Evaluating Sparse Autoencoders: From Shallow Design to Matching Pursuit

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

Abstract

Sparse autoencoders (SAEs) have recently become central tools for interpretability, leveraging dictionary learning principles to extract sparse, interpretable features from neural representations whose underlying structure is typically unknown. This paper evaluates SAEs in a controlled setting using MNIST, which reveals that current shallow architectures implicitly rely on a quasi-orthogonality assumption that limits the ability to extract correlated features. To move beyond this, we compare them with an iterative SAE that unrolls Matching Pursuit (MP-SAE), enabling the residual-guided extraction of correlated features that arise in hierarchical settings such as handwritten digit generation while guaranteeing monotonic improvement of the reconstruction as more atoms are selected.

1. Introduction

Sparse dictionary learning (Mairal et al., 2009; Rubinstein et al., 2010; Tosic & Frossard, 2011) aims to represent data $\boldsymbol{x}^i$ as sparse linear combinations of learned basis vectors ($\boldsymbol{D}$, a.k.a atoms), for $i = 1, \ldots, n$, i.e.,

$$\boldsymbol{x}^i \approx \boldsymbol{D}\boldsymbol{z}^i \quad \text{subject to} \quad \|\boldsymbol{z}^i\|_0 \leq k, \quad \forall\, i = 1, \ldots, n. \tag{1}$$

where the constraint ensures that the sparse code $\boldsymbol{z}^i$ is at most k-sparse. Sparse representations are ubiquitous in science and engineering, with applications in medical imaging (Lustig et al., 2007; Hämäläinen et al., 2013), image restorations (Mairal et al., 2007; Dong et al., 2013), transcriptomics (Cleary et al., 2021), and genomics (Lucas et al., 2006) with origins from computational neuroscience (Olshausen & Field, 1997; 1996). The formulation in (1) leads to a bi-level optimization problem: the *inner* problem performs sparse approximation to estimate the code $\boldsymbol{z}^i$ given the dictionary $\boldsymbol{D}$, while the *outer* problem updates the dictionary based on the current code estimates to better represent the data (Tolooshams, 2023).

Solving this bi-level optimization can be achieved via alternating minimization (Agarwal et al., 2016), alternating between the inner and outer problems until a convergence criterion. The inner problem has been extensively studied in the compressed sensing literature (Donoho, 2006; Candès et al., 2006). Classical approaches include greedy ℓ_0-based algorithms (Tropp, 2004) such as Matching Pursuit (Mallat & Zhang, 1993) and Orthogonal Matching Pursuit (Pati et al., 1993), as well as convex relaxation ℓ_1-based methods (Chen et al., 2001; Tibshirani, 1996). Since sparse recovery lacks a closed-form solution (Natarajan, 1995), the sparse approximation step typically requires multiple residual-based iterations to converge.

Prior work solves dictionary learning by optimizing the outer problem using closed-form least squares (Agarwal et al., 2016), local gradient updates (Chatterji & Bartlett, 2017), or sequential methods like MOD (Engan et al., 1999) and K-SVD (Aharon et al., 2006). A key bottleneck lies in the repeated solution of the inner sparse coding problem, which typically converges sublinearly (Beck & Teboulle, 2009; Moreau & Bruna, 2017). To address this, the unrolling literature proposes turning the iterative solver into a neural network, enabling fixed-complexity approximations. This idea, sparked by LISTA (Gregor & LeCun, 2010), has been shown to achieve linear convergence (Chen et al., 2018; Ablin et al., 2019) and accelerate both inference and learning (Tolooshams & Ba, 2022; Malézieux et al., 2022). Unrolling further allows the inner and outer optimization to be directly mapped to forward inference and backward learning in deep networks (Tolooshams et al., 2020).

[1]École Polytechnique Fédérale de Lausanne [2]Harvard University [3]Kempner Institute [4]CBS-NTT Program in Physics of Intelligence, Harvard University [5]Department of Electrical and Computer Engineering, University of Alberta [6]Alberta Machine Intelligence Institute (Amii) [7]Neuroscience and Mental Health Institute, University of Alberta. Correspondence to: Demba Ba <demba@seas.harvard.edu>.
[*]Work done while visiting Harvard University. [†]Co-senior authors.

Proceedings of the 42nd International Conference on Machine Learning, Vancouver, Canada. PMLR 267, 2025. Copyright 2025 by the author(s).

Sparsity has been established as a useful prior for interpretability (Mairal et al., 2014; Lipton, 2017; Ribeiro et al., 2016). Building on this principle, recent work has proposed the use of sparse autoencoders (SAEs) to extract human-interpretable features from the internal activations of large language models (LLMs) (Elhage et al., 2022; Cunningham et al., 2023; Bricken et al., 2023; Rajamanoharan et al., 2024; Gao et al., 2025). These methods are motivated by the linear representation hypothesis (Arora et al., 2016; Park et al., 2024; Elhage et al., 2022), which posits that internal representations can be modeled as sparse linear combinations of semantic directions. Unlike the classical dictionary learning methods described above, SAEs solve the inner problem in a single iteration.

Although SAEs are widely used for model interpretability (Kim et al., 2018; Cunningham et al., 2023), they have rarely been evaluated in small-scale, controlled settings, despite their strong ties to classical dictionary learning. Among the few studies, (Hindupur et al., 2025) demonstrate that SAEs impose structural assumptions that shape what they can and cannot detect.

Our Contributions In this work, we revisit the relationship between SAEs and dictionary learning by testing common interpretability SAEs architectures on MNIST, a controlled and well-established dataset in sparse coding literature. Despite similarities in architecture, we find that SAEs with different sparsity mechanisms yield structurally distinct dictionaries. These differences may have important implications for interpretability.

We demonstrate that shallow SAEs implicitly favor near-orthogonal dictionaries, due to their one-shot sparse inference. To investigate this further, we introduce MP-SAE, which unrolls Matching Pursuit into a sequential, residual-guided inference process that operates effectively in regimes with highly correlated features and is supported by convergence guarantees. MP-SAE learns a globally correlated set of atoms but uses a low-redundancy subset to represent each input, a property missing in current shallow SAEs.

Compared to shallow SAEs, MP-SAE yields more expressive representations and naturally constructs a representational hierarchy—first selecting atoms that capture coarse structure, then adding finer details. This coarse-to-fine behavior may lead to more interpretable representations, as it mirrors the hierarchical organization of real-world features (Bussmann et al., 2025).

Algorithm 1 Matching Pursuit Sparse Autoencoders (MP-SAE)

Input: Dictionary $\boldsymbol{D}$, bias $\boldsymbol{b}_{\text{pre}}$, data $\boldsymbol{x}$, steps T
Initialize residual: $\boldsymbol{r}^{(0)} = \boldsymbol{x} - \boldsymbol{b}_{\text{pre}}$
Initialize reconstruction: $\hat{\boldsymbol{x}}^{(0)} = \boldsymbol{b}_{\text{pre}}$
Initialize sparse code: $\boldsymbol{z}^{(0)} = \boldsymbol{0}$
for $t = 1, \ldots, T$ **do**
$j^{(t)} = \arg\max_{j=\{1,\ldots,p\}} (\boldsymbol{D}^\top \boldsymbol{r}^{(t-1)})_j$
$\boldsymbol{z}^{(t)}_{j^{(t)}} = \boldsymbol{D}^\top_{j^{(t)}} \boldsymbol{r}^{(t-1)}$
$\hat{\boldsymbol{x}}^{(t)} = \hat{\boldsymbol{x}}^{(t-1)} + \boldsymbol{z}^{(t)}_{j^{(t)}} \boldsymbol{D}_{j^{(t)}}$
$\boldsymbol{r}^{(t)} = \boldsymbol{r}^{(t-1)} - \boldsymbol{z}^{(t)}_{j^{(t)}} \boldsymbol{D}_{j^{(t)}}$
end for
Output: Sparse code $\boldsymbol{z} = \sum_{t=1}^{T} \boldsymbol{z}^{(t)}$
Reconstruction $\hat{\boldsymbol{x}} = \hat{\boldsymbol{x}}^{(T)}$

