AI/CRV 2026

NeurIPS 2025

From Flat to Hierarchical: Extracting Sparse Representations with Matching Pursuit

Valérie Costa[1*] **Thomas Fel**[2*] **Ekdeep Singh Lubana**[3,4*]
Bahareh Tolooshams[5,6†] **Demba Ba**[2,7†]

[1]EPFL [2]Kempner Institute, Harvard University
[3]CBS-NTT Program in Physics of Intelligence, Harvard University
[4]Physics of Artificial Intelligence Group, NTT Research, Inc., Sunnyvale, CA, USA
[5]University of Alberta [6]Alberta Machine Intelligence Institute (Amii) [7]SEAS, Harvard University

From Flat to Hierarchical: Extracting Sparse Representations with Matching Pursuit

Bahareh Tolooshams, Ph.D.
Assistant Professor,
University of Alberta
btolooshams@ualberta.ca
https://btolooshams.github.io

May 28, 2026

NeuBahar Lab

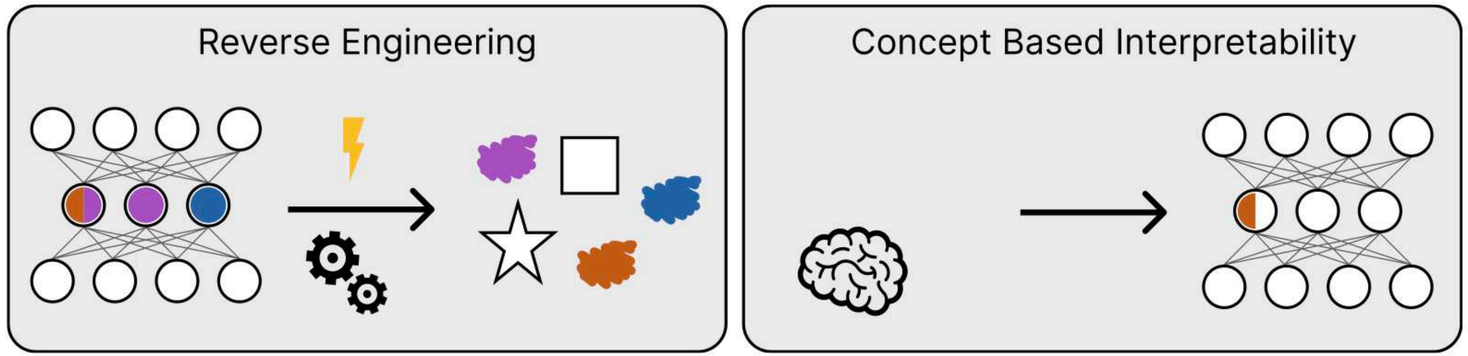
UNIVERSITY OF ALBERTA

amii

CIFAR

How Does Large Neural Networks Represent Data?

Mechanistic Interpretability: Understanding the structures underlying internal model representations.



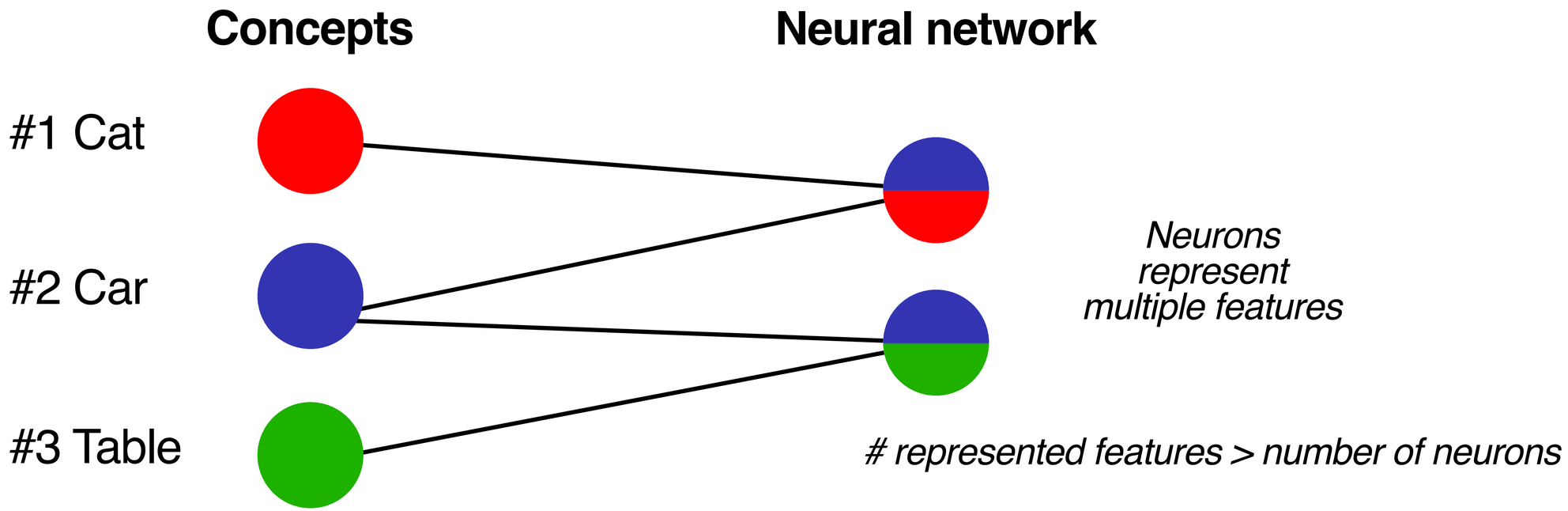
Identify the roles of network components

