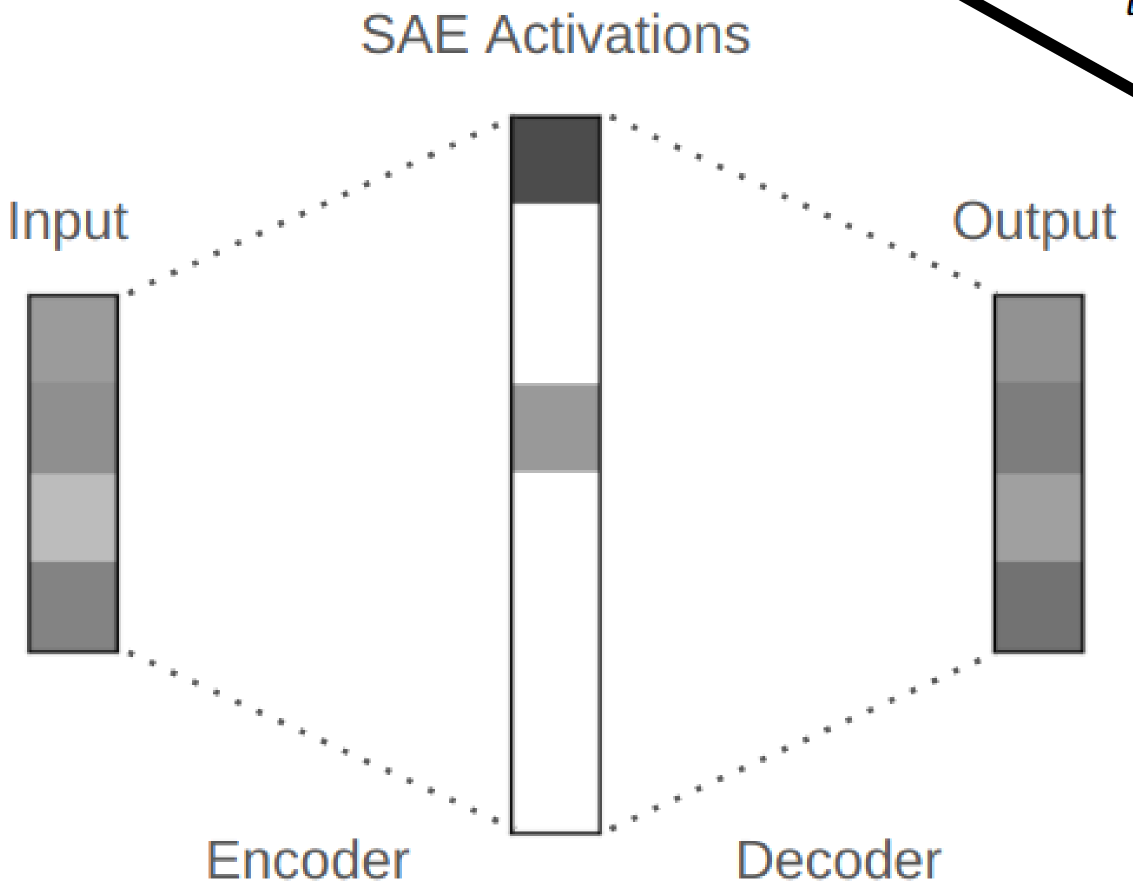
SAE orthogonalization

Anton Korznikov

Skoltech

Sparse Autoencoders

Superposition hypothesis: $\mathbf{x} \approx \mathbf{x}_0 + \sum_{i=1}^{M} f_i(\mathbf{x})\mathbf{d}_i,$



embedding vector of token [Jordan] in residual stream in 8th layer of GPT2

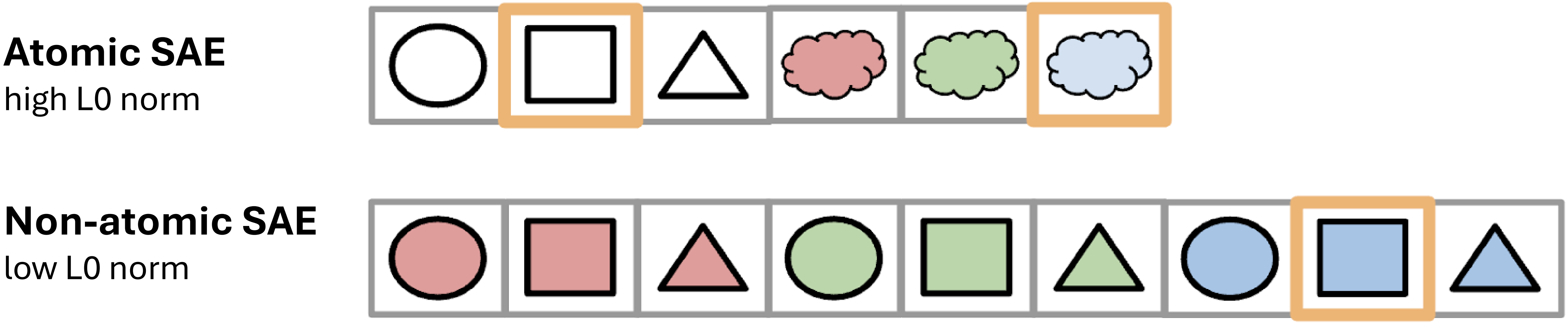
$$\mathbf{f}(\mathbf{x}) := \mathrm{ReLU}\left(\mathbf{W}_{\mathrm{enc}}\left(\mathbf{x} - \mathbf{b}_{\mathrm{dec}}\right) + \mathbf{b}_{\mathrm{enc}}\right)$$

$$\hat{\mathbf{x}}(\mathbf{f}) := \mathbf{W}_{\mathrm{dec}}\mathbf{f} + \mathbf{b}_{\mathrm{dec}}$$

$$\mathcal{L}(\mathbf{x}) := \|\mathbf{x} - \hat{\mathbf{x}}(\mathbf{f}(\mathbf{x}))\|_2^2 + \lambda \, \|\mathbf{f}(\mathbf{x})\|_1 \, .$$

SAE features are not atomic. Theoretical evidence.

