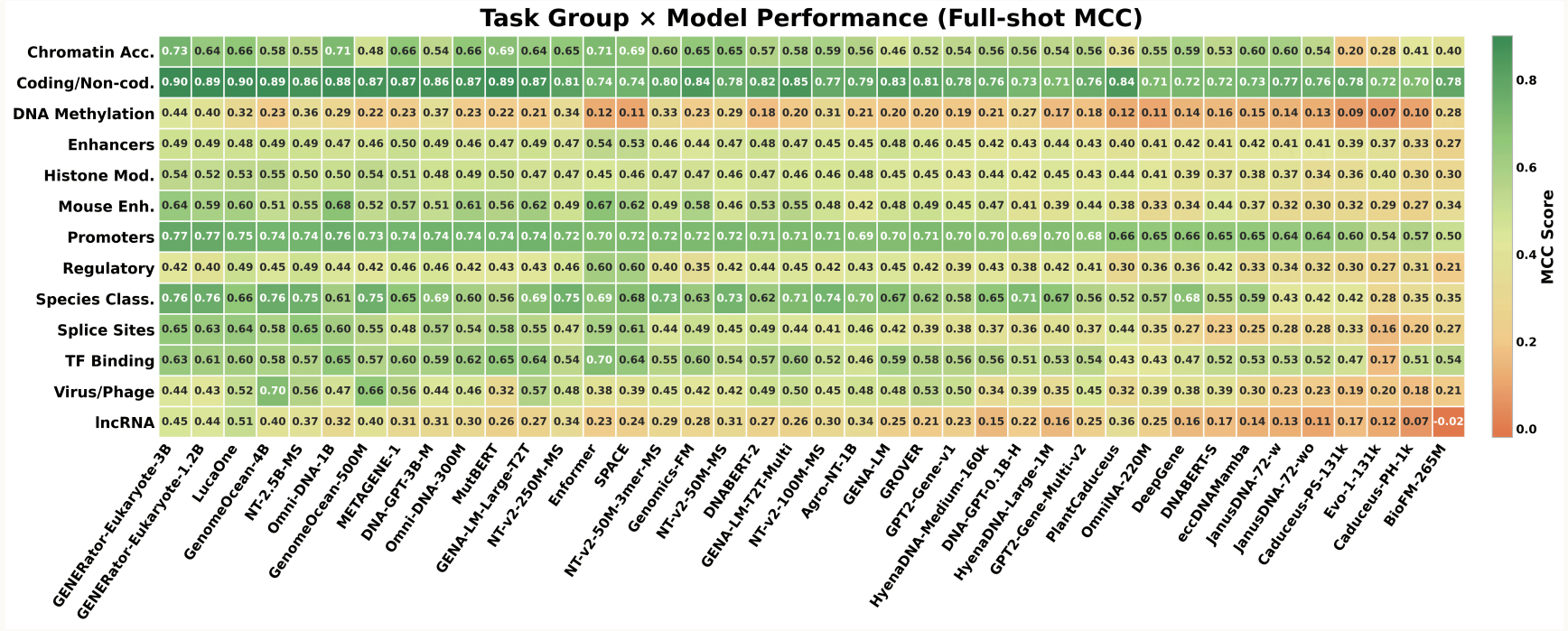
Architecture & pretraining often **outweigh** parameter count



Macro-MCC vs. parameter count. Several large models fall below the Pareto frontier.

Aggregate leaderboards hide the truth.

Rankings vary sharply by category — **Enformer** is only mid-rank overall yet leads on TF binding (0.698), enhancers & regulatory tasks.



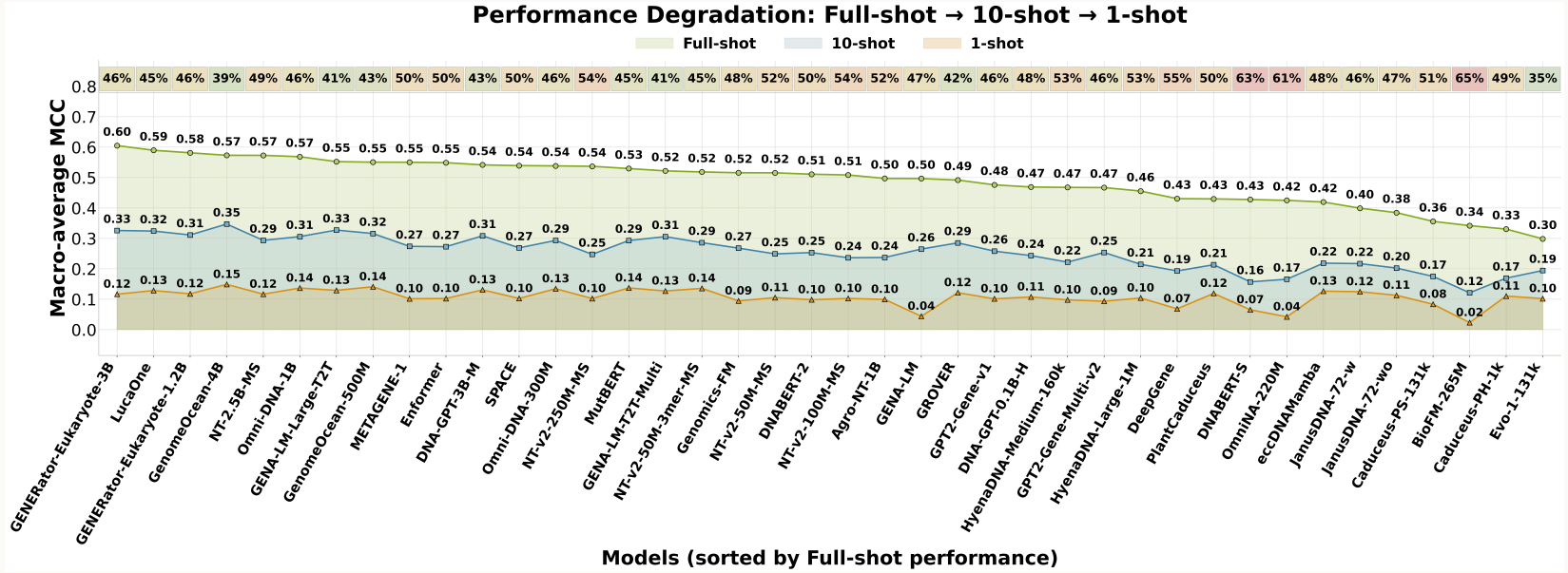
Full-shot MCC per task group × model (green = high, red = low). The structure is the message.

With less data, it gets worse.

Mean macro-MCC collapses from **0.488** (full-shot) to **0.253** (10-shot) to **0.106** (1-shot). The best model reranks in 8 of 13 categories at 10-shot.

−48% · −78%

relative drop at 10- & 1-shot



A small drop signals a *low ceiling*, not robustness - aggregate few-shot ranks mislead.

WHAT TO REMEMBER

Choose models by category, not by leaderboard.

ρ = 0.565

Scale helps on average but is an unreliable guide for any one task.

8 / 13

Categories where the full-data winner is not the few-shot winner.

13 / 13

Per-category model recommendations — from controlled ablations over architecture, tokenization & pretraining.

GENEB replaces single-paper claims with **controlled, category-aware comparison** — a shared reference point for principled progress in genomic ML.

PUBLIC BENCHMARK & LEADERBOARD - RELEASE PLANNED