Identify the network components for given roles

- Decompose the internal AI representation into *interpretable* components.

Sharkey, L., Chughtai, B., Batson, J., Lindsey, J., Wu, J., Bushnaq, L., ... & McGrath, T. (2025). Open problems in mechanistic interpretability. *arXiv.*

How Does AI Represent Data?



- Features are sufficiently **sparse**, so assign features rarely co-occur to one neuron.

- As sparsity increases, models use **superposition** to represent more features than dimensions.

Elhage, N., Hume, T., Olsson, C., Schiefer, N., Henighan, T., Kravec, S., ... & Olah, C. (2022). Toy models of superposition. *arXiv.*

Linear Representation Hypothesis (Model of Data)

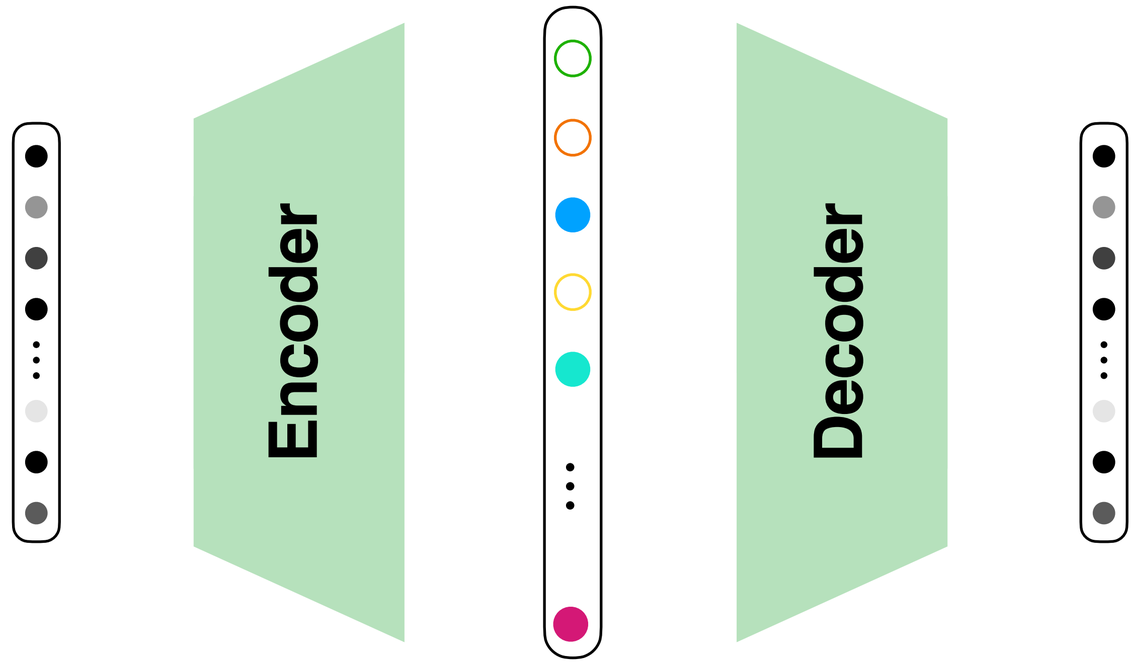
Definition 2.1 (Linear Representation Hypothesis (LRH)). *A representation* $\boldsymbol{x} \in \mathbb{R}^m$ *is said to satisfy the* linear representation hypothesis *(LRH) if there exists a dictionary* $\boldsymbol{D} = [\boldsymbol{D}_1, \ldots, \boldsymbol{D}_p] \in \mathbb{R}^{m \times p}$ *and a coefficient vector* $\boldsymbol{z} \in \mathbb{R}^p$ *such that* $\boldsymbol{x} = \boldsymbol{D}\boldsymbol{z}$*, under the following conditions:*

$$\begin{cases} \textit{(i) Overcompleteness:} & p \gg m; \\ \textit{(ii) Quasi-orthogonality:} & \max_{i \neq j} \left| \boldsymbol{D}_i^\top \boldsymbol{D}_j \right| \leq \varepsilon, \quad \textit{where } \forall i \ \ \|\boldsymbol{D}_i\|_2 = 1 \ ; \textit{ and} \\ \textit{(iii) Sparsity:} & |\operatorname{supp}(\boldsymbol{z})| \leq k \ll p. \end{cases}$$

- Features are represented by **directions**.
- Features are linearly **accessible**.