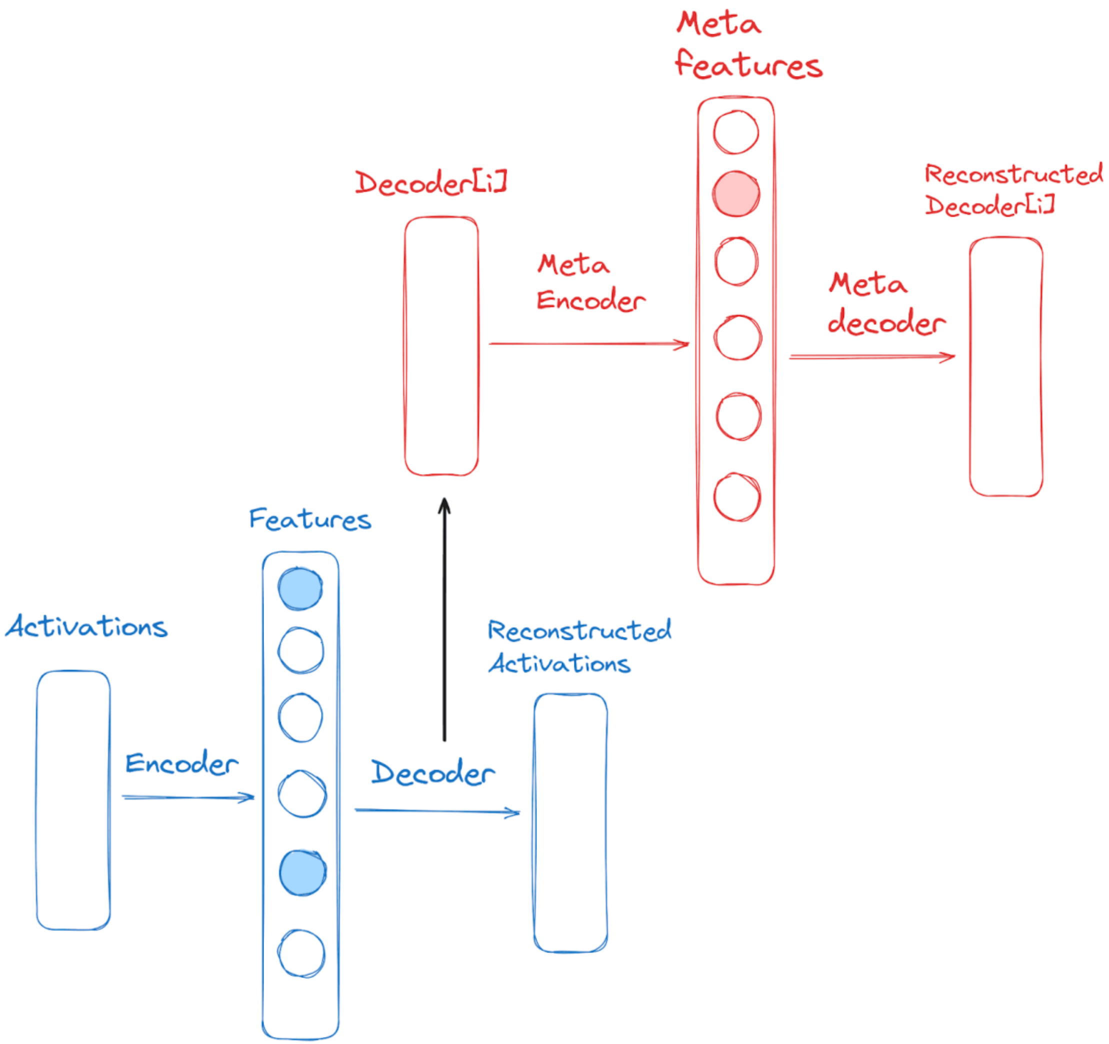
L0 regularization conflicts with reconstruction loss.



SAE features are not atomic. Practical evidence № 1

Meta SAE trained on decoder features shows that many features can be further decomposed.

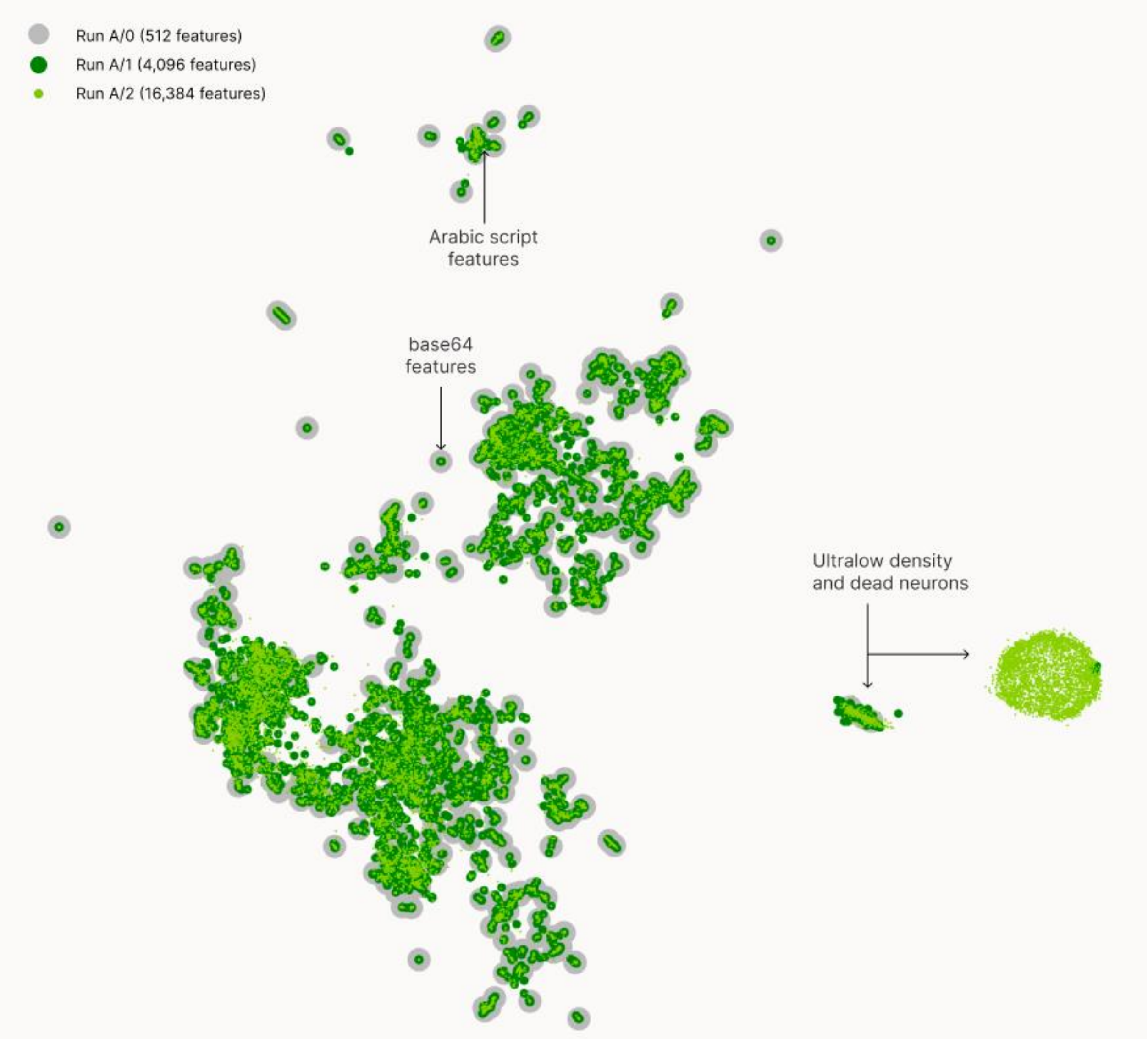
Example:
SAE feature that activates only on notion of Albert Einstein can be decomposed as sum of features “Physics”, “German”, “Noble prize”



