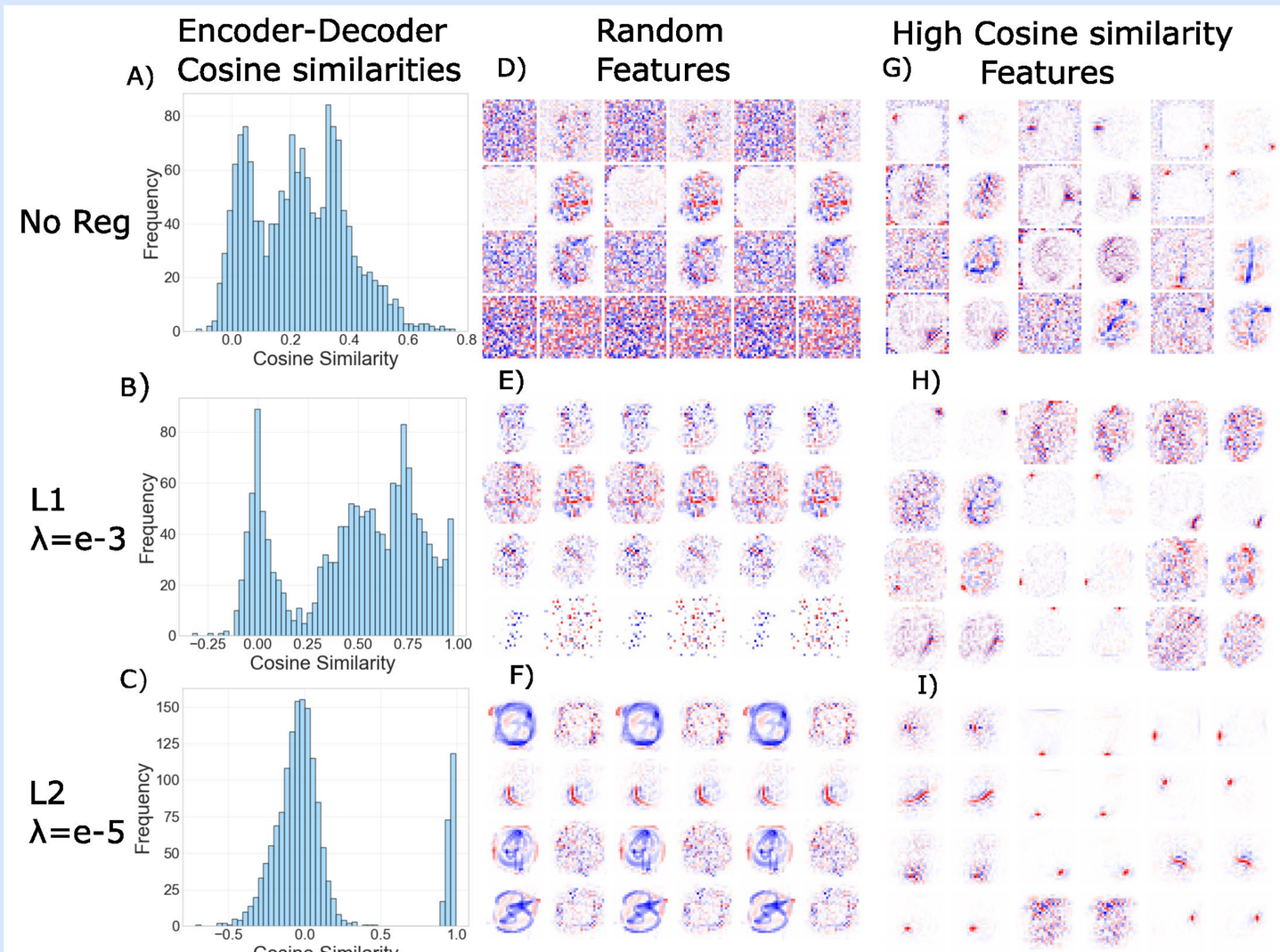


A simple L2 weight penalty makes SAE features reproducible across random seeds and ~2x more steerable

Stable and Steerable Sparse Autoencoders with Weight Regularization

Piotr Jedryszek Oliver M. Crook

Dept. of Biology · Kavli Institute for Nanoscience Discovery · Dept. of Chemistry
University of Oxford
piotr.jedryszek@sjc.ox.ac.uk

1 TL;DR

- Stability: adding a small L2 weight penalty raises cross-seed feature sharedness from <2% to ~35% (TopK, Pythia-70M).
- Steerability: steering success roughly doubles (6.3% → 13.0%, $p<10^{-4}$) while mean auto-interp is unchanged.
- Alignment: under L2, auto-interp scores become a better predictor of steerability (Spearman r: 0.06 → 0.14).