We have "model of data"; we can now construct a sparse autoencoder (SAEs) to extract geometry of representations in large language models.

Sparse Autoencoders (SAEs) to Extract Concepts from Model Representations



Internal model activations

Sparse representation

Reconstruction

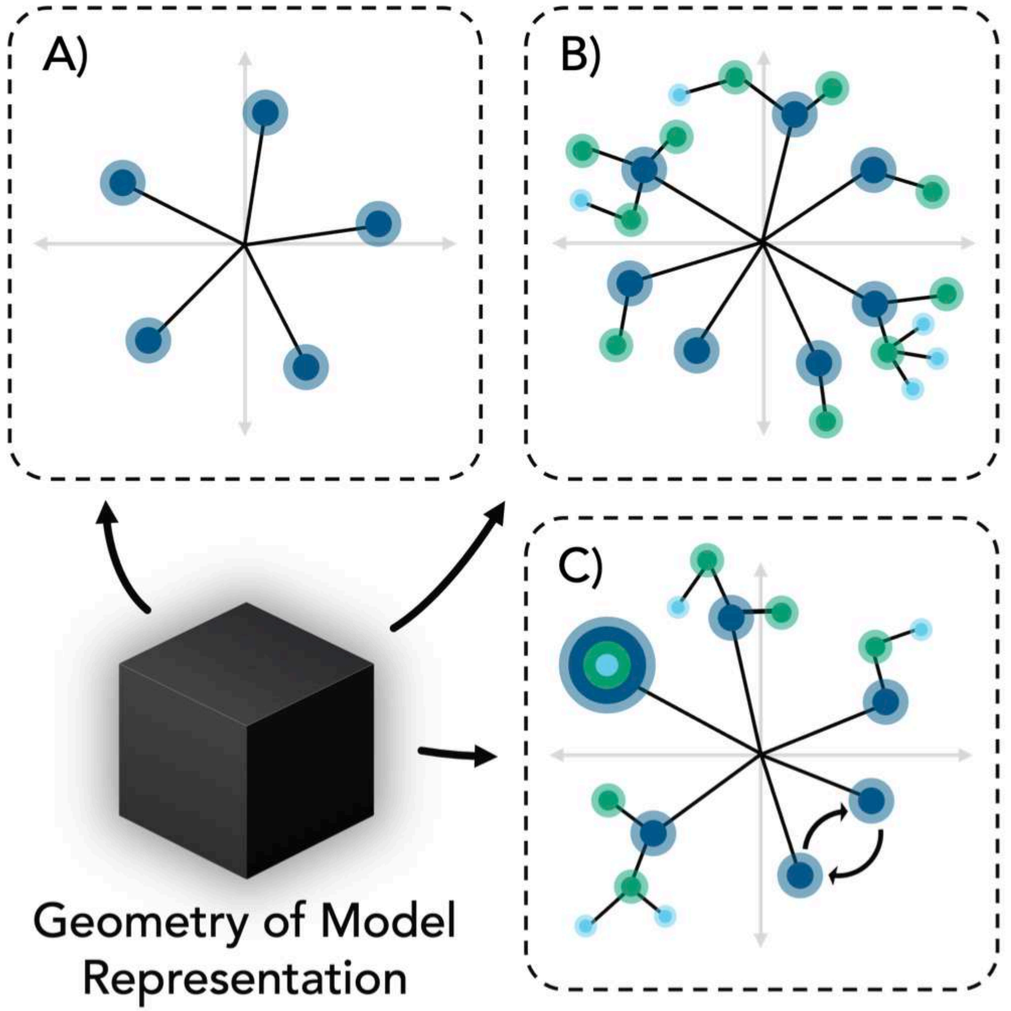
Can SAEs recover features beyond the *Linear Representation Hypothesis*?

Conceptual Organization in Neural Representations

A) ***Linearly accessible:*** approximately orthogonal abstract directions, a.k.a Linear Representation Hypothesis (LRH).

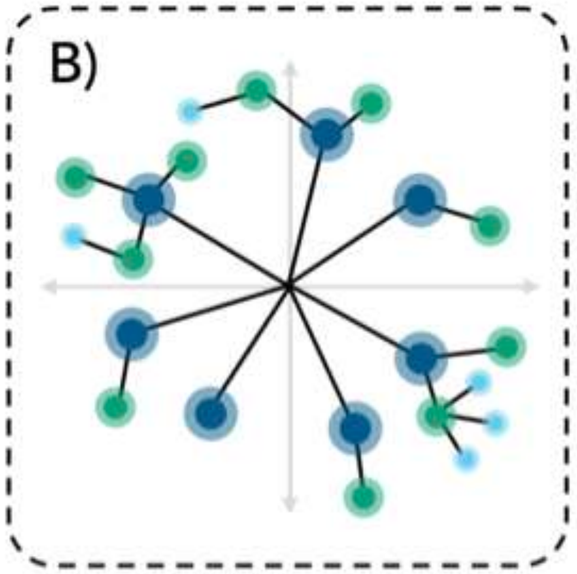
B) ***Hierarchical:*** representations are structured in parent-child relations.