SAE features are not atomic. Practical evidence № 2

SAE features tend to create clusters of highly similar features.

It will take place for non atomic features, because they can share some common atomic features inside each other.

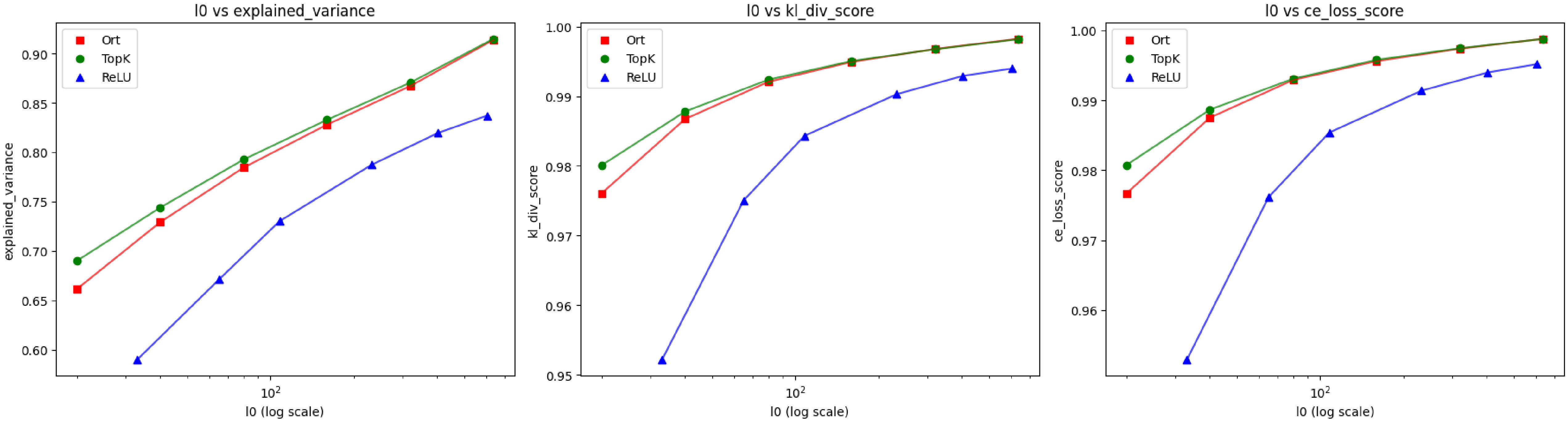


Problem: current SAEs learn non-atomic features

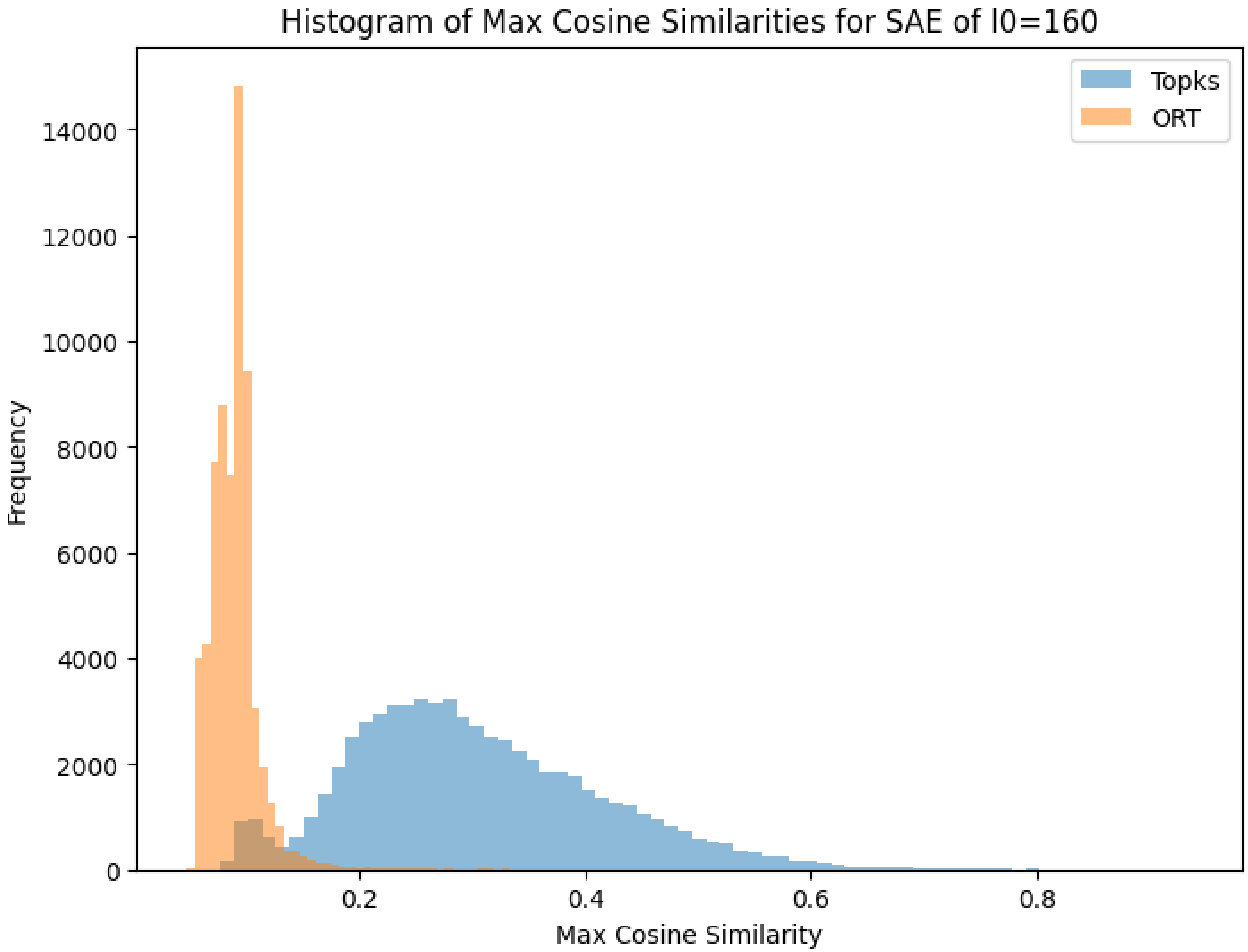
Hypothesis: atomic features are almost orthogonal