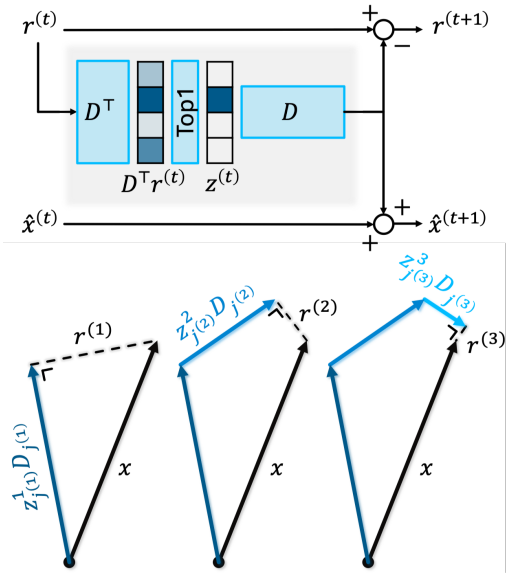

Figure 1: **MP-SAE.** Top : Full forward pass. Bottom: One encoder iteration.

2. Background

Representation Hypotheses Efforts to understand neural representations through interpretability are often motivated by two guiding hypotheses: the *Linear Representation Hypothesis* and the *Superposition Hypothesis* (Arora et al., 2016; Olah et al., 2020; Park et al., 2024; Elhage et al., 2022). These suggest that internal activations of large deep neural networks can be expressed as linear combinations of human-interpretable concept directions $\boldsymbol{D}$. In practice, the number of possible concepts p far exceeds the dimensionality m of the representation: $m \ll p$, leading to superposition—multiple concepts overlapping within the same activation (Elhage et al., 2022). Despite this, meaningful disentanglement is possible under the sparsity assumption: $\|\boldsymbol{z}\|_0 \leq k$, where only a small number $k \ll p$ of concepts are active (Donoho, 2006).

Sparse Concept Extraction as Dictionary Learning As formalized in (Fel et al., 2023), the task of concept extraction in interpretability can be cast as a dictionary learning problem: learn a set of interpretable directions $\boldsymbol{D}$ such that activations $\boldsymbol{x}$ can be approximated by sparse linear combinations with sparse code $\boldsymbol{z}$ (See equation 1). In practice, this is most often implemented using shallow SAEs, which have been shown to extract meaningful and monosemantic concepts across a variety of architectures and domains (Elhage et al., 2022; Cunningham et al., 2023; Bricken et al., 2023).

Sparse Autoencoders Given an input $\boldsymbol{x} \in \mathbb{R}^m$, an SAE computes a sparse code using an encoder $\boldsymbol{z} = \sigma(\boldsymbol{W}^\top(\boldsymbol{x}-\boldsymbol{b}_{\text{pre}})+\boldsymbol{b})$ and reconstructs the data as $\hat{\boldsymbol{x}} = \boldsymbol{D}\boldsymbol{z} + \boldsymbol{b}_{\text{pre}}$, where $\boldsymbol{W} \in \mathbb{R}^{m\times p}$ and $\boldsymbol{b} \in \mathbb{R}^p$ are the encoder parameters, $\boldsymbol{b}_{\text{pre}} \in \mathbb{R}^m$ is a bias, and $\boldsymbol{D} \in \mathbb{R}^{m\times p}$ is the learned dictionary of interest with normalized atoms (i.e., $\|\boldsymbol{D}_i\|_2 = 1$). The nonlinearity $\sigma(\cdot)$ enforces sparsity; common choices include ReLU, TopK, and JumpReLU. These models are trained to minimize a sparsity-augmented reconstruction loss: $\mathcal{L} = \|\boldsymbol{x} - \hat{\boldsymbol{x}}\|_2^2 + \lambda\,\mathcal{R}(\boldsymbol{z}) + \alpha\,\mathcal{L}_{\text{aux}}$, where $\mathcal{R}(\boldsymbol{z})$ promotes sparsity, e.g., via ℓ_1 regularization for ReLU (Cunningham et al., 2023; Bricken et al., 2023) or target-ℓ_0 penalties (Rajamanoharan et al., 2024).

Orthogonality and Limitations of Shallow Recovery Sparse recovery (i.e., the inner problem optimization) in one iteration is theoretically guaranteed only when the dictionary $\boldsymbol{D}$ is sufficiently incoherent, i.e., when the mutual coherence $\mu(\boldsymbol{D}) = \max_{i\neq j} |\boldsymbol{D}_i^\top \boldsymbol{D}_j|$ is small (Makhzani & Frey, 2014; Arora et al., 2015). However, concepts underlying natural data may exhibit high coherence. This limitation motivates the use of unrolled sparse coding networks, which address the inner problem through multiple iterative updates, progressively refining the estimate of the sparse code z as in classical sparse recovery methods. Shallow SAEs can be seen as approximating unrolled networks by performing only a single iteration: they rely on the same nonlinearities – JumpReLU implementing hard thresholding as in NOODL (Rambhatla et al., 2018), and ReLU implementing soft thresholding as in PUDLE (Tolooshams & Ba, 2022) – but do not perform iterative updates.

3. Unrolling Matching Pursuit into MP-SAE

We propose the *Matching Pursuit Sparse Autoencoder* (MP-SAE), which unrolls Matching Pursuit (MP) (Mallat & Zhang, 1993) into an iterative sparse autoencoder architecture. Previously used for image restoration tasks under the name Learned-MP (Khatib et al., 2021), our work focuses on the dictionary properties learned by MP-SAE and compares them to those obtained with shallow SAEs.

Iterative Inference via Residual Updates MP-SAE can be seen as an unrolled variant of TopK. Instead of selecting k atoms in a single iteration, as in TopK, it selects one atom at each of the k iterations. This allows the model to subtract the contribution of each selected atom from the input before the next selection, so that subsequent atoms are chosen to explain what remains unexplained. This unexplained part is captured by the residual, defined as the difference between the input $\boldsymbol{x}$ and its current reconstruction $\hat{\boldsymbol{x}}$, i.e., $\boldsymbol{r} = \boldsymbol{x} - \hat{\boldsymbol{x}}$. Initially, the residual is equal to the input (or $\boldsymbol{x} - \boldsymbol{b}_{\text{pre}}$).

At each iteration, MP greedily selects the dictionary atom that best aligns with the current residual. This is done by computing the inner product between the residual and each atom, and selecting the one with the highest projection. Once the best-matching atom is selected, the algorithm projects the residual onto that atom to determine its contribution to the reconstruction. This contribution is then added to the current approximation of the input and subtracted from the residual. Over time, this greedy procedure iteratively reduces the residual and improves the reconstruction, step by step (see Figure 1 and Algorithm 1).

Theoretical Properties of Matching Pursuit At each step, the residual $\boldsymbol{r}^{(t)}$ is orthogonal to the selected atom $\boldsymbol{D}_{j^{(t)}}$ (Proposition 8.1). Moreover, the norm of the residual decreases monotonically at each iteration (Proposition 8.2) and converges asymptotically to the component of the input orthogonal to the span of the dictionary $\boldsymbol{D}$ (Proposition 8.3).