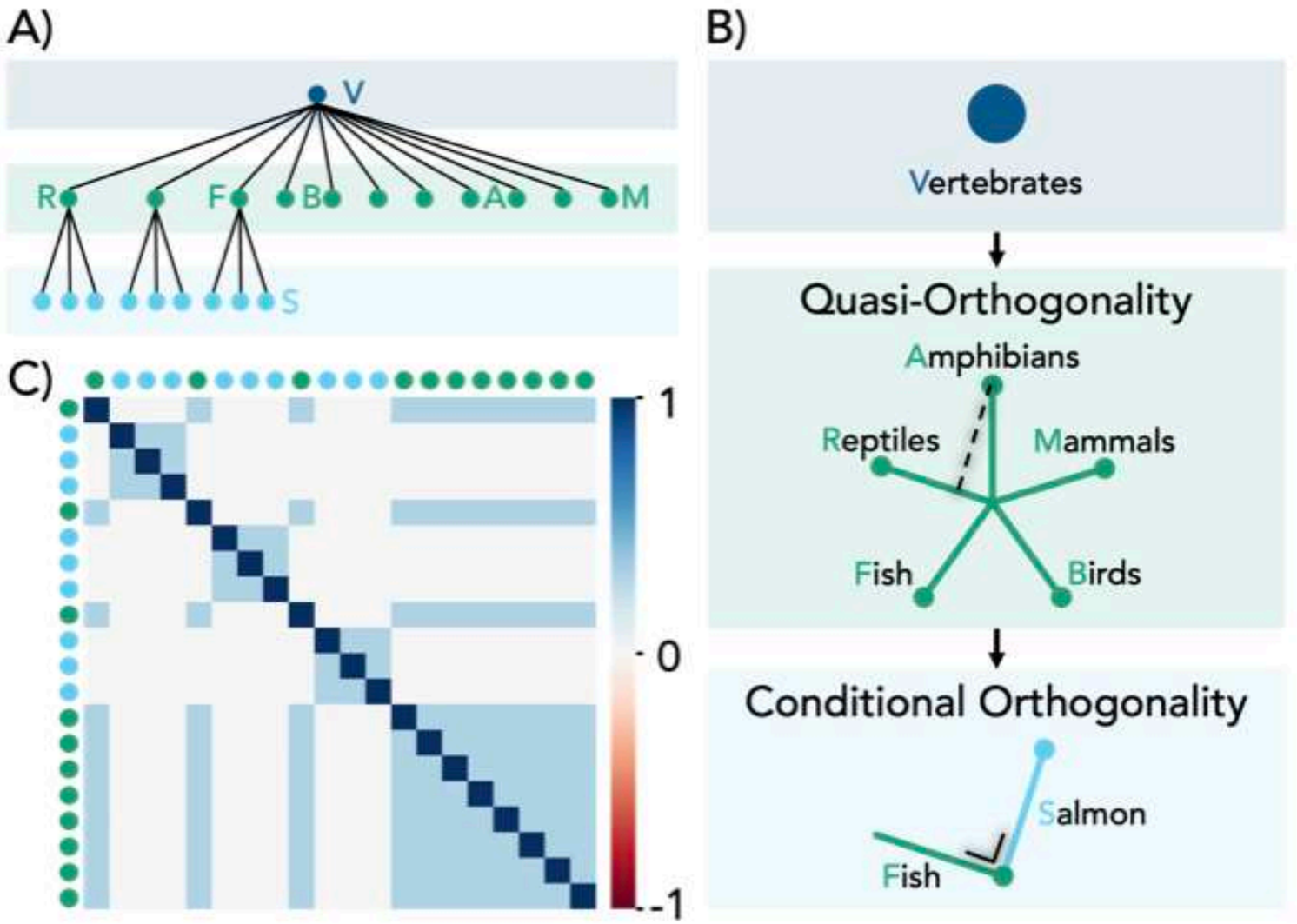
C) *Nonlinear, multidimensional, and temporally structured …*



Costa, V., Fel, T., Lubana, E. S., Tolooshams, B., & Ba, D. (2025). From Flat to Hierarchical: Extracting Sparse Representations with Matching Pursuit. *NeurIPS.*