Solution: add orthogonal regularization on features

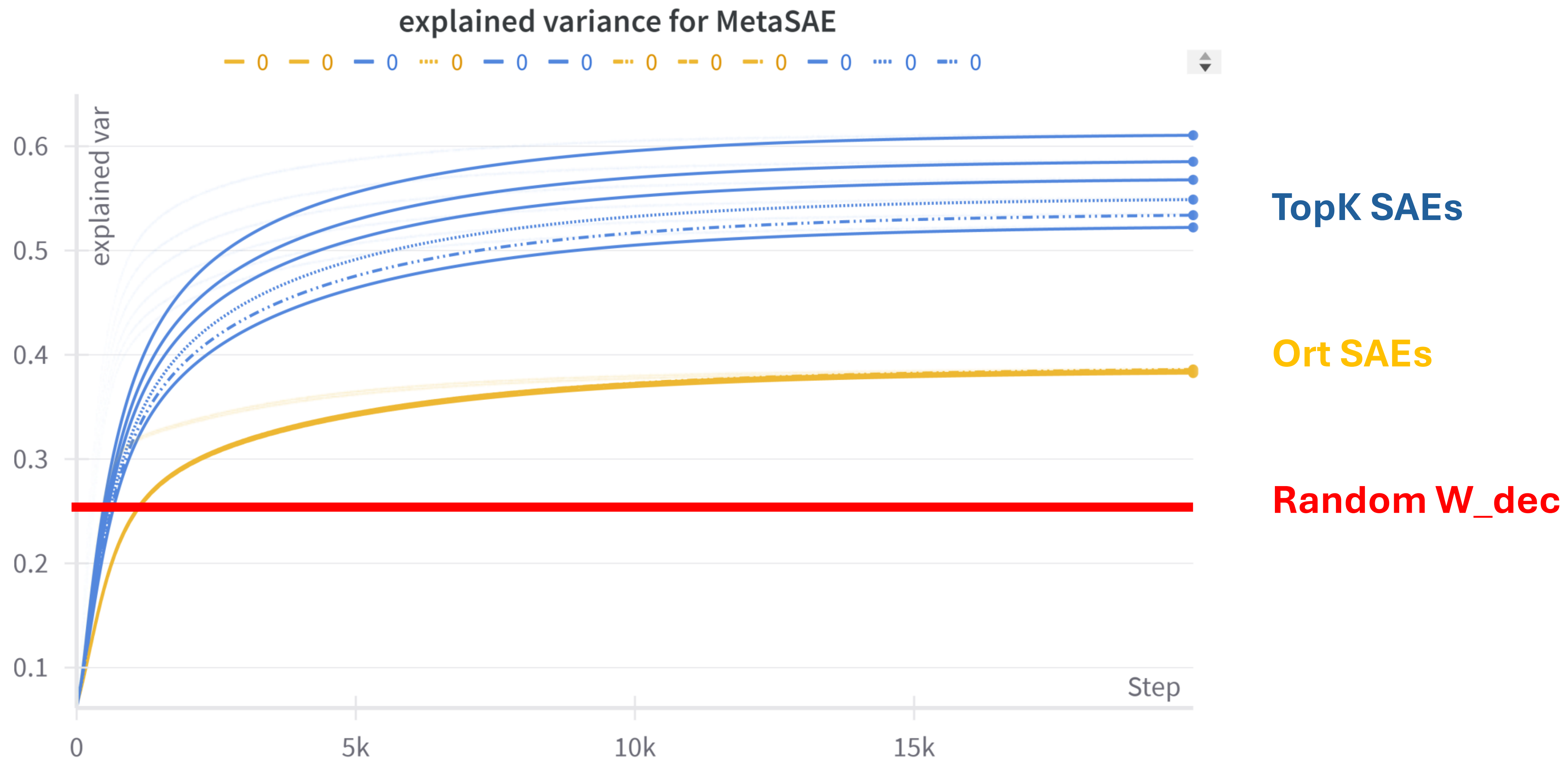
Experiments. Core metrics.



