DMSAEs distill a core of stable encoder directions that can be transferred to new SAE trainings across sparsity levels.

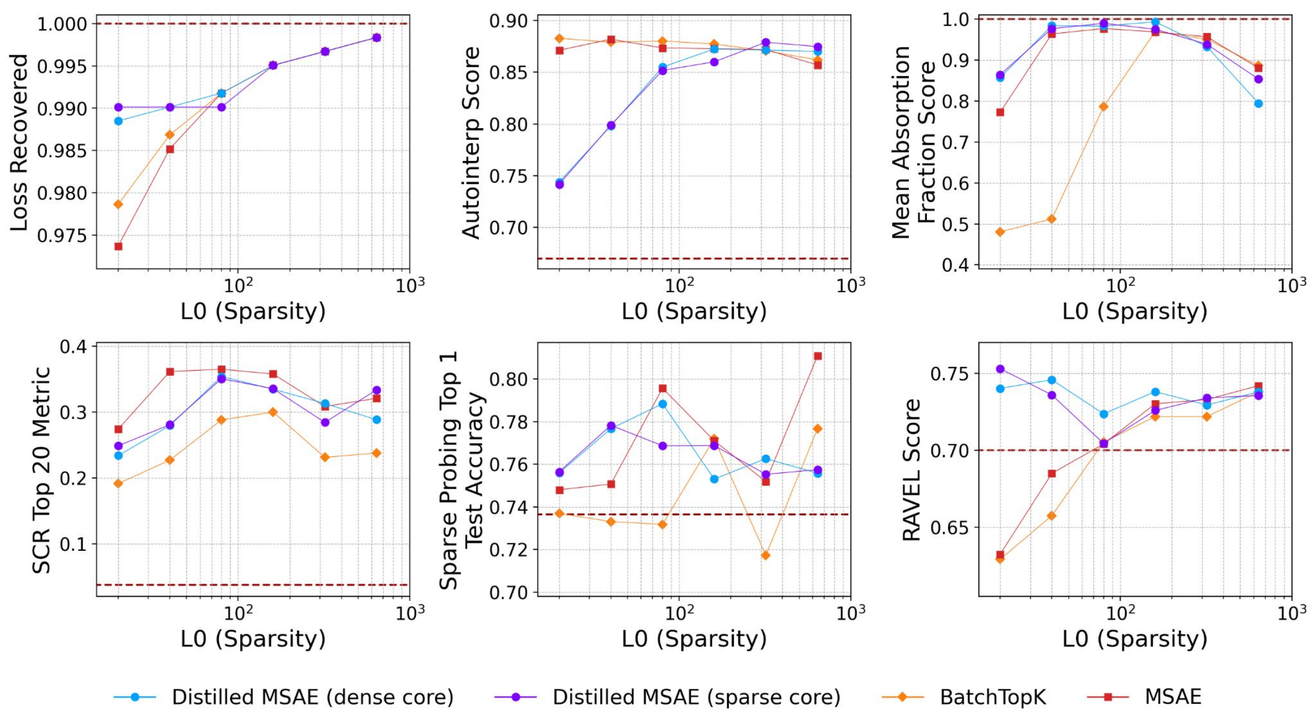
Iterative Distillation for Feature-Consistent Sparse Autoencoders

SAEs can discover interpretable features, but learned latents are often redundant, unstable across runs, and inconsistent across sparsity levels. Can we distill a compact set of features that remains useful after retraining?

Results

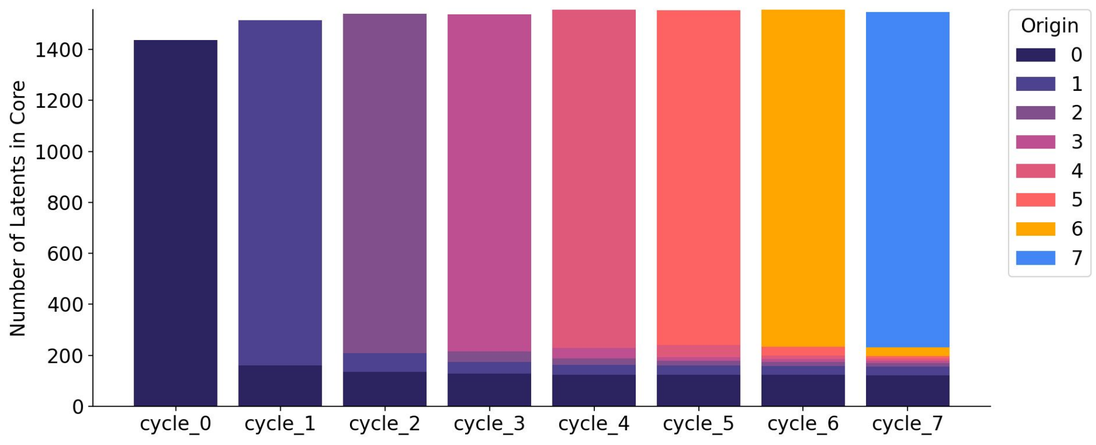
Improves SAEBench Metrics:

DMSAEs match or improve several SAEBench metrics across sparsity targets, with strongest gains in absorption and RAVEL.