From Flat to Hierarchical Representations





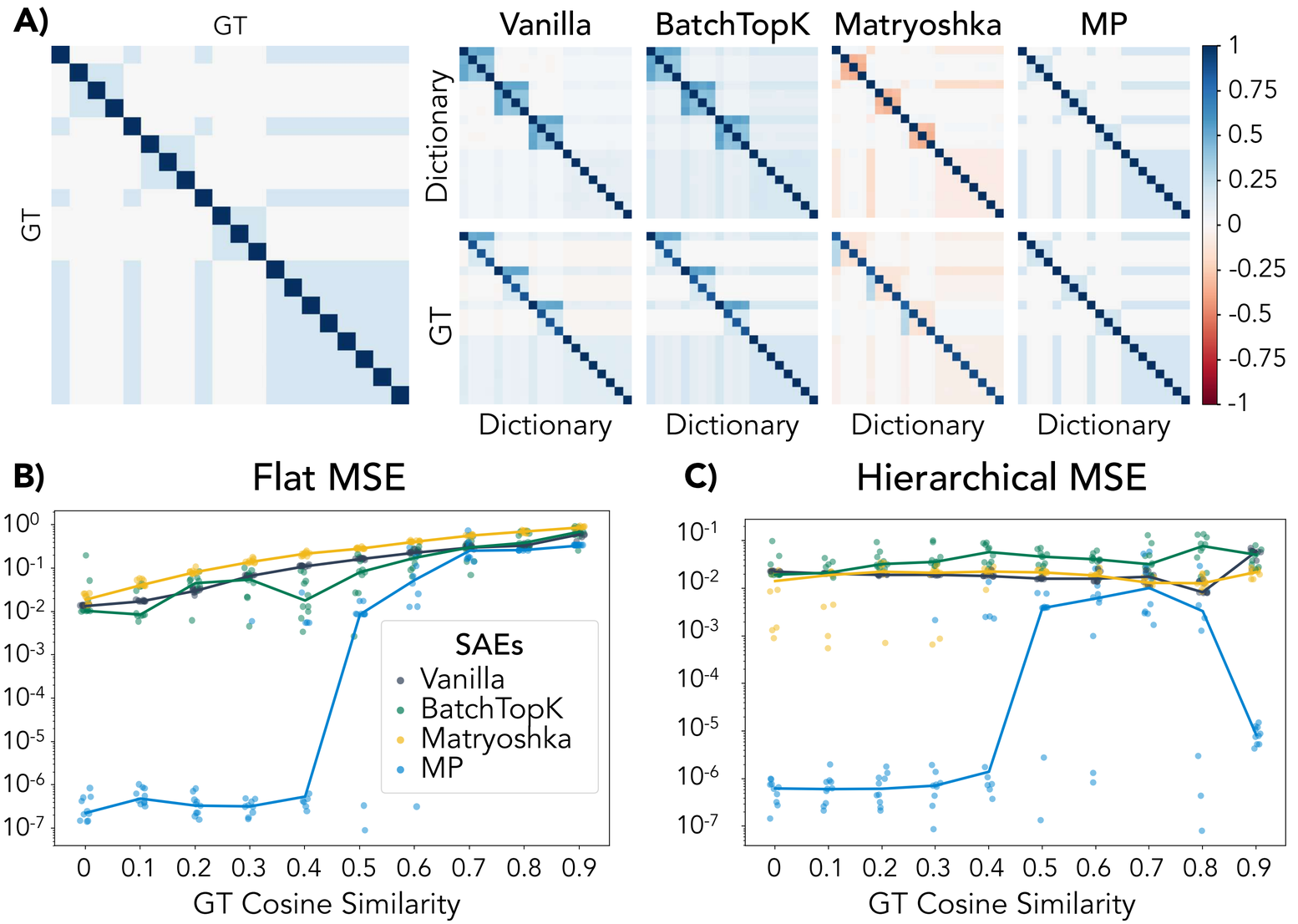
Costa, V., Fel, T., Lubana, E. S., Tolooshams, B., & Ba, D. (2025). From Flat to Hierarchical: Extracting Sparse Representations with Matching Pursuit. *NeurIPS.*

Extracting Hierarchical Representations with Matching Pursuit Sparse Autoencoder (MP-SAE)

$$\boldsymbol{x} = \boldsymbol{r}^{(0)} = \boldsymbol{\varphi}(\boldsymbol{x}) + \boldsymbol{r}^{(1)} = \underbrace{\boldsymbol{\varphi}(\boldsymbol{x})}_{\textit{linearly accessible}} + \underbrace{\sum_{t=1}^{T} \boldsymbol{\varphi}(\boldsymbol{r}^{(t)})}_{\textit{nonlinearly accessible}} + \underbrace{\boldsymbol{r}^{(T+1)}}_{\textit{residual error}},$$

Costa, V., Fel, T., Lubana, E. S., Tolooshams, B., & Ba, D. (2025). From Flat to Hierarchical: Extracting Sparse Representations with Matching Pursuit. *NeurIPS.*