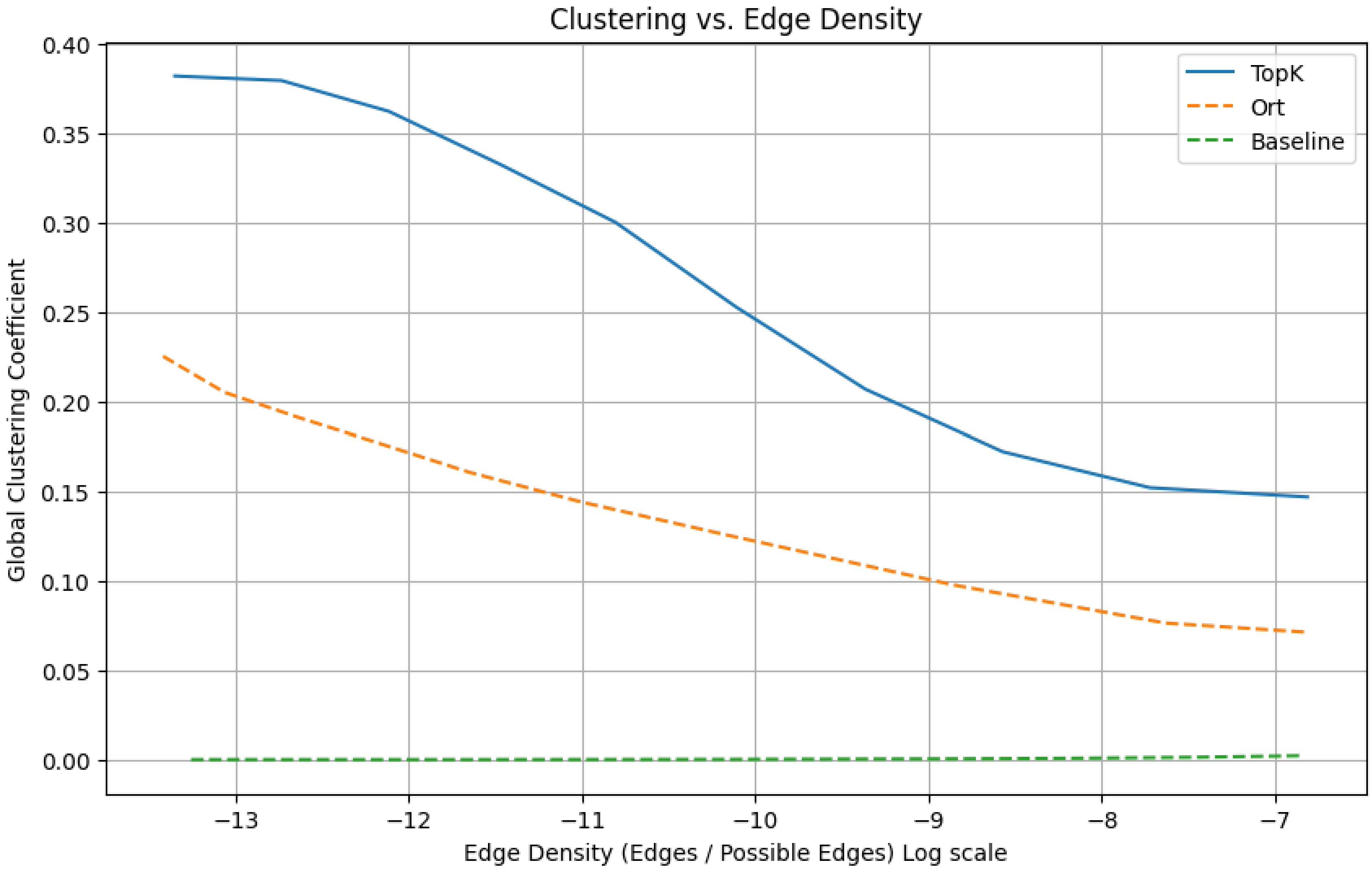
Experiments. Core metrics.

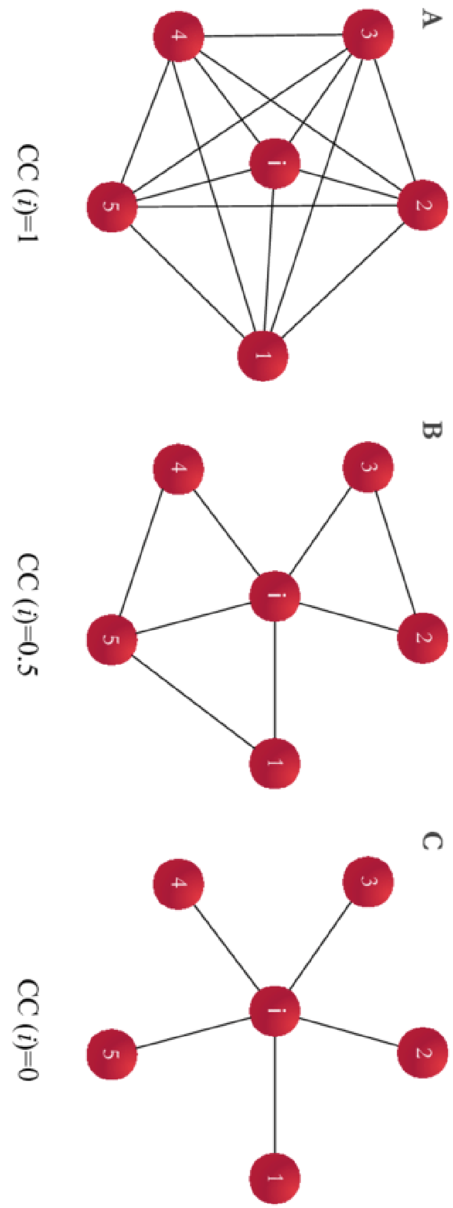


Experiments. Atomicity metrics.



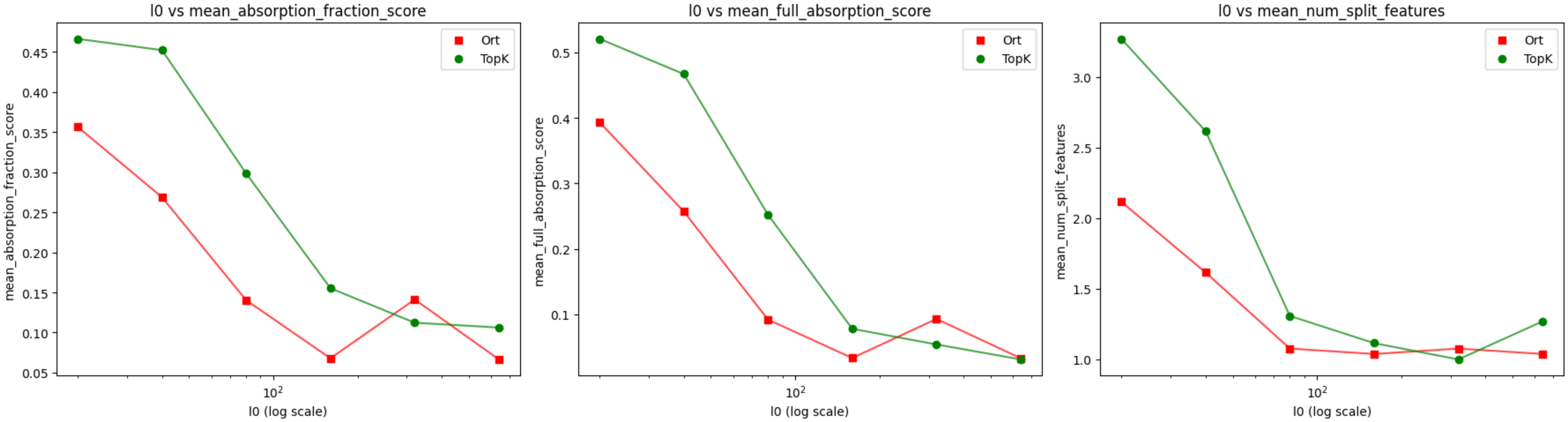
Experiments. Clustering metrics.