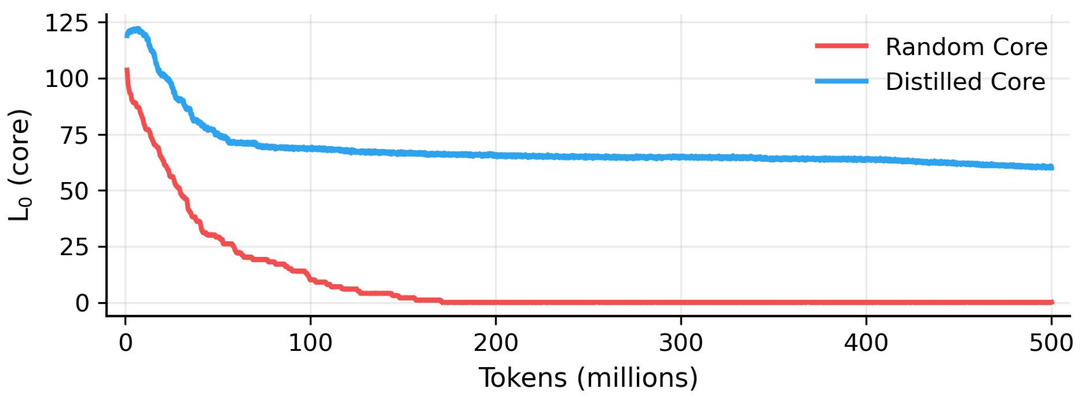
Distills a Compact Reusable Core:

7 cycles over 500M tokens each distilled a 197-latent core from a 65,536-feature MSAE.



Core is Reused After Retraining:

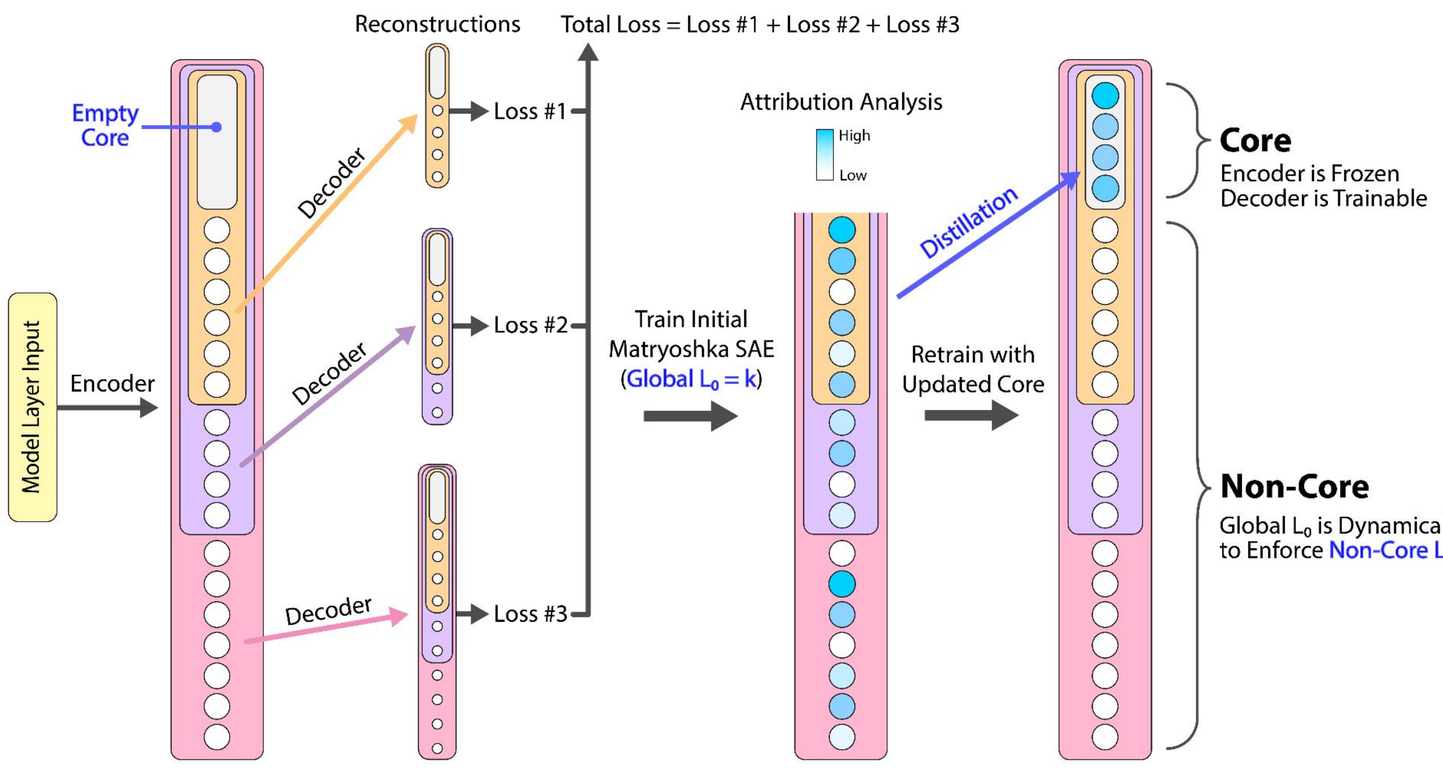
Random same-size cores are ignored during training, while distilled cores continue to activate.



Methods

1. Train MSAE
2. Score latents by gradient × activation to next-token loss
3. Keep top latents covering 90% attribution
4. Restart with only encoder directions frozen

Distilled Matryoshka Sparse Autoencoder



Limitation: Core discovery is expensive with seven full train-and-select cycles were used here. The final distilled core can then be reused at standard single-run training cost.

PAPER

Cristina P. Martin-Linares[1], and Jonathan P. Ling[2]
[1]Dept. of Biomedical Eng.; [2]Dept. of Pathology, Johns Hopkins

Mechanistic Interpretability Workshop, NeurIPS 2025