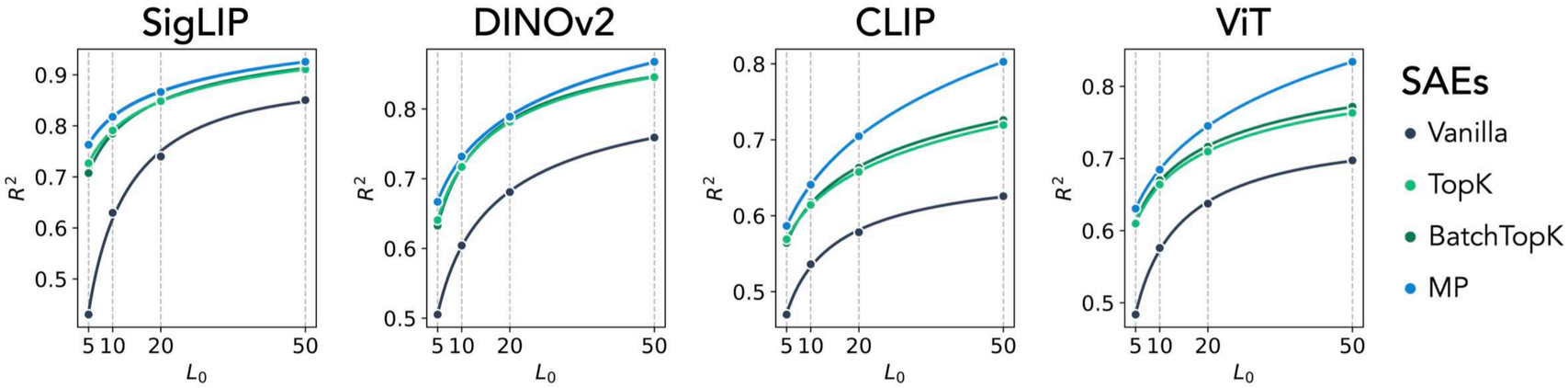
MP-SAE Recovers Hierarchical Correlated Features



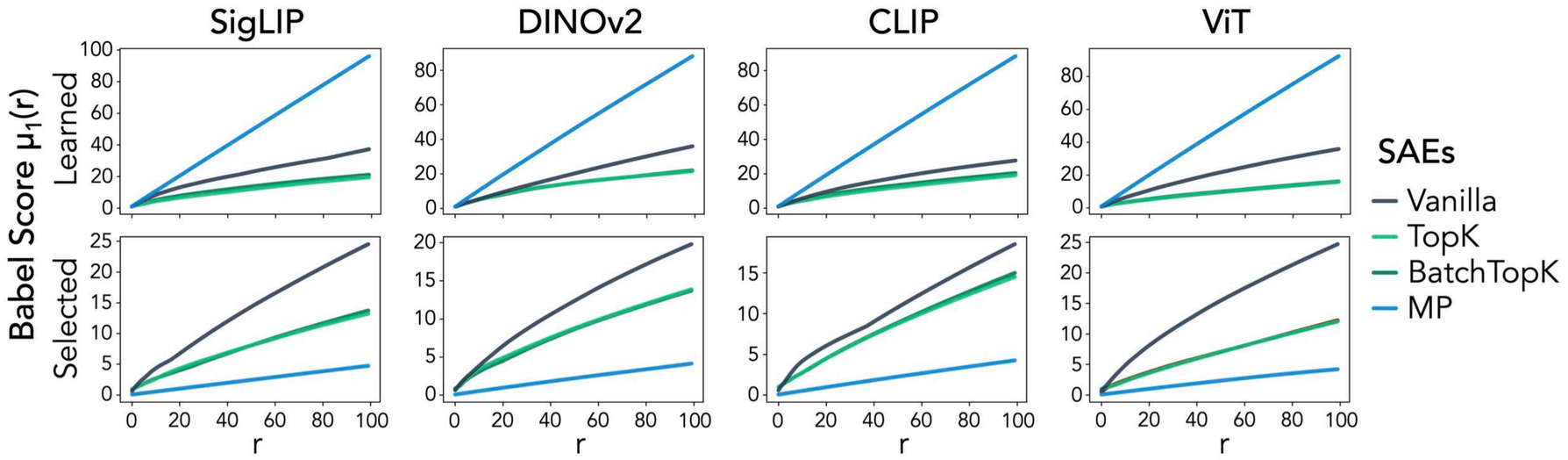
Costa, V., Fel, T., Lubana, E. S., Tolooshams, B., & Ba, D. (2025). From Flat to Hierarchical: Extracting Sparse Representations with Matching Pursuit. *NeurIPS.*

Extracting Features from Internal Representations of Vision Models

MP-SAE recovers more expressive features than standard SAEs.

