SPARSE AUTOENCODERS DO NOT FIND CANONICAL UNITS OF ANALYSIS

Patrick Leask[*]
Department of Computer Science
Durham University
patrickaaleask@gmail.com

Bart Bussmann[*]
Independent
bartbussmann@gmail.com

Michael Pearce
Independent

Joseph Bloom
Decode Research

Curt Tigges
Decode Research

Noura Al Moubayed
Department of Computer Science
Durham University

Lee Sharkey
Apollo Research

Neel Nanda

ABSTRACT

A common goal of mechanistic interpretability is to decompose the activations of neural networks into features: interpretable properties of the input computed by the model. Sparse autoencoders (SAEs) are a popular method for finding these features in LLMs, and it has been postulated that they can be used to find a *canonical* set of units: a unique and complete list of atomic features. We cast doubt on this belief using two novel techniques: SAE stitching to show they are incomplete, and meta-SAEs to show they are not atomic. SAE stitching involves inserting or swapping latents from a larger SAE into a smaller one. Latents from the larger SAE can be divided into two categories: *novel latents*, which improve performance when added to the smaller SAE, indicating they capture novel information, and *reconstruction latents*, which can replace corresponding latents in the smaller SAE that have similar behavior. The existence of novel features indicates incompleteness of smaller SAEs. Using meta-SAEs - SAEs trained on the decoder matrix of another SAE - we find that latents in SAEs often decompose into combinations of latents from a smaller SAE, showing that larger SAE latents are not atomic. The resulting decompositions are often interpretable; e.g. a latent representing "Einstein" decomposes into "scientist", "Germany", and "famous person". Even if SAEs do not find canonical units of analysis, they may still be useful tools. We suggest that future research should either pursue different approaches for identifying such units, or pragmatically choose the SAE size suited to their task. We provide an interactive dashboard to explore meta-SAEs: https://metasaes.streamlit.app/

1 INTRODUCTION

Mechanistic interpretability aims to reverse-engineer neural networks into human-interpretable algorithms (Olah et al., 2020; Meng et al., 2022; Geva et al., 2023; Nanda et al., 2023; Elhage et al., 2021). A key challenge of mechanistic interpretability is identifying the correct units of analysis — fundamental components that can be individually understood and collectively explain the network's function. Ideally, these units would be *unique*, with no variations (Bricken et al., 2023); *complete*, encompassing all necessary features (Elhage et al., 2022); and *atomic* or *irreducible*, indivisible into smaller components (Engels et al., 2024). We refer to a set of units with all of these properties as **canonical**.

Initially, researchers hoped that individual MLP neurons (Meng et al., 2022; Olah et al., 2020) and attention heads (Wang et al., 2022; Olsson et al., 2022) could serve as these units. However, these

[*] These authors contributed equally to this work.

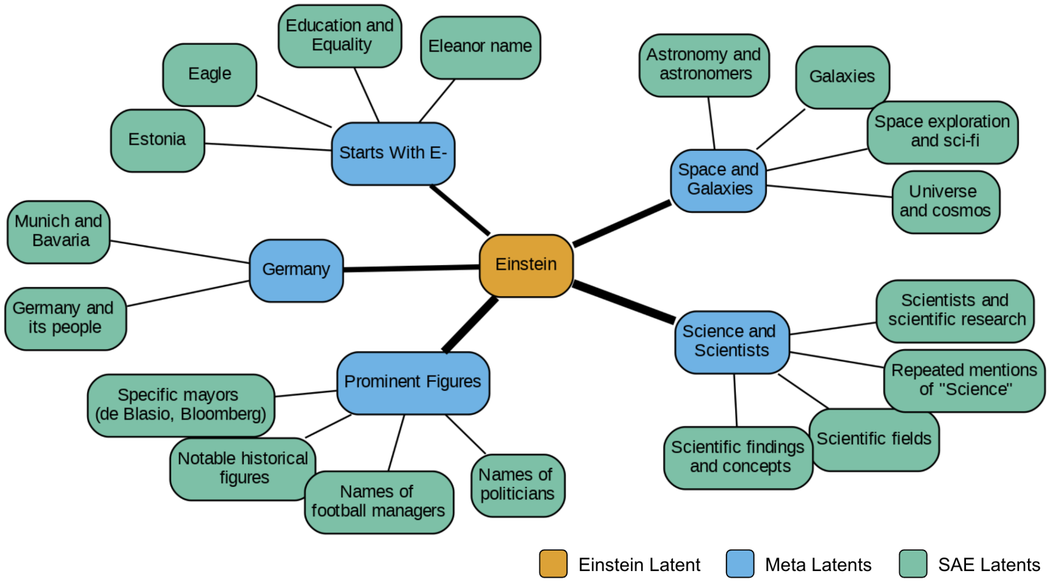

Figure 1: Decomposition of an SAE latent representing "Einstein" into a set of interpretable meta-latents. The edges connecting the nodes indicate shared activation by a meta-latent, with thicker lines representing stronger connections. It demonstrates the ability of meta-SAEs to uncover the underlying compositional structure of SAE latents, revealing how a complex concept can be represented as a sparse combination of meta-latents. We built a dashboard where you can explore all meta-latents: `https://metasaes.streamlit.app`

proved insufficient for interpretability due to polysemanticity, where a single neuron responds to multiple unrelated concepts (Olah et al., 2020; Elhage et al., 2022).

Recently, sparse autoencoders (SAEs) have emerged as a promising alternative by decomposing the activations of LLMs into a dictionary of interpretable and monosemantic features (Bricken et al., 2023; Cunningham et al., 2023). A key hyperparameter when training SAEs is the dictionary size, i.e. number of latent units. Previous work conjectured that SAEs might identify a set of "true features" with sufficient dictionary size (Bricken et al., 2023), i.e. the canonical features that are the goal of much mechanistic interpretability research.

