SynthSAEBench

Evaluating Sparse Autoencoders on Scalable Realistic Synthetic Data

David Chanin, Adrià Garriga-Alonso

Decode

MATS

UCL

SAEs struggle even when the Linear Representation Hypothesis holds perfectly.

We built **SynthSAEBench**: a large-scale synthetic data benchmark with 16k ground-truth features, correlation, hierarchy, and superposition. We trained 5 SAE architectures on it.

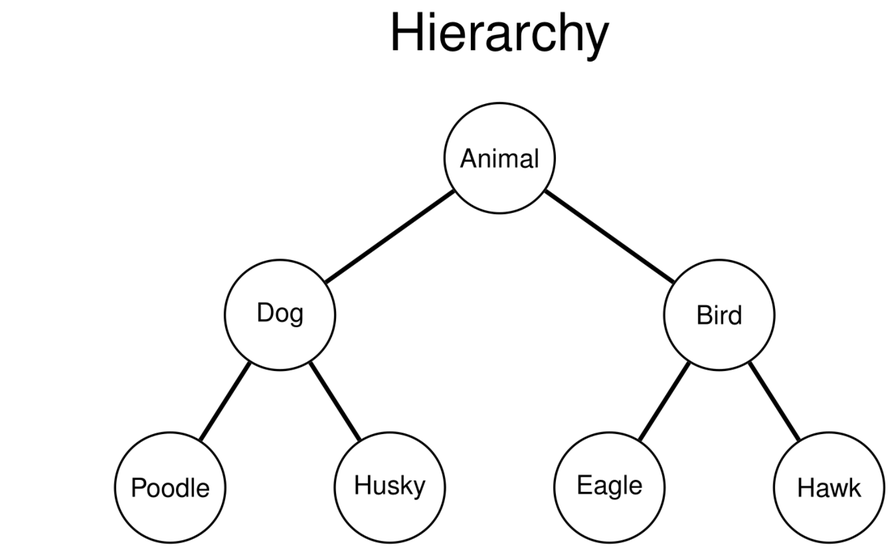
None achieve perfect feature recovery.

Why synthetic data?

SAE eval is stuck between two extremes. LLM benchmarks are noisy, expensive, and have no ground truth, but typical toy models experiments have tiny scale and unrealistic assumptions like perfect independence.

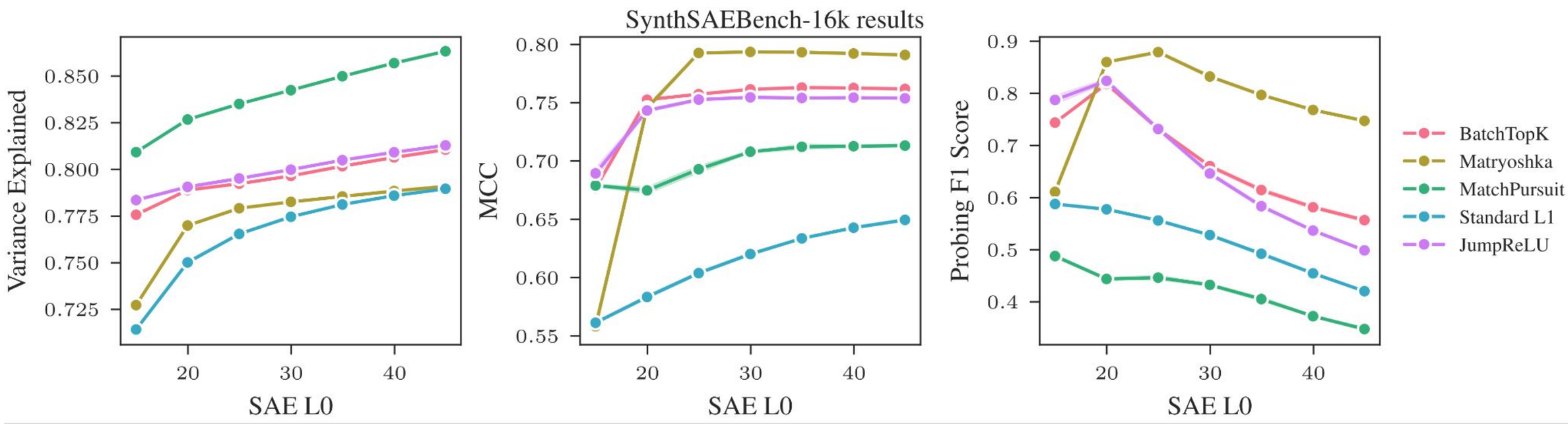
SynthSAEBench sits in the middle: 16k features, realistic properties, ground-truth access, 15 min per SAE on one GPU.

SynthSAEBench feature characteristics



Training on SynthSAEBench reproduces known LLM SAE phenomena

Matryoshka SAEs have great quality despite poor reconstruction, MP-SAEs having great reconstruction but poor quality, and SAEs underperform linear probes



SynthSAEBench is easy to use and extend

Load SynthSAEBench and start building in minutes, now build-in to **SAELens**. Train SAEs on the official benchmark model, or create your own custom synthetic data setups

```
from sae_lens.synthetic import SyntheticModel, SyntheticSAERunner

model = SyntheticModel.from_pretrained(
    "decoderesearch/synth-sae-bench-16k-v1"
)
```

Check out the Paper

SynthSAEBench: Evaluating Sparse Autoencoders on Scalable Realistic Synthetic Data

https://arxiv.org/pdf/2602.14687