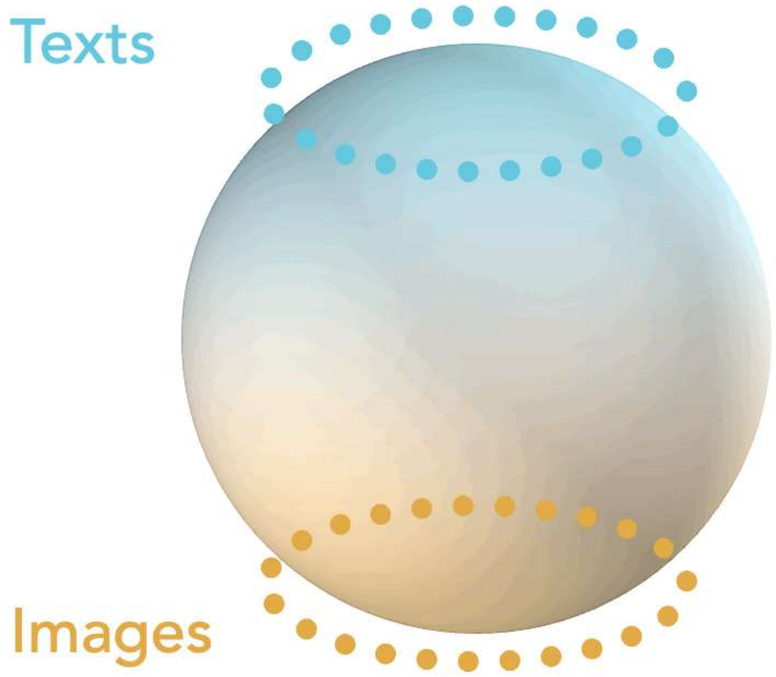


MP-SAE promotes conditional orthogonality at inference.



Costa, V., Fel, T., Lubana, E. S., Tolooshams, B., & Ba, D. (2025). From Flat to Hierarchical: Extracting Sparse Representations with Matching Pursuit. *NeurIPS.*