Because each selected atom is subtracted from the residual at each iteration, the algorithm naturally explores new directions – each selection is driven by what remains unexplained, pushing the model toward atoms that are less redundant and more complementary. This residual-driven mechanism improves diversity in the selected features and enhances robustness in the presence of dictionary coherence. Indeed, MP can achieve exponential convergence and accurately recover active components even when the dictionary is highly coherent, as long as it satisfies a block-incoherence condition, where correlations are localized within small groups of atoms (Peotta & Vandergheynst, 2007).

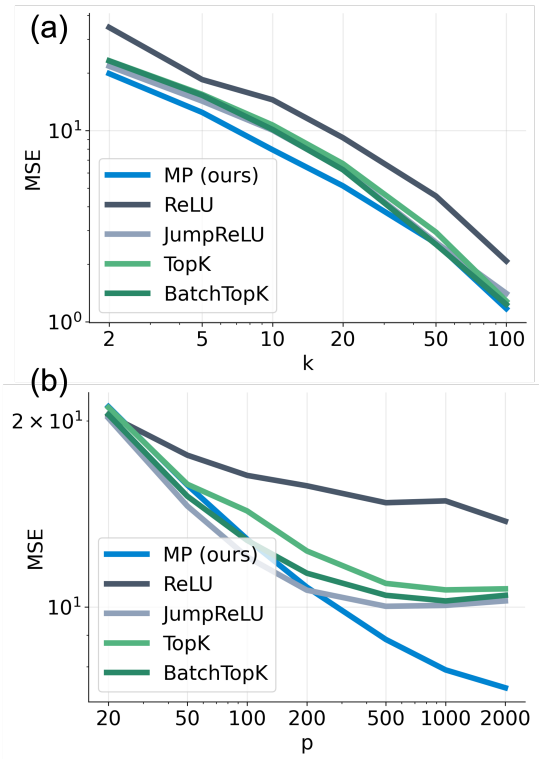

Figure 2: **Expressivity.**

4. Results

We evaluate MP-SAE against four shallow SAEs: ReLU (Elhage et al., 2022; Cunningham et al., 2023), JumpReLU (Rajamanoharan et al., 2024), TopK (Gao et al., 2025), and BatchTopK (Bussmann et al., 2024), using the MNIST dataset. Additional results on large vision models, including expressivity and coherence, are provided in Appendix 7.

Expressivity We assess reconstruction performance by varying two key parameters: the sparsity level k and the dictionary size p. Figure 2(a) shows that, when fixing $p = 1000$ and varying k, MP-SAE is consistently more expressive across sparsity levels—despite having half the capacity of other SAEs due to the absence of encoder parameters from weight tying. As shown in Figure 2(b), when $k = 10$ is fixed and p is swept, MP-SAE continues to improve as the dictionary size grows, while shallow SAEs plateau. This highlights the efficiency of MP-SAE in leveraging additional capacity under the same sparsity constraint.

Learned Concepts In the following, we focus on dictionaries trained with $k = 10$ and $p = 1000$. The top row in Figure 3 shows 25 atoms with the highest activation frequency for each SAE. All methods except ReLU appear to learn pen-stroke-like patterns, while ReLU learns atoms resembling full digits. The pen strokes learned by MP-SAE appear more precise. Interestingly, all shallow methods include a "negative" atom that closely resembles $-\boldsymbol{b}_{\text{pre}}$, which is the most frequently activated atom in ReLU and JumpReLU.

When sorting atoms by their average activation value $\mathbb{E}(\boldsymbol{z}_j)$ rather than frequency $\sum_i \boldsymbol{z}_j^i \neq 0$, a structural shift unique to MP-SAE emerges. As shown in the bottom row of Figure 3, the most heavily weighted atoms in MP-SAE resemble clean, idealized digit prototypes, in contrast to the more detailed pen strokes observed among its most frequently activated atoms. In comparison, shallow SAEs exhibit little variation between these two rankings; ReLU continues to activate full-digit atoms, while the others still primarily activate stroke-like patterns.

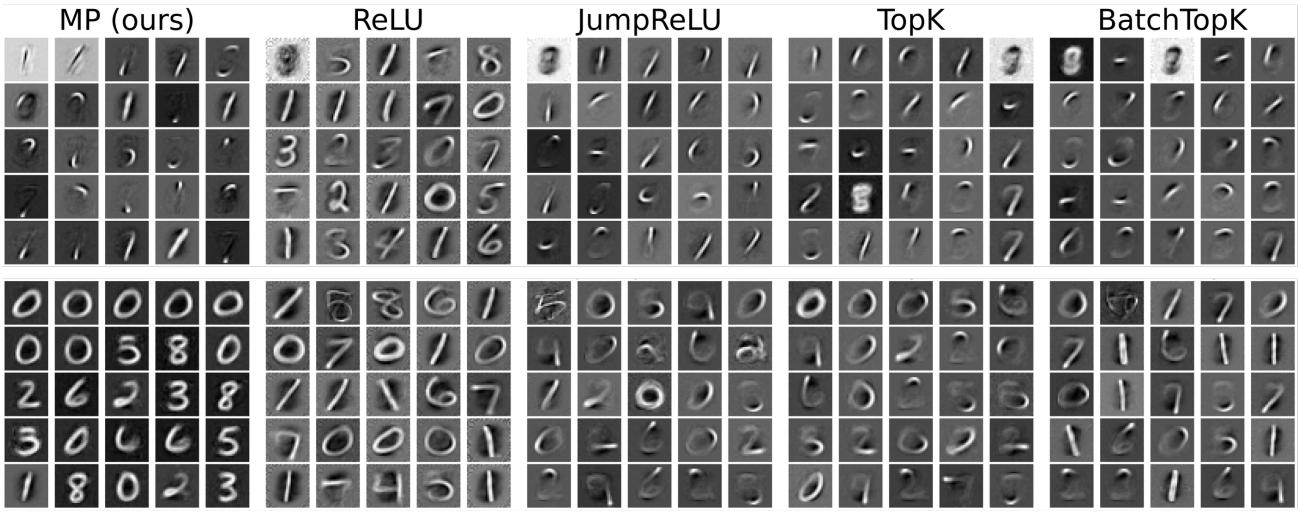

Figure 3: **Feature Selection vs. Activation Levels**. Top: atoms with highest activation frequency (ℓ_0). Bottom: atoms with highest activation $\mathbb{E}[\boldsymbol{z}_j]$ (ℓ_1).

Hints of Hierarchy To capture this shift, the distributions of activation frequency and average activation value $\mathbb{E}[\boldsymbol{z}_j]$ are shown in Figure 4(a). JumpReLU, TopK, and BatchTopK exhibit high variance in activation frequency—some atoms are rarely used, while others are activated very frequently. In contrast, the variance in activation values remains low, with slightly higher values observed for the more frequently used atoms. By comparison, ReLU activates its dictionary atoms uniformly, both in frequency and magnitude. MP displays a perpendicular trend to the other ℓ_0-based methods and JumpReLU: its atoms are activated with roughly equal frequency—similar to ReLU—but their activation values vary widely. Some atoms contribute much more to the reconstruction than others. A subtle inverse relationship is also observed: atoms with higher activation values tend to be used less frequently.

This hierarchical behavior is supported by Figure 4(b), which shows that atoms with higher average activation values tend to be selected in the early layers of MP-SAE (the first iterations of Matching Pursuit). These atoms correspond to digit-like patterns in the bottom row of Figure 3, and are refined by later atoms capturing more localized pen strokes (top row). This progression suggests a hierarchical structure in MP-SAE, reconstructing from coarse to fine features (see Appendix 6).