2 Background & Question

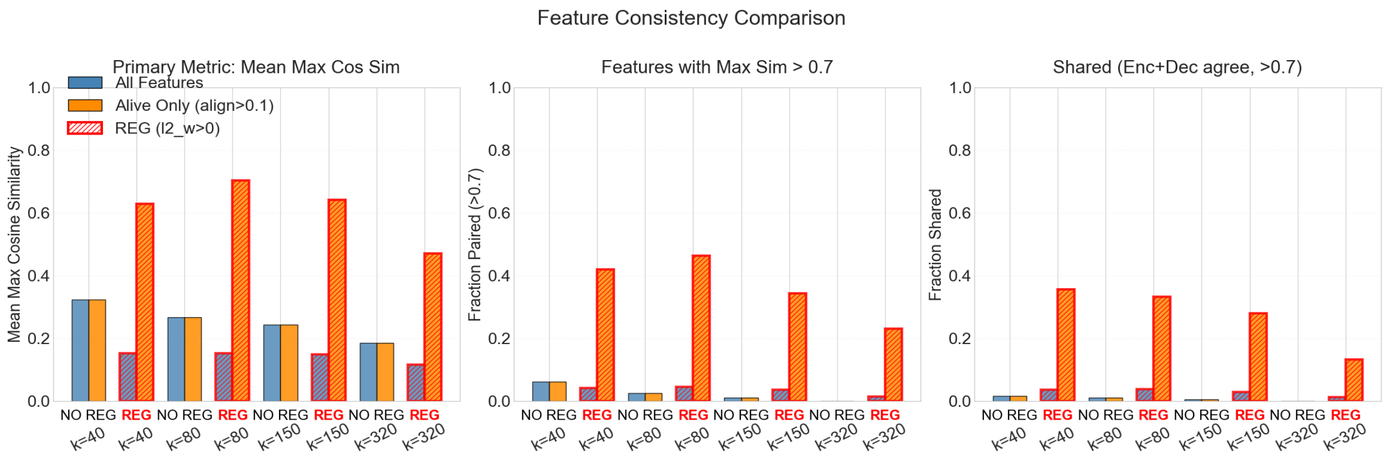
SAEs are underdetermined — different seeds learn different features (Paulo & Belrose, 2025) and results swing with the L0 choice. Classical ML cures underdetermination with regularization.

Q: Can a simple weight penalty make SAEs more stable? What effects does it have on downstream utility and how does it interact with standard SAE design choices?

3 Method

Loss = reconstruction + λ_sparse · sparsity + λ_w (||W_enc||ᵖ + ||W_dec||ᵖ), p ∈ {1, 2}

- Tested on the MNIST dataset as a toy model and on Pythia-70M-deduped (layer-3 residual) via SAEBench; TopK / BatchTopK / Matryoshka, k ∈ {40,80,150,320}.
- Consistency: Hungarian-matched decoder cosine over 3 seeds; a feature is 'alive' if enc–dec cosine > 0.1.
- Steering: inject α·d_i into the residual stream during generation; auto-interp explanations scored by a GPT-5.1 judge.