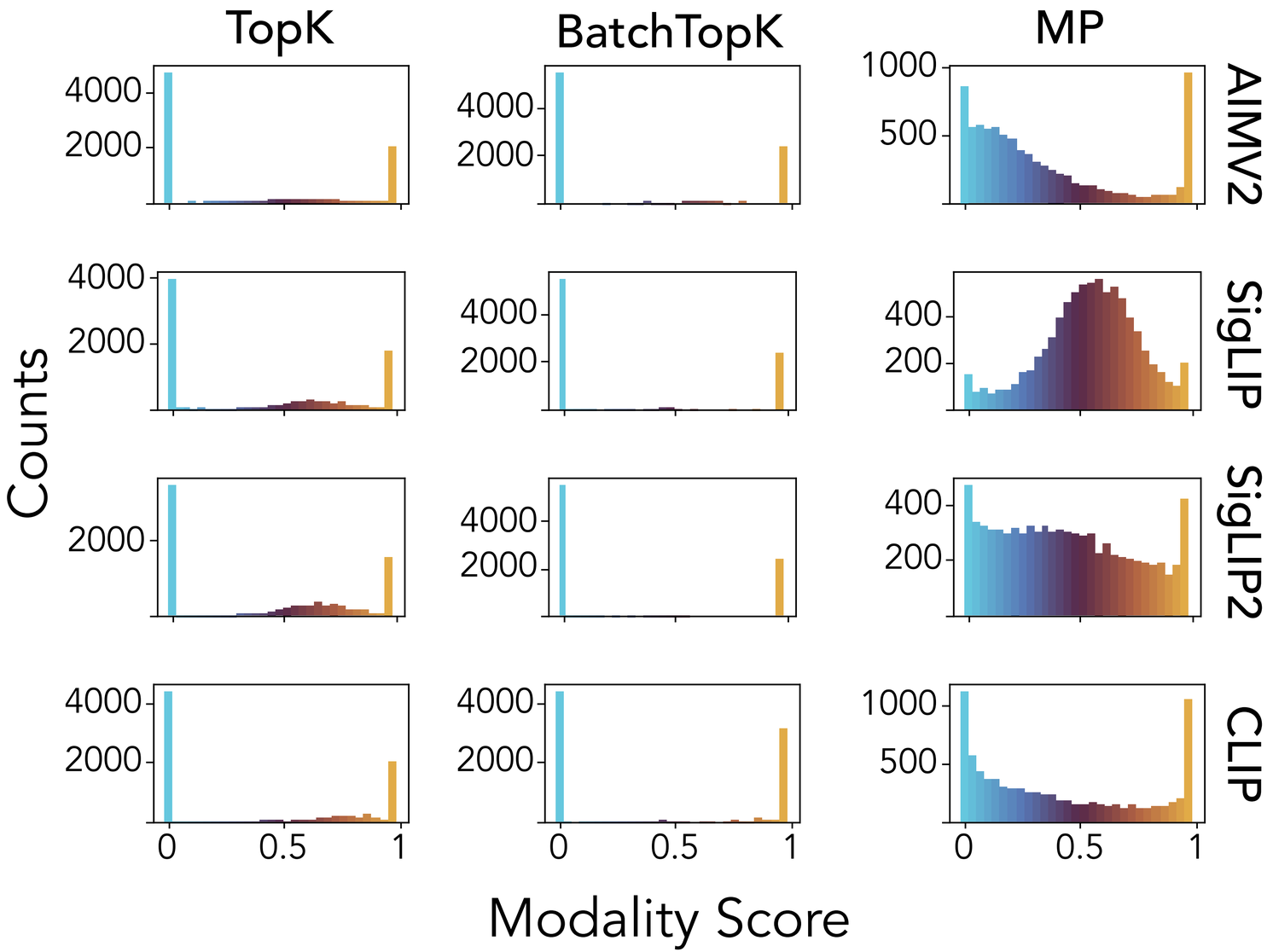
Interpreting Multi-Modal Large Models



Costa, V., Fel, T., Lubana, E. S., Tolooshams, B., & Ba, D. (2025). From Flat to Hierarchical: Extracting Sparse Representations with Matching Pursuit. *NeurIPS.*

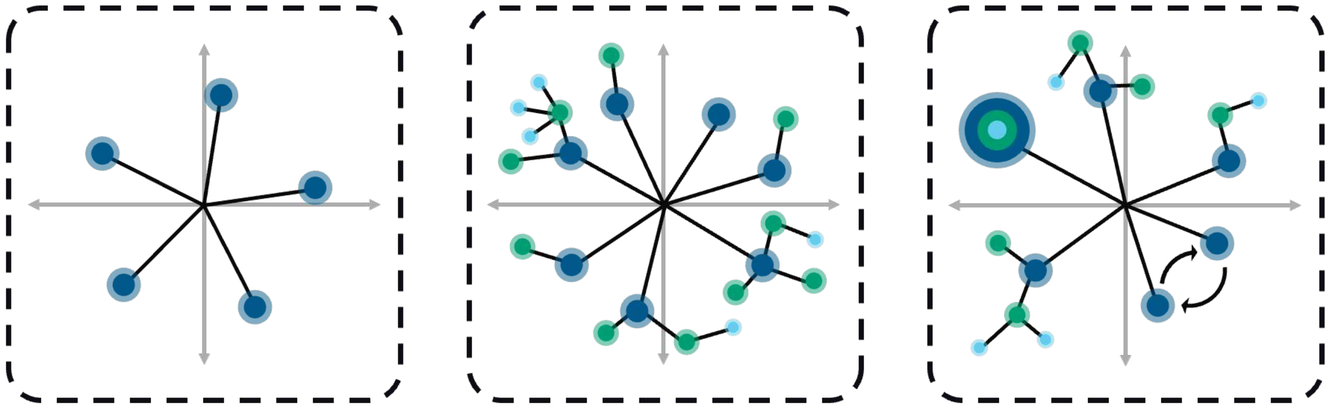
Identifying Shared Structures Across Modalities

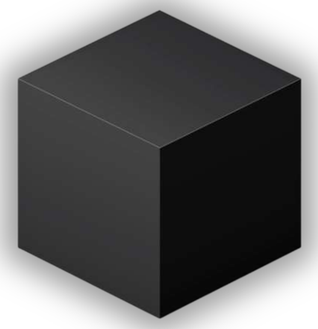
Emergent modality subtraction from abstract meaning



Costa, V., Fel, T., Lubana, E. S., Tolooshams, B., & Ba, D. (2025). From Flat to Hierarchical: Extracting Sparse Representations with Matching Pursuit. *NeurIPS.*

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

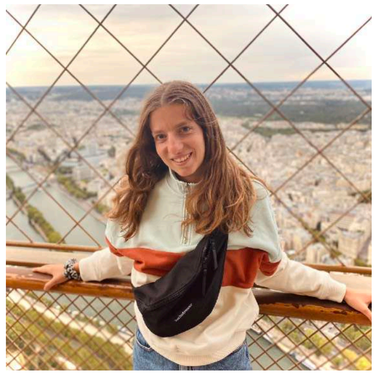




Geometry of Model Representation ?

Costa, V., Fel, T., Lubana, E. S., Tolooshams, B., & Ba, D. (2025). From Flat to Hierarchical: Extracting Sparse Representations with Matching Pursuit. *NeurIPS.*

Valérie Costa



Thomas Fel

Ekdeep Singh Lubana

Demba Ba

HARVARD UNIVERSITY

Kempner INSTITUTE

EPFL

NTTResearch

NeuBahar Lab

UNIVERSITY OF ALBERTA

amii

CIFAR

Bertarelli Foundation