To enable comparison with shallow encoders, atoms are reordered by activation value to simulate a sequential selection. ℓ_0-based methods show a concentration of atoms toward the end, likely due to the auxiliary loss. This also highlights that ReLU exhibits an amplitude bias.

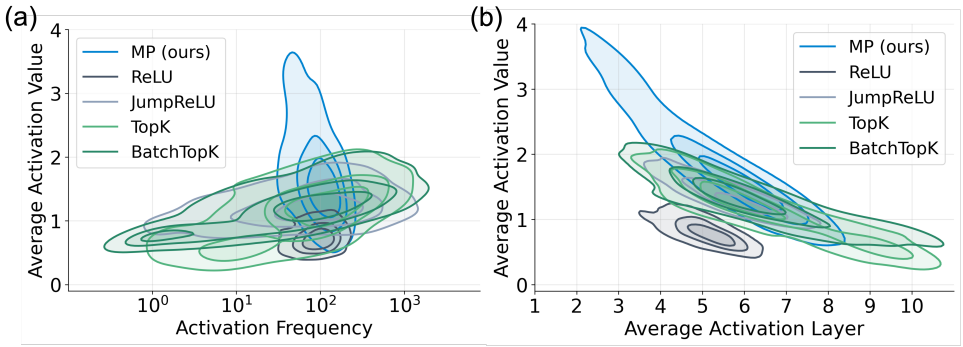

Figure 4: **Activation distributions**.

Coherence Finally, we assess the coherence of both the learned dictionary and the atoms selected at inference time using the Babel score (Tropp, 2004). Unlike mutual coherence, which captures only the maximum pairwise similarity between atoms, the Babel function provides a more comprehensive measure of redundancy by quantifying cumulative coherence. Given a dictionary $\boldsymbol{D} = [\boldsymbol{D}_1, \ldots, \boldsymbol{D}_p] \in \mathbb{R}^{m \times p}$ with unit-norm columns, the Babel function of order r is defined as:

$$\mu_1(r) = \max_{S \subset [p], |S|=r} \left(\max_{j \notin S} \sum_{i \in S} \left| \boldsymbol{D}_i^\top \boldsymbol{D}_j \right| \right)$$

The value $\mu_1(r)$ reflects how well a single atom can be approximated by a group of r others—lower values indicate lower redundancy. Further details are provided in Appendix 7.

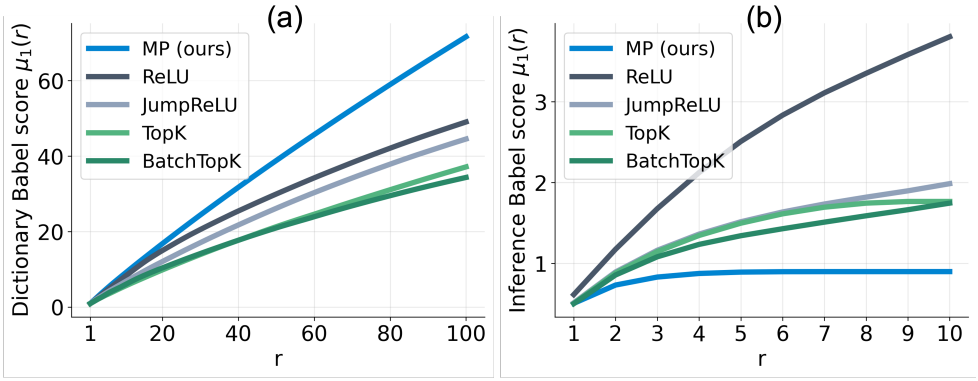

Figure 5: **Coherence analysis of learned concepts.**

Figure 5(a) shows that MP-SAE exhibits a more coherent dictionary than the shallow SAEs. However, as shown in Figure 5(b), it selects more incoherent atoms at inference. This highlights MP's ability to draw incoherent subsets from a globally coherent dictionary. Interestingly, for shallow SAEs, the trends for the learned dictionary and the selected atoms align: ReLU consistently exhibits the highest coherence, while TopK remains the least coherent. This suggests that shallow SAEs are constrained to select more correlated atoms when the dictionary itself is more coherent.

5. Conclusion

We introduce MP-SAE and show through small-scale experiments on MNIST that it improves expressivity, learns hierarchical features, and overcomes coherence limitations of shallow SAEs. These experiments also reveal distinct representation behaviors across SAEs and offer a foundation for building more interpretable sparse autoencoders.

References

Ablin, P., Moreau, T., Massias, M., and Gramfort, A. Learning step sizes for unfolded sparse coding. In *Proceedings of Advances in Neural Information Processing Systems*, volume 32, pp. 1–11, 2019.

Agarwal, A., Anandkumar, A., Jain, P., and Netrapalli, P. Learning sparsely used overcomplete dictionaries via alternating minimization. *SIAM Journal on Optimization*, 26(4):2775–2799, 2016.

Aharon, M., Elad, M., and Bruckstein, A. K-svd: An algorithm for designing overcomplete dictionaries for sparse representation. *IEEE Transactions on Signal Processing*, 54(11):4311–4322, 2006.

Arora, S., Ge, R., Ma, T., and Moitra, A. Simple, efficient, and neural algorithms for sparse coding. In Grünwald, P., Hazan, E., and Kale, S. (eds.), *Proceedings of Conference on Learning Theory*, volume 40 of *Proceedings of Machine Learning Research*, pp. 113–149, Paris, France, 03–06 Jul 2015. PMLR.

Arora, S., Li, Y., Liang, Y., Ma, T., and Risteski, A. A Latent Variable Model Approach to PMI-based Word Embeddings. *Transactions of the Association for Computational Linguistics*, 4:385–399, 2016. doi: 10.1162/tacl_a_00106. URL https://aclanthology.org/Q16-1028/. Place: Cambridge, MA Publisher: MIT Press.

Beck, A. and Teboulle, M. A fast iterative shrinkage-thresholding algorithm for linear inverse problems. *SIAM journal on imaging sciences*, 2(1):183–202, 2009.

Bricken, T., Templeton, A., Batson, J., Chen, B., Jermyn, A., Conerly, T., Turner, N., Anil, C., Denison, C., Askell, A., Lasenby, R., Wu, Y., Kravec, S., Schiefer, N., Maxwell, T., Joseph, N., Hatfield-Dodds, Z., Tamkin, A., Nguyen, K., McLean, B., Burke, J. E., Hume, T., Carter, S., Henighan, T., and Olah, C. Towards monosemanticity: Decomposing language models with dictionary learning. *Transformer Circuits Thread*, 2023. https://transformer-circuits.pub/2023/monosemantic-features/index.html.

Bussmann, B., Leask, P., and Nanda, N. Batchtopk sparse autoencoders. *preprint arXiv:2412.06410*, 2024.

Bussmann, B., Nabeshima, N., Karvonen, A., and Nanda, N. Learning multi-level features with matryoshka sparse autoencoders. *arXiv preprint arXiv:2503.17547*, 2025.

Candès, E. J., Romberg, J., and Tao, T. Robust uncertainty principles: Exact signal reconstruction from highly incomplete frequency information. *IEEE Transactions on information theory*, 52(2):489–509, 2006.

Chatterji, N. S. and Bartlett, P. L. Alternating minimization for dictionary learning: Local convergence guarantees. *arXiv:1711.03634*, pp. 1–26, 2017.

Chen, S. S., Donoho, D. L., and Saunders, M. A. Atomic decomposition by basis pursuit. *SIAM review*, 43(1):129–159, 2001.

Chen, X., Liu, J., Wang, Z., and Yin, W. Theoretical linear convergence of unfolded ista and its practical weights and thresholds. In *Proceedings of Advances in Neural Information Processing Systems*, volume 31, pp. 1–11, 2018.

