From Flat to Hierarchical: Extracting Sparse Representations with Matching Pursuit

NeurIPS 2025

Valérie Costa*[1] Thomas Fel*[2,3] Ekdeep Singh Lubana*[2,4] Bahareh Tolooshams†[5,6] Demba Ba†[2,3]

[1] EPFL [2] Harvard University [3] Kempner Institute [4] NTT Research [5] University of Alberta [6] amii

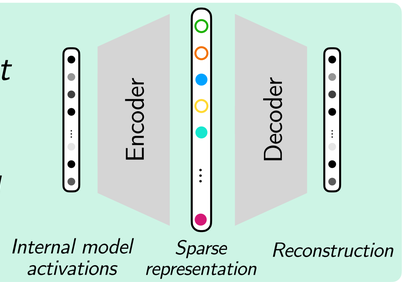
Nectar spotlight AI/CRV 2026

Motivated by the linear representation hypothesis (LRH), **sparse autoencoders** (SAEs) have become a popular tool for **mechanistic interpretability**.

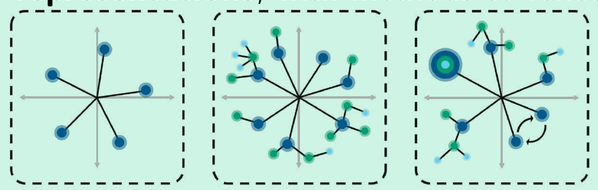
Can SAEs recover features beyond the LRH?

We propose **MP-SAE** to extract hierarchical features.

SAEs to extract interpretable features from internal model activations.

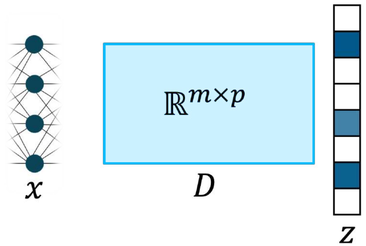


Interpretability should begin with **geometry of model representations**, with methods to reflect those **assumptions**.



Geometry of Model Representation ?

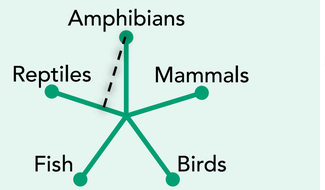
Method: Conditional Orthogonality and Matching Pursuit Sparse Autoencoder (MP-SAE)



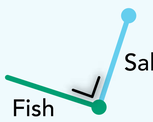
Linear Representation Hypothesis

1. Overcompleteness $p \gg m$
2. Quasi-orthogonality $\max_{i\neq j} |D_i^\top D_j| = \mu(D) \leq \epsilon$
3. Sparsity $\|z\|_0 \leq k \ll p$

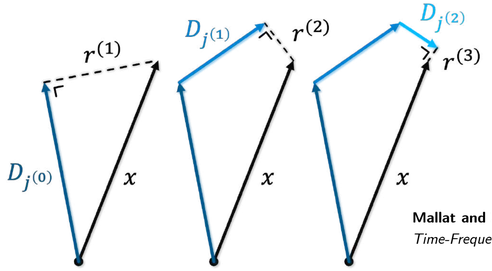
Quasi-Orthogonality

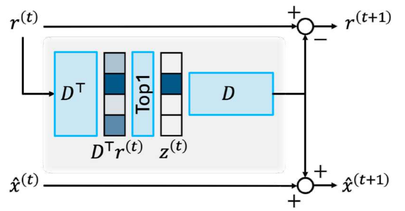


Conditional Orthogonality



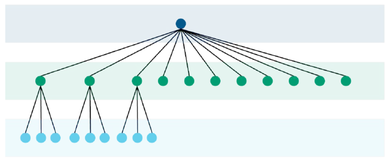
Park et al. *The Geometry of Categorical and Hierarchical Concepts in Large Language Models* 2024.





Mallat and Zhang. *Matching Pursuits with Time-Frequency Dictionaries 1993.*

MP-SAE Recovers Hierarchical Correlated Features



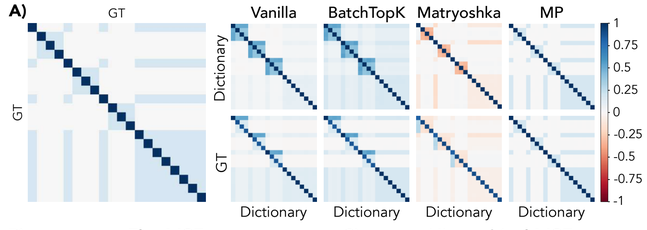
Bussmann et al. *Learning Multi-Level Features with Matryoshka Sparse Autoencoders 2025.*