One challenge to the theory that SAEs identify a canonical set of units is the phenomenon of *feature splitting*, where latents from smaller SAEs "split" into multiple, more fine-grained latents in larger SAEs (Bricken et al., 2023). For example, Bricken et al. (2023) find a base64 feature that splits into three features in a larger SAE: activating on letters, digits, and encoded ASCII in base64 text. Furthermore, Templeton (2024) finds that a larger SAE has latents that activate on certain specific individual chemical elements that a smaller SAE did not represent. Currently, the effect of dictionary size on the features has not been systematically studied, in part because we lack good methods to compare latents found in SAEs of different sizes.

To better understand how SAEs of different sizes capture features, we develop a method called SAE stitching. When stitching SAEs, we systematically swap clusters of latents between SAEs of different dictionary sizes based on their cosine similarity. Through this method, we observe that larger SAEs learn both more fine-grained versions of latents found in smaller SAEs, but also entirely novel latents. The existence of novel latents suggests that the reconstruction error of smaller SAEs is partially due to missing out information altogether, not just imperfect approximations to features or overly coarse latents, indicating *incompleteness*.

Contrary to the story of feature splitting, we observed that some reconstruction latents had split from *multiple* latents in the smaller SAE, indicating those latents were composing into more complex latents. Previous work has predicted that the sparsity penalty incentivizes latents to represent composed features, even if they are independent (Wattenberg & Viégas, 2024; Bricken et al., 2023; Anders et al., 2024). For instance, consider a neural network that represents color and shape fea-

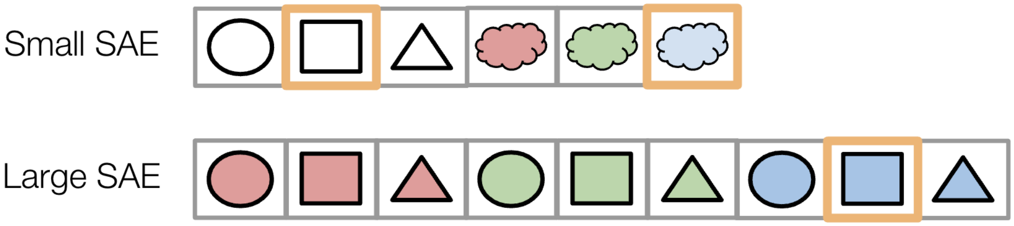

Figure 2: Example of composition of latents in SAEs of different sizes. The smaller SAE has six latents, three of which reconstruct shape features, and three of which reconstruct color features. Reconstructing a shape of a specific color requires two active latents (e.g. blue and square). On the other hand, the larger SAE has nine latents, each of which reconstructs a different color and shape combination. In the larger SAE, only a single active latent is required to reconstruct the colored shape (e.g. blue square). The sparsity penalty incentivizes larger SAEs to learn compositions of latents rather than atomic latents.

tures, each with three values (red/green/blue and circle/square/triangle). A small SAE might learn a latent for each value. A large SAE, however, might learn latents for all 9 color-shape combinations (i.e. blue square) instead of the 6 fundamental features (Smith, 2024). The large SAE can represent this is sparser as only one latent activates per input rather than two, see Figure 2.

To investigate this, we introduce meta-SAEs, which are SAEs trained to find sparse decompositions of the decoder directions of another SAE. These decompositions are often interpretable, e.g. a latent representing “Einstein” decomposes into meta-latents representing “scientist”, “Germany”, and “prominent figures”, among others (see Figure 1). This shows that latents are often not *atomic*, especially in larger SAEs. We find that meta-latents are similar to latents in smaller SAEs, demonstrating that latents from larger SAEs can be interpreted as the composition of latents from smaller SAEs.

In summary, our contributions are:

1. **SAE stitching**, as a method for comparing latents across different sizes of SAE. Latents in a larger SAE are either novel latents, missing in smaller SAEs, or reconstruction latents, similar to some latents in smaller SAEs.
2. **Meta-SAEs**, as an approach for decomposing the decoder directions of SAEs into interpretable, monosemantic meta-latents.

Our empirical results suggest that simply training larger SAEs is unlikely to result in a canonical set of units for all mechanistic interpretability tasks, and that the choice of dictionary size is subjective. We suggest taking a pragmatic approach to applying SAEs to mechanistic interpretability tasks, trying SAEs of several widths to see which is best suited. We are uncertain whether canonical units of analysis exist, but our results suggest that alternative approaches should be explored.

2 Sparse Autoencoders

Sparse dictionary learning is the problem of finding a decomposition of a signal that is both sparse and overcomplete (Olshausen & Field, 1997). Lee et al. (2007) initially applied the sparsity constraint to deep belief networks, with SAEs later being applied to the reconstruction of neural network activations (Bricken et al., 2023; Cunningham et al., 2023). In the context of large language models, SAEs decompose model activations $\mathbf{x} \in \mathbb{R}^n$ into sparse linear combinations of learned directions, which are often interpretable and monosemantic.

An SAE consists of an encoder and a decoder:

$$\mathbf{f}(\mathbf{x}) := \sigma(\mathbf{W}^{\text{enc}}\mathbf{x} + \mathbf{b}^{\text{enc}}), \tag{1}$$

$$\hat{\mathbf{x}}(\mathbf{f}) := \mathbf{W}^{\text{dec}}\mathbf{f} + \mathbf{b}^{\text{dec}}. \tag{2}$$