Cleary, B., Simonton, B., Bezney, J., Murray, E., Alam, S., Sinha, A., Habibi, E., Marshall, J., Lander, E. S., Chen, F., et al. Compressed sensing for highly efficient imaging transcriptomics. *Nature Biotechnology*, pp. 1–7, 2021.

Cunningham, H., Ewart, A., Riggs, L., Huben, R., and Sharkey, L. Sparse Autoencoders Find Highly Interpretable Features in Language Models, October 2023. URL http://arxiv.org/abs/2309.08600. arXiv:2309.08600 [cs].

Deng, J., Dong, W., Socher, R., Li, L.-J., Li, K., and Fei-Fei, L. ImageNet: A Large-Scale Hierarchical Image Database. *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2009.

Dong, W., Zhang, L., Shi, G., and Li, X. Nonlocally Centralized Sparse Representation for Image Restoration. *IEEE Transactions on Image Processing*, 22(4):1620–1630, April 2013. ISSN 1941-0042. doi: 10.1109/TIP.2012.2235847. URL https://ieeexplore.ieee.org/abstract/document/6392274.

Donoho, D. L. Compressed sensing. *IEEE Transactions on information theory*, 52(4):1289–1306, 2006.

Dosovitskiy, A., Beyer, L., Kolesnikov, A., Weissenborn, D., Zhai, X., Unterthiner, T., Dehghani, M., Minderer, M., Heigold, G., Gelly, S., et al. An image is worth 16x16 words: Transformers for image recognition at scale. *arXiv preprint arXiv:2010.11929*, 2020.

Elhage, N., Hume, T., Olsson, C., Schiefer, N., Henighan, T., Kravec, S., Hatfield-Dodds, Z., Lasenby, R., Drain, D., Chen, C., Grosse, R., McCandlish, S., Kaplan, J., Amodei, D., Wattenberg, M., and Olah, C. Toy models of superposition, 2022. URL https://arxiv.org/abs/2209.10652.

Engan, K., Aase, S., and Hakon Husoy, J. Method of optimal directions for frame design. In *Proceedings of IEEE International Conference on Acoustics, Speech, and Signal Processing*, volume 5, pp. 2443–2446 vol.5, 1999.

Fel, T., Boutin, V., Moayeri, M., Cadène, R., Bethune, L., andéol, L., Chalvidal, M., and Serre, T. A Holistic Approach to Unifying Automatic Concept Extraction and Concept Importance Estimation, October 2023. URL http://arxiv.org/abs/2306.07304. arXiv:2306.07304 [cs].

Gao, L., la Tour, T. D., Tillman, H., Goh, G., Troll, R., Radford, A., Sutskever, I., Leike, J., and Wu, J. Scaling and evaluating sparse autoencoders. In *The Thirteenth International Conference on Learning Representations*, 2025. URL https://openreview.net/forum?id=tcsZt9ZNKD.

Gregor, K. and LeCun, Y. Learning fast approximations of sparse coding. In *Proceedings of the 27th International Conference on International Conference on Machine Learning*, ICML'10, pp. 399–406, Madison, WI, USA, 2010. Omnipress. ISBN 9781605589077.

Hindupur, S. S. R., Lubana, E. S., Fel, T., and Ba, D. Projecting assumptions: The duality between sparse autoencoders and concept geometry. *arXiv preprint arXiv:2503.01822*, 2025.

Hämäläinen, K., Kallonen, A., Kolehmainen, V., Lassas, M., Niinimäki, K., and Siltanen, S. Sparse Tomography. *SIAM Journal on Scientific Computing*, 35(3):B644–B665, January 2013. ISSN 1064-8275. doi: 10.1137/120876277. URL https://epubs.siam.org/doi/abs/10.1137/120876277. Publisher: Society for Industrial and Applied Mathematics.

Khatib, R., Simon, D., and Elad, M. Learned greedy method (lgm): A novel neural architecture for sparse coding and beyond. *Journal of Visual Communication and Image Representation*, 77:103095, 2021.

Kim, B., Wattenberg, M., Gilmer, J., Cai, C., Wexler, J., Viegas, F., et al. Interpretability beyond feature attribution: Quantitative testing with concept activation vectors (tcav). In *International conference on machine learning*, pp. 2668–2677. PMLR, 2018.

Lipton, Z. C. The Mythos of Model Interpretability, March 2017. URL http://arxiv.org/abs/1606.03490. arXiv:1606.03490 [cs].

Lucas, J., Carvalho, C., Wang, Q., Bild, A., Nevins, J. R., and West, M. Sparse statistical modelling in gene expression genomics. *Bayesian inference for gene expression and proteomics*, 1(1):1644, 2006.

Lustig, M., Donoho, D., and Pauly, J. M. Sparse MRI: The application of compressed sensing for rapid MR imaging. *Magnetic Resonance in Medicine*, 58(6):1182–1195, 2007. ISSN 1522-2594. doi: 10.1002/mrm.21391. URL https://onlinelibrary.wiley.com/doi/abs/10.1002/mrm.21391. _eprint: https://onlinelibrary.wiley.com/doi/pdf/10.1002/mrm.21391.

Mairal, J., Elad, M., and Sapiro, G. Sparse representation for color image restoration. *IEEE Transactions on image processing*, 17(1):53–69, 2007.

Mairal, J., Bach, F., Ponce, J., and Sapiro, G. Online dictionary learning for sparse coding. In *Proceedings of the 26th Annual International Conference on Machine Learning*, ICML '09, pp. 689–696, New York, NY, USA, June 2009. Association for Computing Machinery. ISBN 978-1-60558-516-1. doi: 10.1145/1553374.1553463. URL https://dl.acm.org/doi/10.1145/1553374.1553463.

Mairal, J., Bach, F., and Ponce, J. Sparse Modeling for Image and Vision Processing, December 2014. URL http://arxiv.org/abs/1411.3230. arXiv:1411.3230 [cs].

Makhzani, A. and Frey, B. k-sparse autoencoders, 2014. URL https://arxiv.org/abs/1312.5663.

Malézieux, B., Moreau, T., and Kowalski, M. Understanding approximate and unrolled dictionary learning for pattern recovery. In *International Conference on Learning Representations*, 2022. URL https://openreview.net/forum?id=rI0LYgGeYaw.

Mallat, S. and Zhang, Z. Matching pursuits with time-frequency dictionaries. *IEEE Transactions on Signal Processing*, 41 (12):3397–3415, 1993. doi: 10.1109/78.258082.

Moreau, T. and Bruna, J. Understanding Trainable Sparse Coding via Matrix Factorization, May 2017. URL http://arxiv.org/abs/1609.00285. arXiv:1609.00285 [stat].

Natarajan, B. K. Sparse Approximate Solutions to Linear Systems. *SIAM Journal on Computing*, 24(2):227–234, April 1995. ISSN 0097-5397. doi: 10.1137/S0097539792240406. URL https://epubs.siam.org/doi/10.1137/S0097539792240406. Publisher: Society for Industrial and Applied Mathematics.

Olah, C., Cammarata, N., Schubert, L., Goh, G., Petrov, M., and Carter, S. Zoom In: An Introduction to Circuits. *Distill*, 5 (3):e00024.001, March 2020. ISSN 2476-0757. doi: 10.23915/distill.00024.001. URL https://distill.pub/2020/circuits/zoom-in.

Olshausen, B. A. and Field, D. J. Emergence of simple-cell receptive field properties by learning a sparse code for natural images. *Nature*, 381(6583):607–609, 1996.

Olshausen, B. A. and Field, D. J. Sparse coding with an overcomplete basis set: A strategy employed by v1? *Vision research*, 37(23):3311–3325, 1997.