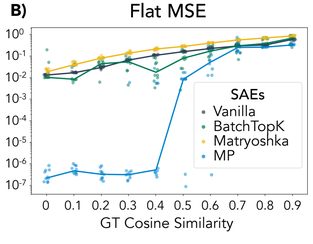
Hierarchical Generative Process

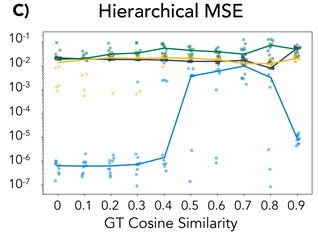
$$x = z_p \cdot D_p + z_c \cdot D_c \quad \text{with}$$

$$\mathbb{P}(z_p > 0 \mid z_c > 0) = 1$$

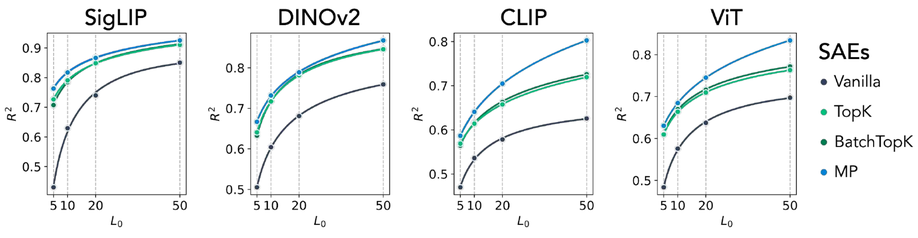
$$\mathbb{P}(z_c > 0 \mid z_p > 0) < 1$$



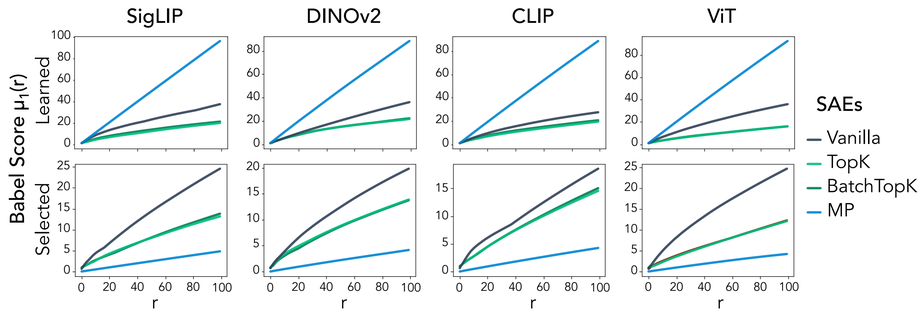




MP-SAE is More Expressive

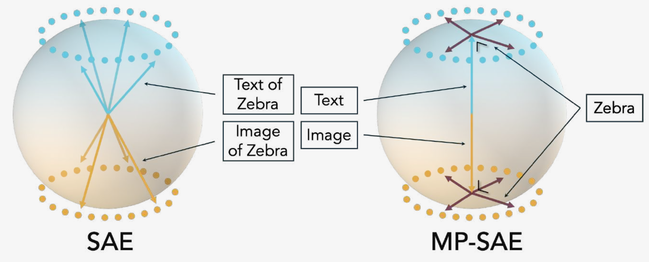


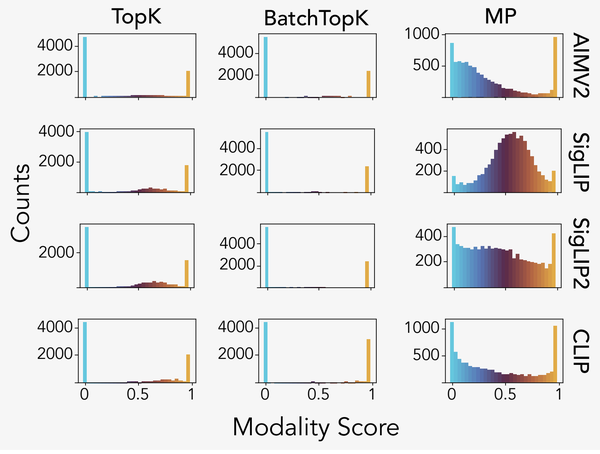
MP-SAE *Learns* Correlated Features, *Selects* Conditionally Orthogonal Ones



Interpreting Multi-Modal Large Models

MP-SAE Identifies Shared Structures Across Modalities





NeuBahar Lab

NeurIPS 2025

MP-SAE on MNIST

CRISP Group at Harvard