SaeBench results. Feature absorption.

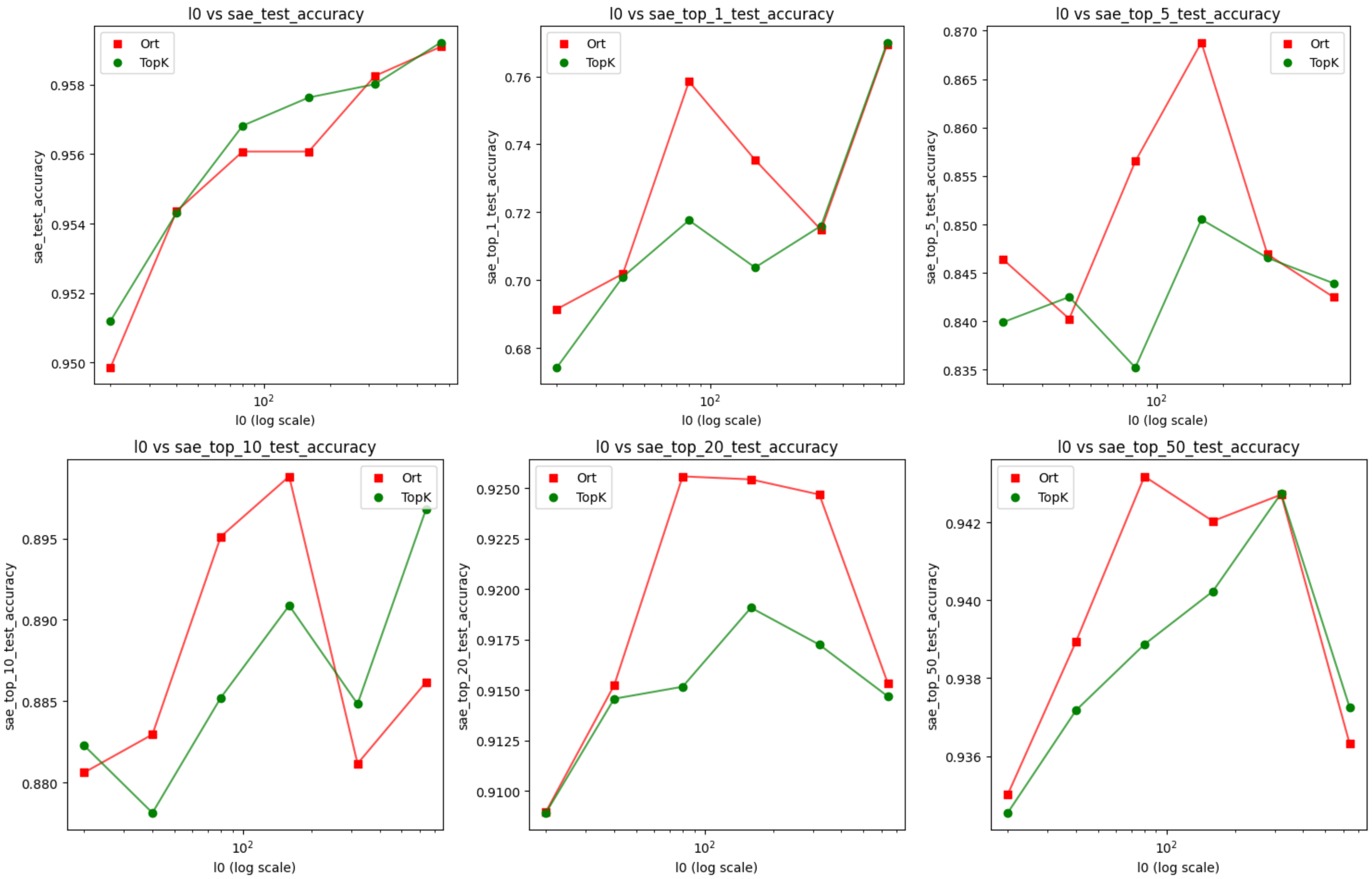
Feature absorption is a phenomenon where sparsity incentivizes SAEs to learn undesirable feature representations. This occurs with hierarchical concepts where A implies B (e.g., pig implies mammal, or red implies color)—rather than learning separate latents for both concepts, the SAE is incentivized to learn a latent for A and a latent for “B except A” to improve sparsity.



SaeBench results. Sparse probing.

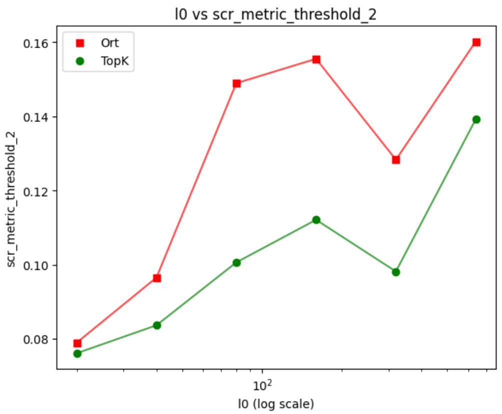
Sparse probing evaluates whether SAEs isolates prespecified concepts.

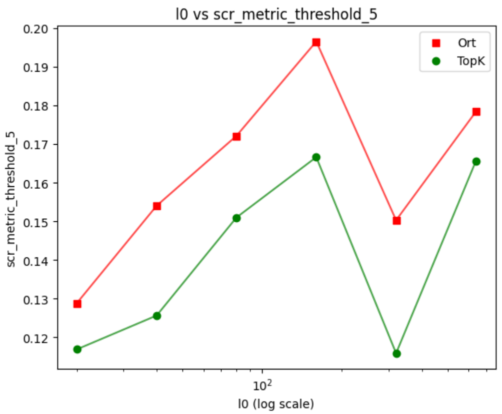
For each concept (e.g., sentiment), we identify the k most relevant latents by comparing their mean activations on positive versus negative examples and train a linear probe on top k latents.