Oquab, M., Darcet, T., Moutakanni, T., Vo, H., Szafraniec, M., Khalidov, V., Fernandez, P., Haziza, D., Massa, F., El-Nouby, A., et al. Dinov2: Learning robust visual features without supervision. *arXiv preprint arXiv:2304.07193*, 2023.

Park, K., Choe, Y. J., and Veitch, V. The Linear Representation Hypothesis and the Geometry of Large Language Models, July 2024. URL http://arxiv.org/abs/2311.03658. arXiv:2311.03658 [cs].

Pati, Y., Rezaiifar, R., and Krishnaprasad, P. Orthogonal matching pursuit: recursive function approximation with applications to wavelet decomposition. In *Proceedings of 27th Asilomar Conference on Signals, Systems and Computers*, pp. 40–44 vol.1, 1993. doi: 10.1109/ACSSC.1993.342465.

Peotta, L. and Vandergheynst, P. Matching pursuit with block incoherent dictionaries. *IEEE transactions on signal processing*, 55(9):4549–4557, 2007.

Radford, A., Kim, J. W., Hallacy, C., Ramesh, A., Goh, G., Agarwal, S., Sastry, G., Askell, A., Mishkin, P., Clark, J., et al. Learning transferable visual models from natural language supervision. In *International conference on machine learning*, pp. 8748–8763. PmLR, 2021.

Rajamanoharan, S., Lieberum, T., Sonnerat, N., Conmy, A., Varma, V., Kramár, J., and Nanda, N. Jumping ahead: Improving reconstruction fidelity with jumprelu sparse autoencoders. *arXiv preprint arXiv:2407.14435*, 2024.

Rambhatla, S., Li, X., and Haupt, J. Noodl: Provable online dictionary learning and sparse coding. In *Proceedings of International Conference on Learning Representations*, pp. 1–11, 2018.

Ribeiro, M. T., Singh, S., and Guestrin, C. "Why Should I Trust You?": Explaining the Predictions of Any Classifier, August 2016. URL http://arxiv.org/abs/1602.04938. arXiv:1602.04938 [cs].

Rubinstein, R., Bruckstein, A. M., and Elad, M. Dictionaries for Sparse Representation Modeling. *Proceedings of the IEEE*, 98(6):1045–1057, June 2010. ISSN 1558-2256. doi: 10.1109/JPROC.2010.2040551. URL https://ieeexplore.ieee.org/document/5452966/.

Tibshirani, R. Regression shrinkage and selection via the lasso. *Journal of the Royal Statistical Society. Series B (Methodological)*, 58(1):267–288, 1996. ISSN 00359246.

Tolooshams, B. *Deep Learning for Inverse Problems in Engineering and Science*. PhD thesis, Harvard University, 2023.

Tolooshams, B. and Ba, D. E. Stable and interpretable unrolled dictionary learning. *Transactions on Machine Learning Research*, 2022.

Tolooshams, B., Dey, S., and Ba, D. Deep residual autoencoders for expectation maximization-inspired dictionary learning. *IEEE Transactions on neural networks and learning systems*, 32(6):2415–2429, 2020.

Tosic, I. and Frossard, P. Dictionary Learning. *IEEE Signal Processing Magazine*, 28(2):27–38, March 2011. ISSN 1558-0792. doi: 10.1109/MSP.2010.939537. URL https://ieeexplore.ieee.org/abstract/document/5714407.

Tropp, J. Greed is good: algorithmic results for sparse approximation. *IEEE Transactions on Information Theory*, 50(10): 2231–2242, October 2004. ISSN 1557-9654. doi: 10.1109/TIT.2004.834793. URL https://ieeexplore.ieee.org/abstract/document/1337101.

Zhai, X., Mustafa, B., Kolesnikov, A., and Beyer, L. Sigmoid loss for language image pre-training. In *Proceedings of the IEEE/CVF international conference on computer vision*, pp. 11975–11986, 2023.

6. Sequential Reconstruction on MNIST

Figure 6 illustrates how each model reconstructs the input step by step, starting from the pre-activation bias $\boldsymbol{b}_{\text{pre}}$. For shallow SAEs, atoms are reordered by their activation values in the sparse code $\boldsymbol{z}$ to simulate a sequential inference process. Note that for ReLU, JumpReLU, and BatchTopK, the number of selected atoms may differ from $k = 10$, as these methods do not enforce a fixed sparsity level.

MP-SAE exhibits a clear coarse-to-fine reconstruction pattern: with just two atoms, the model already recovers the input's global structure—a zero with an internal pen stroke. Subsequent atoms progressively refine the digit's contour using precise pen-stroke components, highlighting the hierarchical behavior of MP-SAE.

In contrast, ReLU fails to recover the inner stroke, likely because its dictionary contains few atoms resembling pen strokes and is dominated by full-digit prototypes. JumpReLU, TopK, and BatchTopK reconstruct the digit by combining multiple pen-stroke atoms, both for the outer zero shape and the internal stroke, relying on distributed, part-based representations.