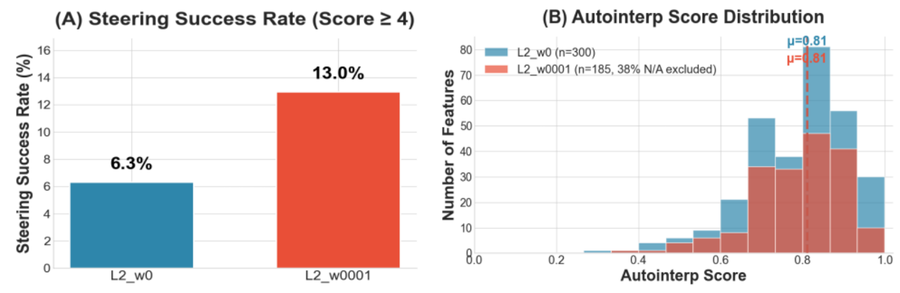


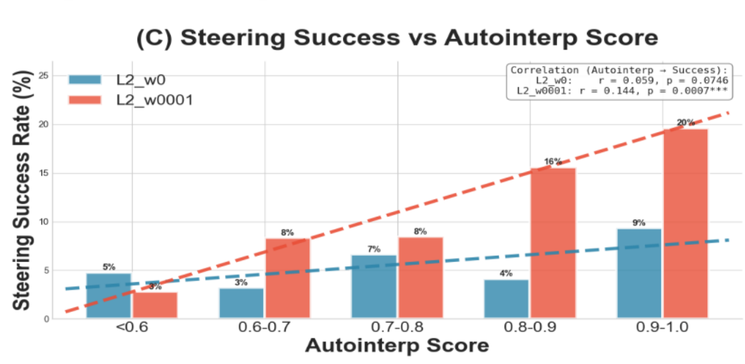
Cross-seed consistency (Pythia-70M TopK). L2 (red) sharply increases mean max-cosine, fraction paired, and strictly shared features — especially among alive features — across all sparsity levels k.

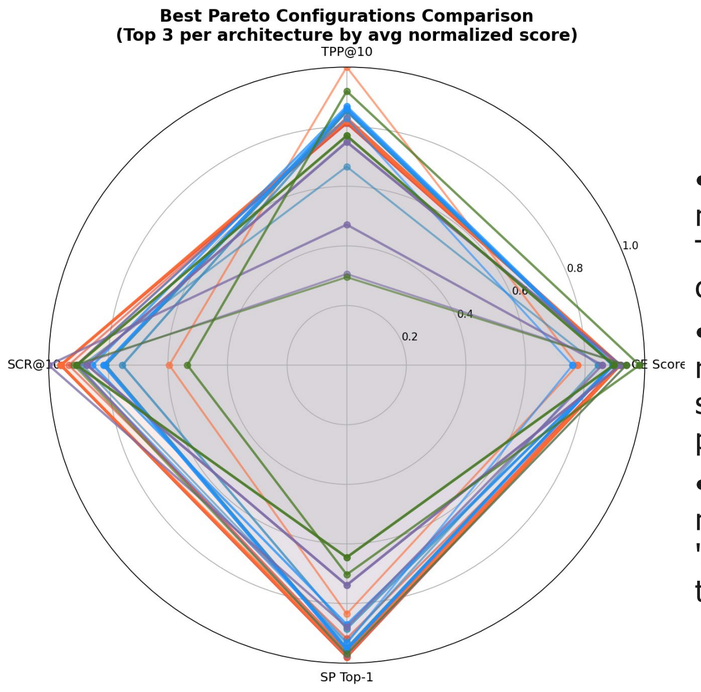
4 Results

Steering doubles

- L2 lifts steering success 6.3% → 13.0%
- Mean auto-interp score is unchanged.
- The auto-interp and steering link strengthens (r: 0.06 → 0.14). Explanations better predict what a feature does.







Competitive on SAEBench

- Radar = normalized SAEBench metrics (TPP@10, CE, SCR@10, SP Top-1); each line is a top Pareto config (full legend in paper).
- Regularized configs frequently reach the outer frontier — stability & steering gains don't cost benchmark performance.
- On MNIST, L2 + tied init + unit-norm decoder gives the cleanest 'aligned core' (see key figure at the top).

5 Limitations & Discussion

- L2 collapses ~90% of latents — yet >1,500 alive features remain (3× the 512-dim residual stream).
- Likely partly implicit dictionary-size selection; converges with MDL and attribution-distillation (Martin-Linares & Ling, 2025) on compact feature sets.
- Limits: Pythia-70M only; steering tested at k=40; single LLM judge.

Takeaway: a simple weight penalty is a cheap lever toward stable, controllable SAEs.

PAPER

CODE

Mechanistic Interpretability Workshop ICML 2026 | https://arxiv.org/abs/2603.04198 | https://github.com/oxPJ/sae-weight-regularization