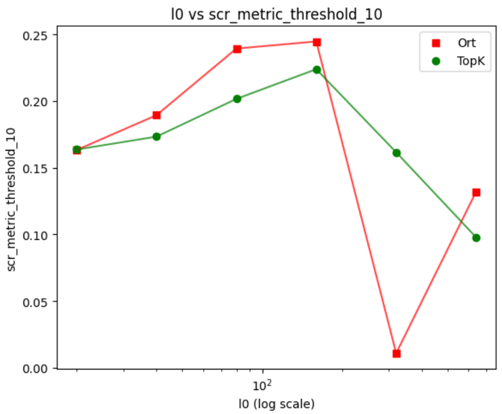


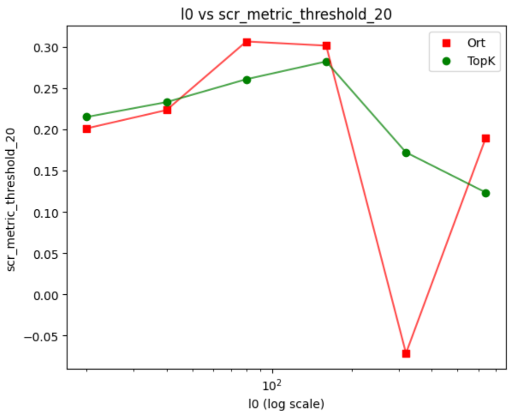
SaeBench results. Spurious Correlation Removal.

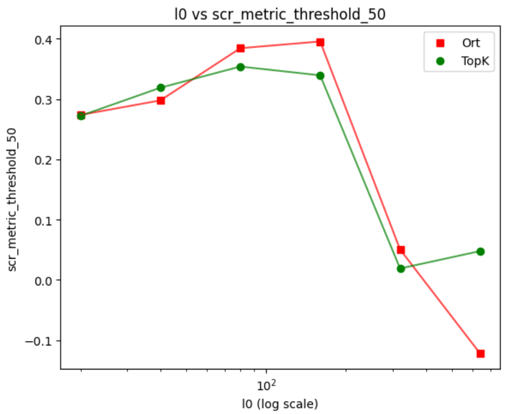
Starting with a biased linear probe classifier that has learned both intended signals (e.g., profession) and spurious correlations (e.g., gender), we measure how effectively zero-ablating a small number of SAE latents can remove the unwanted correlation from the SAE's output. If these latents cleanly isolate the spurious concept, removing them should significantly improve the classifier's accuracy on the intended signal.





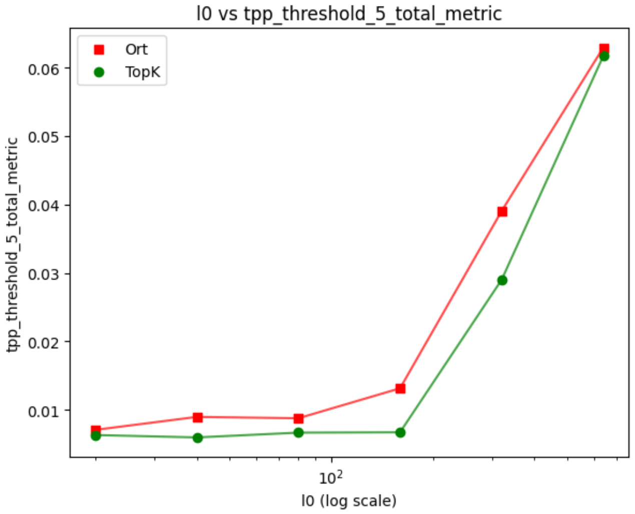


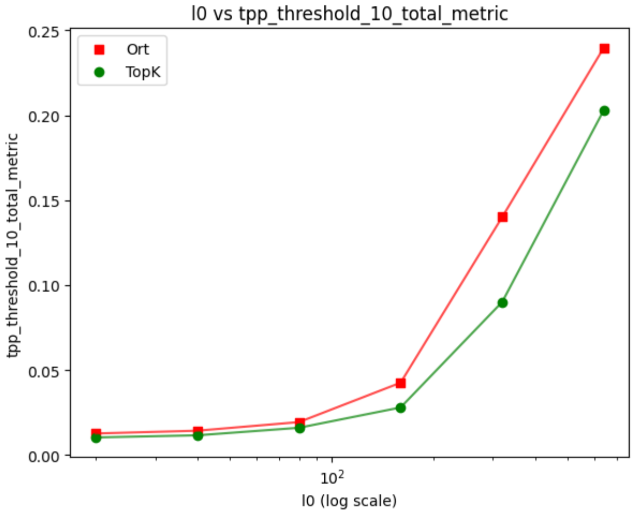


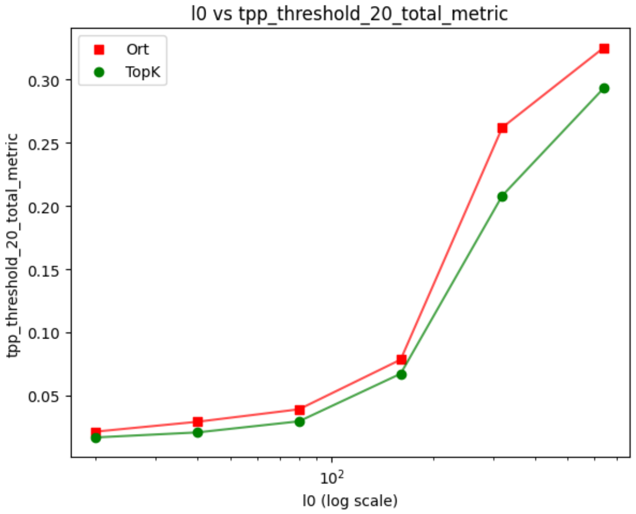


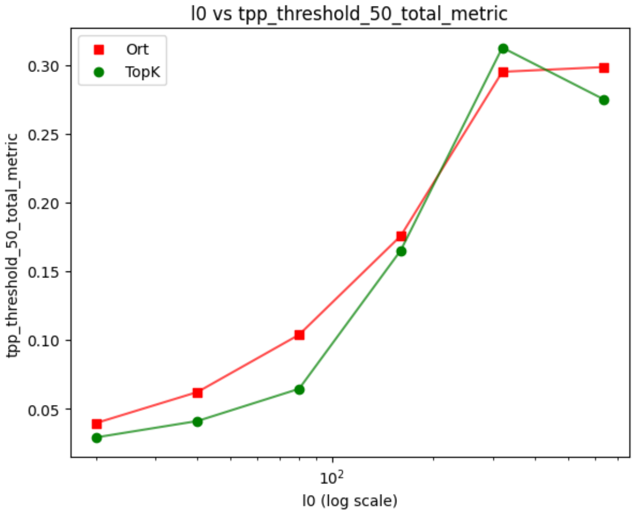
SaeBench results. Targeted Probe Perturbation.

Targeted Probe Perturbation (TPP) generalizes this approach to multi-class settings. For each class, we train binary classification probes and identify its most relevant latents. We then measure how zero-ablating these latents affects probe accuracy across all classes. A high TPP score indicates that concepts are captured by distinct sets of latents—ablating latents relevant to one class should primarily degrade that class's probe accuracy while leaving other class probes unaffected.