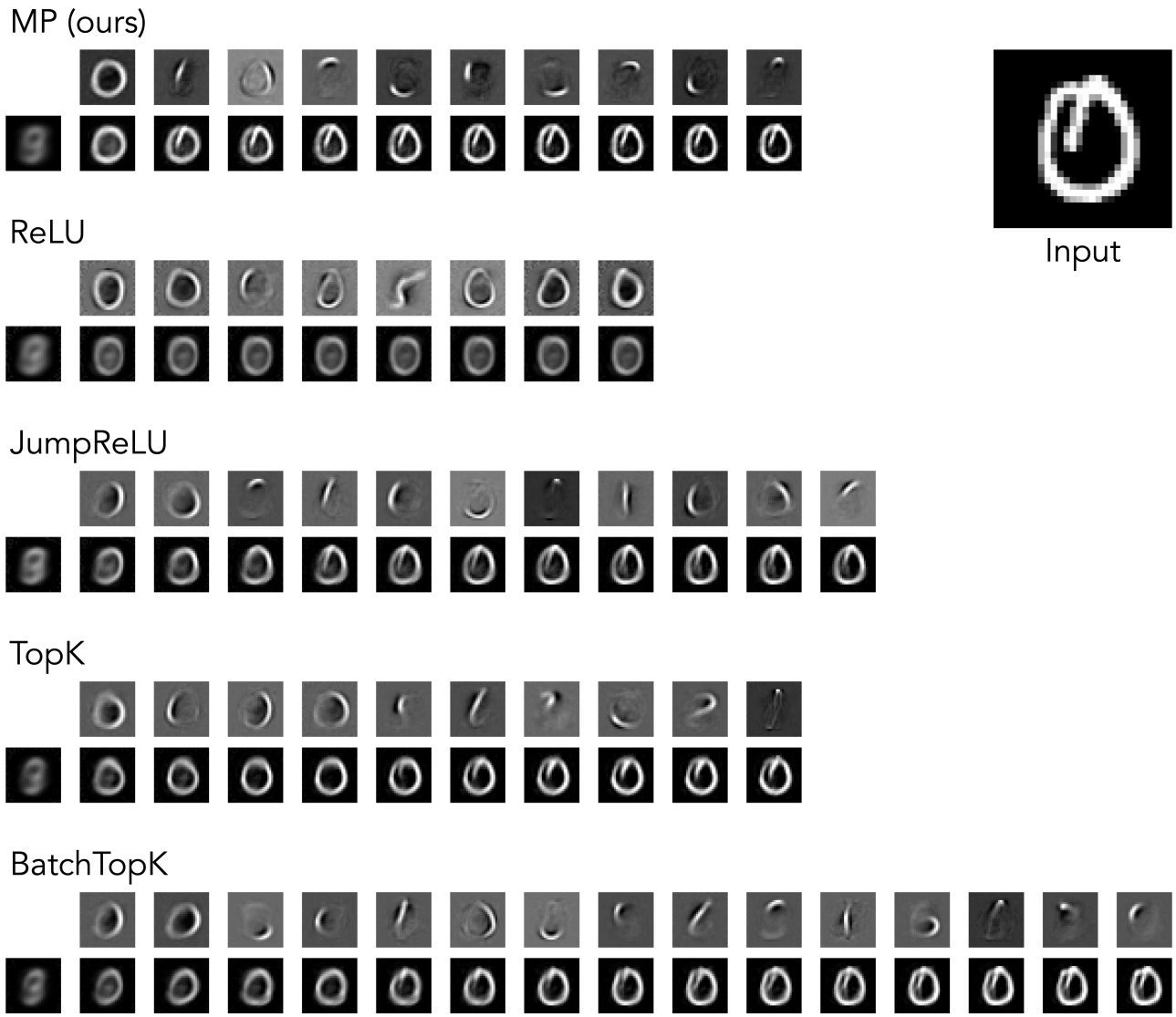

Figure 6: Example of Sequential Reconstruction for $k = 10$.

7. Results on Large Vision Models

We evaluate MP-SAE on large vision models and compare it to three shallow SAEs: Vanilla (ReLU), TopK, and BatchTopK. Our results show that the findings observed on MNIST generalize to this setting.

Expressivity We first assess the representational expressivity of MP-SAE relative to standard SAEs. Figure 7 presents the Pareto frontier obtained by varying the sparsity level while keeping the dictionary size p fixed. Across all evaluated models—SigLIP (Zhai et al., 2023), DINOv2 (Oquab et al., 2023), CLIP (Radford et al., 2021), and ViT (Dosovitskiy et al., 2020)—MP-SAE consistently achieves higher R^2 values at similar sparsity levels, indicating more efficient reconstructions.

Training was conducted for 50 epochs using the Adam optimizer, with an initial learning rate of $5 \cdot 10^{-4}$ decayed to 10^{-6} via cosine annealing with warmup. All SAEs used an expansion factor of 25 ($p = 25m$). Models were trained on the ImageNet-1k (Deng et al., 2009) training set, using frozen features from the final layer of each backbone. For ViT-style models (e.g., DINOv2), we included both the CLS token and all spatial tokens (approximately 261 tokens per image for DINOv2), resulting in roughly 25 billion training tokens overall.

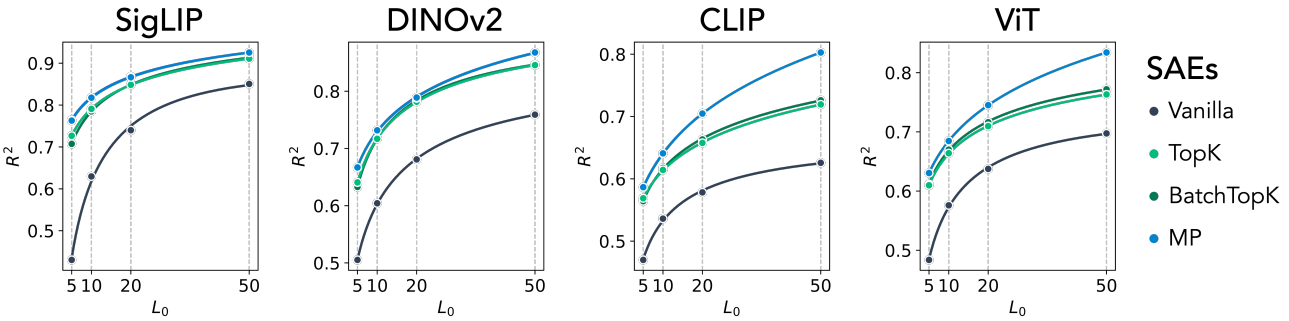

Figure 7: **MP-SAE recovers more expressive atoms than standard SAEs.** Reconstruction performance (R^2) as a function of sparsity level across four pretrained vision models: SigLIP, DINOv2, CLIP, and ViT. MP-SAE consistently yields higher R^2 at comparable sparsity, suggesting more informative and efficient decompositions.

Coherence To evaluate coherence beyond pairwise similarity, we use the Babel function (Tropp, 2004), a standard metric in sparse approximation that captures cumulative interference among dictionary atoms. Mutual coherence reflects only the maximum absolute inner product between pairs of atoms; as a result, it can be misleadingly high if just one pair has a large inner product while the rest are small. In contrast, the Babel function measures the total interference between an atom and a group of others, offering a more comprehensive view of redundancy. Intuitively, $\mu_1(r)$ quantifies how well a single atom can be approximated by a group of r others. Lower values indicate lower redundancy, reflecting the degree of overlap between an atom and its r nearest neighbors in the dictionary.

Figure 8 reports $\mu_1(r)$ for both the full dictionary (top) and for subsets of atoms co-activated at inference (bottom). As observed on MNIST, MP-SAE learns globally coherent dictionaries with low Babel scores. However, the subsets of atoms selected during inference exhibit higher Babel values, indicating local incoherence. This duality in coherence persists even when training on large vision models.

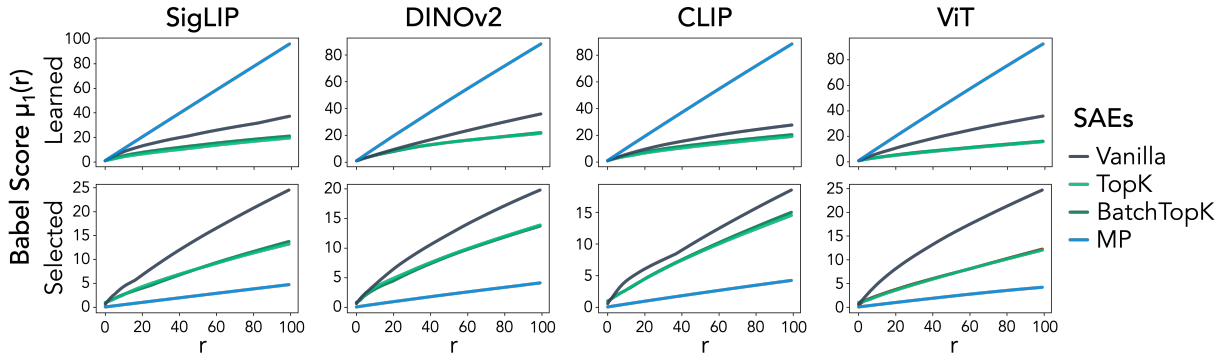

Figure 8: **MP-SAE learns more coherent dictionaries but selects incoherent atoms.** Babel scores for the full dictionaries (top) and co-activated subsets at inference time (bottom).

8. Theoretical Properties of Matching Pursuit

We restate three foundational properties of Matching Pursuit—originally established in the sparse coding literature (Mallat & Zhang, 1993)—and interpret them in the context of sparse autoencoders. These properties help elucidate the structure and dynamics of the representations learned by MP-SAE.

- **Stepwise orthogonality** (Proposition 8.1): at each iteration, the residual becomes orthogonal to the atom most recently selected by the greedy inference rule. This sequential orthogonalization mechanism gives rise to a locally disentangled structure in the representation and reflects the conditional independence induced by MP-SAE inference.
- **Monotonic decrease of residual energy** (Proposition 8.2): the ℓ_2 norm of the residual decreases whenever it retains a nonzero projection onto the span of the dictionary. This guarantees that inference steps lead to progressively refined reconstructions, and enables sparsity to be adaptively tuned at inference time without retraining.
- **Asymptotic convergence** (Proposition 8.3): in the limit of infinite inference steps, the reconstruction converges to the orthogonal projection of the input onto the subspace defined by the dictionary. Thus, MP-SAE asymptotically recovers all structure that is representable within its learned basis.

