Sparse Concept Anchoring

for Interpretable and Controllable Neural Representations

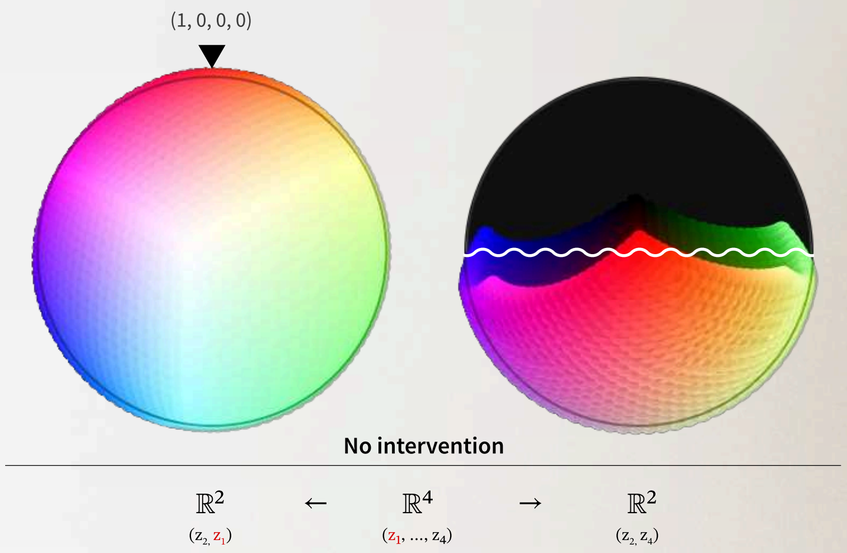
Sandy Fraser & Patryk Wielopolski, independent

Current interpretability methods attempt to discover safety-critical concept locations post-hoc. We anchor them during training, enabling reliable intervention with no search and **<0.1%** labeled data.

Single Anchor

Anchor only concepts of interest

Few & noisy labels suffice

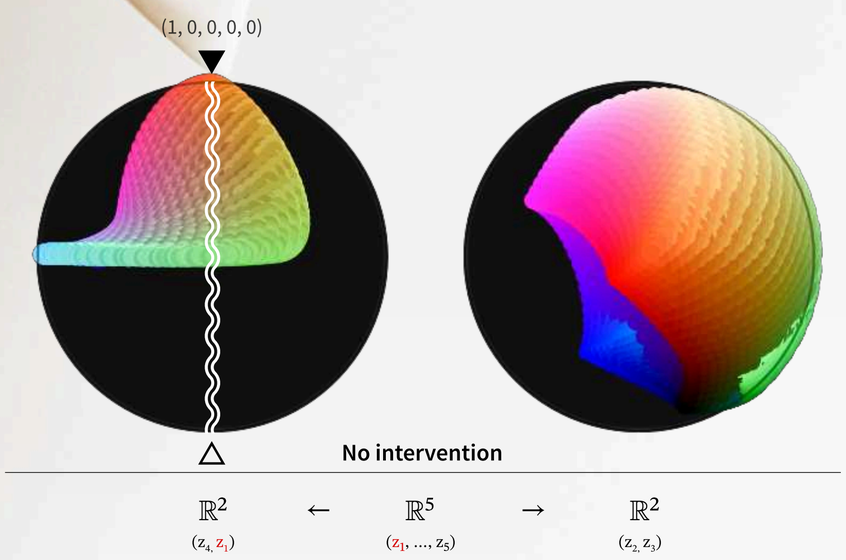


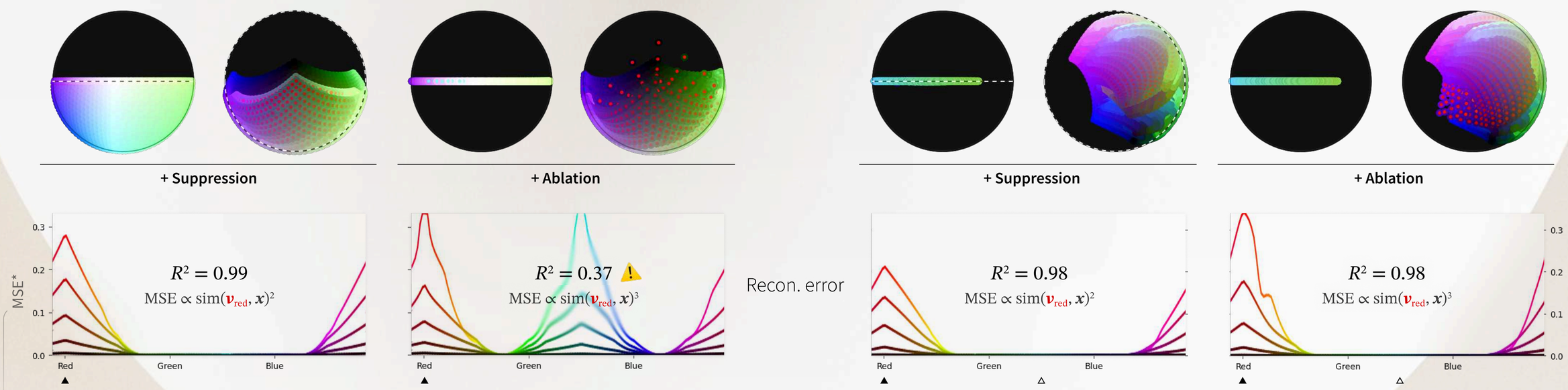
Axis-aligned projections of latent space

Isolation w/ Repulsion

Clear subspaces for ablation

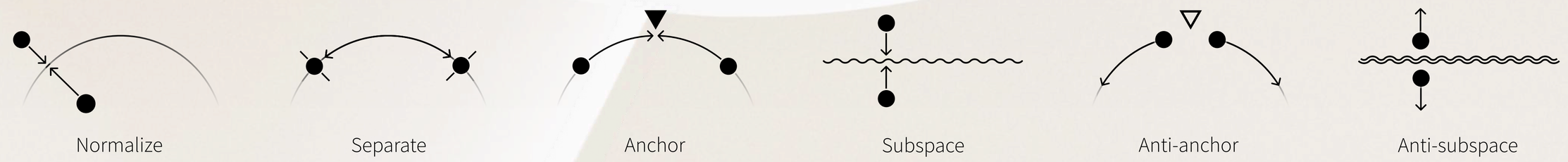
No extra labels needed





Targeted inductive biases on activations:

$$\mathcal{L}_{\text{total}}(\cdot) = \mathcal{L}_{\text{task}}(\cdot) + \underline{\mathcal{L}_{\text{structural}}(\hat{\boldsymbol{z}})} + \underline{\mathcal{L}_{\text{concept}}(\hat{\boldsymbol{z}}, \ell_{\mathcal{C}})}$$



Structural constraints

provide a geometric basis for latent representations. *Normalize* places activations on the hypersphere, while *separate* prevents excessive clustering (pairwise). These constraints create structure that supports concept organization and intervention.

Attractive regularizers

pull rare labeled samples to predetermined locations. *Anchor* positions simple linear concepts, while *subspace* collects multidimensional and cyclic concepts. The network is free to learn optimal representations for others.

Repulsive regularizers

push all samples away from dimensions intended for ablation. Weights vary to a schedule: strong early in training to clear regions while the space is malleable, and weak later to allow attractive terms to dominate.

arXiv:2512.12469

*** Task:** Color autoencoder on RGB color space. Anchored concepts: red (linear), and in some experiments, hue (cyclic). Bottleneck dim ∈ [4,5].