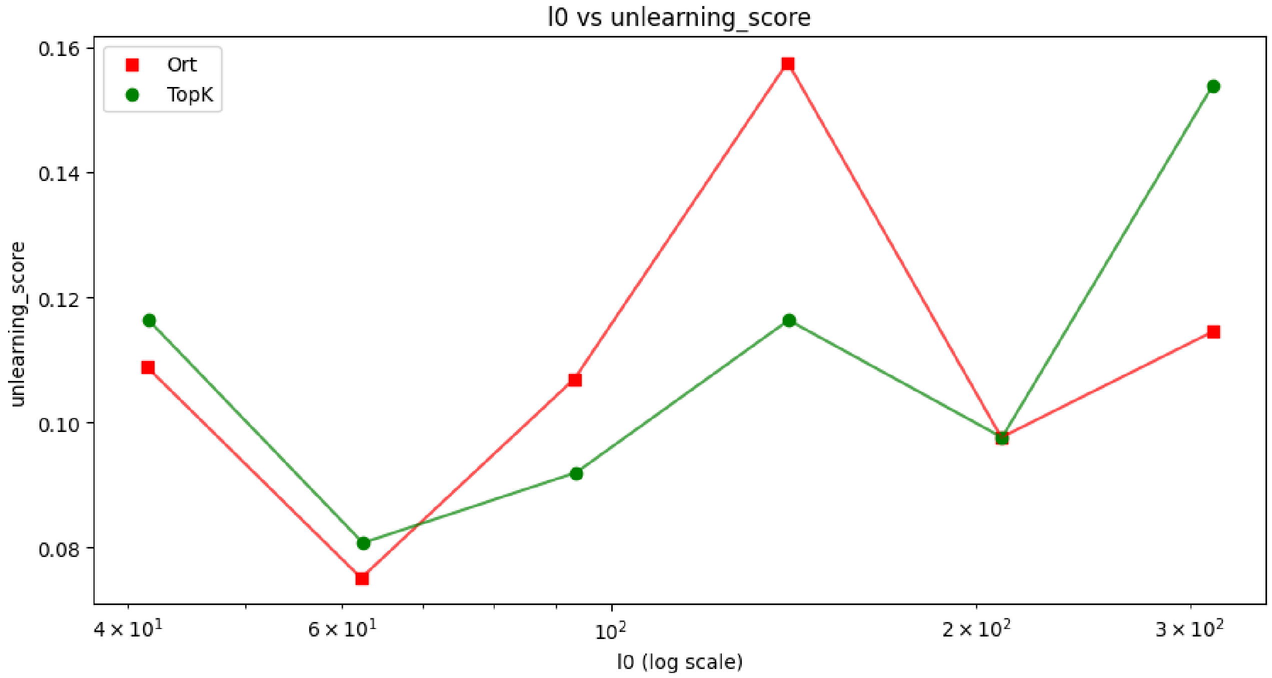




SaeBench results. Unlearning.

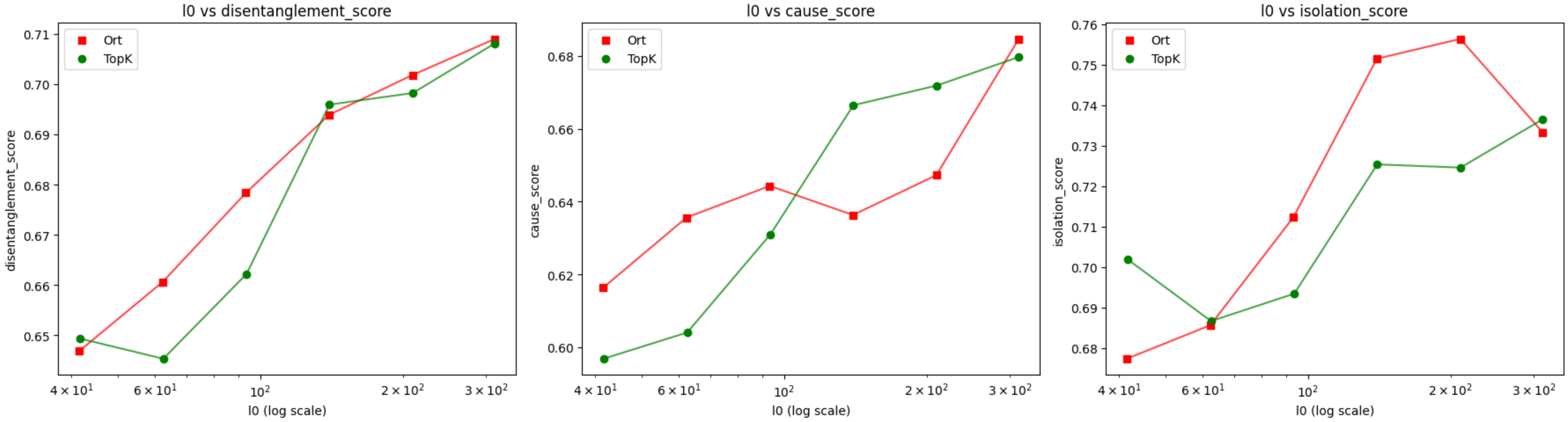
We identify relevant latents by comparing their activation frequencies between a forget set (biology-related text in the WMDP-bio corpus) and retain set (WikiText), then clamp these latents to negative values whenever they activate.

We build on their methodology and report an unlearning score for each individual SAE, measuring unlearning success via degraded accuracy on WMDP-bio test questions while using MMLU categories to verify retained capabilities. Models that achieve strong unlearning of the target domain with minimal side effects on other domains score higher.



SaeBench results. RAVEL.

RAVEL evaluates whether targeted interventions on SAE latents can selectively change a model's predictions for specific attributes without unintended side effects—for instance, making the model believe Paris is in Japan while preserving the knowledge that the language spoken remains French.


l0 vs disentanglement_score
disentanglement_score
🟥 Ort
● TopK
4×10^1
6×10^1
10^2
2×10^2
3×10^2
l0 (log scale)
l0 vs cause_score
cause_score
🟥 Ort
● TopK
4×10^1
6×10^1
10^2
2×10^2
3×10^2
l0 (log scale)
l0 vs isolation_score
isolation_score
🟥 Ort
● TopK
4×10^1
6×10^1
10^2
2×10^2
3×10^2
l0 (log scale)