Proposition 8.1 (Stepwise Orthogonality of MP Residuals)**.** *Let $\boldsymbol{r}^{(t)}$ denote the residual at iteration t of MP-SAE inference, and let $j^{(t)}$ be the index of the atom selected at step t. If the column $j^{(t)}$ of the dictionary $\boldsymbol{D}$ satisfy $\|\boldsymbol{D}_{j^{(t)}}\|_2 = 1$, then the residual becomes orthogonal to the previously selected atom:*

$$\boldsymbol{D}_{j^{(t)}}^\top \boldsymbol{r}^{(t)} = 0.$$

Proof. This follows from the residual update:

$$\boldsymbol{r}^{(t)} = \boldsymbol{r}^{(t-1)} - \boldsymbol{D}_{j^{(t)}} \boldsymbol{z}_{j^{(t)}}^{(t)},$$

with $\boldsymbol{z}_{j^{(t)}}^{(t)} = \boldsymbol{D}_{j^{(t)}}^\top \boldsymbol{r}^{(t-1)}$. Taking the inner product with $\boldsymbol{D}_{j^{(t)}}$ gives:

$$\boldsymbol{D}_{j^{(t)}}^\top \boldsymbol{r}^{(t)} = \boldsymbol{D}_{j^{(t)}}^\top \boldsymbol{r}^{(t-1)} - \|\boldsymbol{D}_{j^{(t)}}\|^2 \boldsymbol{z}_{j^{(t)}}^{(t)} = \boldsymbol{z}_{j^{(t)}}^{(t)} - \boldsymbol{z}_{j^{(t)}}^{(t)} = 0. \qquad \square$$

This result captures the essential inductive step of Matching Pursuit: each update removes variance along the most recently selected atom, producing a residual that is orthogonal to it. Applied iteratively, this localized orthogonality promotes the emergence of conditionally disentangled structure in MP-SAE. In contrast, other sparse autoencoders lack this stepwise orthogonality mechanism, which helps explain the trend observed in the Babel function during inference in Figure 5.

Proposition 8.2 (Monotonic Decrease of MP Residuals)**.** *Let $\boldsymbol{r}^{(t)}$ denote the residual at iteration t of MP-SAE inference, and let $\boldsymbol{z}_{j^{(t)}}^{(t)}$ be the nonzero coefficient selected at that step, Then the squared residual norm decreases monotonically:*

$$\|\boldsymbol{r}^{(t)}\|_2^2 - \|\boldsymbol{r}^{(t-1)}\|_2^2 = -\|\boldsymbol{D}_{j^{(t)}} \boldsymbol{z}_{j^{(t)}}^{(t)}\|_2^2 \leq 0.$$

Proof. From the residual update:

$$\boldsymbol{r}^{(t)} = \boldsymbol{r}^{(t-1)} - \boldsymbol{D}_{j^{(t)}} \boldsymbol{z}_{j^{(t)}}^{(t)},$$

we can rearrange to write:

$$\boldsymbol{r}^{(t-1)} = \boldsymbol{r}^{(t)} + \boldsymbol{D}_{j^{(t)}} \boldsymbol{z}_{j^{(t)}}^{(t)}.$$

Taking the squared norm of both sides:

$$\begin{aligned}\|\boldsymbol{r}^{(t-1)}\|_2^2 &= \|\boldsymbol{r}^{(t)} + \boldsymbol{D}_{j^{(t)}} \boldsymbol{z}_{j^{(t)}}^{(t)}\|_2^2 \\ &= \|\boldsymbol{r}^{(t)}\|_2^2 + 2\langle \boldsymbol{r}^{(t+1)}, \boldsymbol{D}_{j^{(t)}} \rangle \boldsymbol{z}_{j^{(t)}}^{(t)} + \|\boldsymbol{D}_{j^{(t)}} \boldsymbol{z}_{j^{(t)}}^{(t)}\|_2^2.\end{aligned}$$

By Proposition 8.1, the cross term vanishes:

$$\langle \boldsymbol{r}^{(t)}, \boldsymbol{D}_{j^{(t)}} \rangle = 0,$$

yielding:

$$\|\boldsymbol{r}^{(t-1)}\|_2^2 = \|\boldsymbol{r}^{(t)}\|_2^2 + \|\boldsymbol{D}_{j^{(t)}} \boldsymbol{z}_{j^{(t)}}^{(t)}\|_2^2. \qquad \square$$

The monotonic decay of residual energy ensures that each inference step yields an improvement in reconstruction, as long as the residual lies within the span of the dictionary. Crucially, this property enables MP-SAE to support adaptive inference-time sparsity: the number of inference steps can be varied at test time—independently of the training setup—while still allowing the model to progressively refine its approximation.

As shown in Figure 9, MP-SAE exhibits a continuous decay in reconstruction error—a behavior explained by the proposition and not guaranteed by other sparse autoencoders. All models were trained on DINOv2 representations with different training-time ℓ_0 sparsity levels. At inference, the sparsity k is varied to assess generalization. MP-SAE shows monotonic improvement, as guaranteed by Proposition 8.2. This stands in contrast to TopK-based SAEs, which often degrade under sparsity mismatch: when trained with fixed k, the decoder implicitly specializes to superpositions of exactly k features, leading to instability—particularly when the inference-time k is much larger than the training value. ReLU-based SAEs, by contrast, cannot expand their support beyond the features activated during training and thus exhibit flat or plateaued performance as k increases.

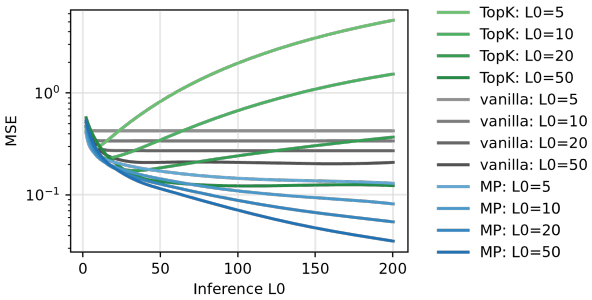

Figure 9: **Reconstruction error vs. inference-time sparsity k.**

Proposition 8.3 (Asymptotic Convergence of MP Residuals)**.** *Let $\hat{\boldsymbol{x}}^{(t)} = \boldsymbol{x} - \boldsymbol{r}^{(t)}$ denote the reconstruction at iteration t, and let $\mathbf{P}_{\boldsymbol{D}}$ be the orthogonal projector onto* $\mathrm{span}(\boldsymbol{D})$*. Then:*

$$\lim_{t\to\infty} \|\hat{\boldsymbol{x}}^{(t)} - \mathbf{P}_{\boldsymbol{D}}\boldsymbol{x}\|_2 = \lim_{t\to\infty} \|\mathbf{P}_{\boldsymbol{D}}\boldsymbol{r}^{(t)}\|_2 = 0.$$

This convergence result is formally established in the original Matching Pursuit paper (Mallat & Zhang, 1993)[Theorem 1]. This result implies that MP-SAE progressively reconstructs the component of $\boldsymbol{x}$ that lies within the span of the dictionary, converging to its orthogonal projection in the limit of infinite inference steps. When the dictionary is complete (i.e., $\mathrm{rank}(\boldsymbol{D}) = m$), this guarantees convergence to the input signal $\boldsymbol{x}$.