

Figure 5: Interpolating between SAE pairs of increasing dictionary size (768→1536→3072→6144→12288) through two steps per phase: adding novel latents (increasing L0) then swapping groups of reconstruction latents (decreasing L0 on average). Both steps on average improve reconstruction (MSE). The L0 and MSE are averages over input samples.

Figure 5 shows four transitions between pairs of SAEs of increasing size. For each transition, we first identify novel latents from the larger SAE and add them one by one, increasing the average L0 since we are inserting extra latents. We then identify groups of related reconstruction latents as defined by our bipartite graph and swap each group - removing similar latents from the smaller SAE and replacing them with their counterparts in the larger SAE. This decreases L0 on average since the larger SAE represents similar information more sparsely. Each step leads on average to improved reconstruction performance, allowing us to interpolate between different SAEs in terms of dictionary size, sparsity, and reconstruction performance.

The existence of the novel group of latents demonstrates that smaller SAEs are *incomplete*, and that larger SAE learn features that are missed by the smaller SAE. The reconstruction group seems to largely reconstruct the same features as the latents in the smaller SAE, but uses more features to achieve lower sparsity (Appendix Figure 14). In some cases, a single latent splits into two more specific latents - this is the feature splitting observed in Bricken et al. (2023). However, we find some reconstruction group latents have high decoder cosine similarity to *multiple* latents in the smaller SAE, suggesting the large SAE latent is an interpolation or composition of the smaller SAE latents (Appendix A.4). This supports the predictions of Wattenberg & Viégas (2024); Bricken et al. (2023); Anders et al. (2024) that the sparsity penalty results in the undesirable composition of features that may be sparser, but do not add any new information. We explore this phenomenon further in Section 5.

5 META-SAEs

In Section 4, we demonstrated through SAE stitching that increasing dictionary size leads to larger SAEs learning not only novel features, but also reconstruction latents that encode similar information to the latents in smaller SAEs. We found that some of these reconstruction latents have high cosine similarity with *multiple* latents in the smaller SAE, see Appendix A.4. This suggests that these smaller SAE latents are composing into more complex latents, such as in the example of a latent representing “blue” and another latent representing “square” combining in a “blue square”-latent (see Figure 2). If large SAE features are indeed compositions rather than *atomic*, it may be possible to decompose them into more fundamental units.

Table 1: Example GPT-2 SAE latents and their meta-latent decompositions, with model-generated explanations (human-summarized) of what the latents activate on. More latent decompositions can be found on our interactive dashboard: <https://metasaes.streamlit.app>

SAE Latent Description	Meta-Latent Descriptions
Albert Einstein	Science & Scientists, Famous People, Space & Astronomy, Germany, Electricity, Words starting with a capital E
Rugby	Sports activities, Words starting with ‘R’, References to Ireland, References to sports leagues, activities & actions
Android Operating System	Mobile phones, operating systems, Californian cities

To decompose the latents of larger SAEs, we introduce meta-SAEs. Meta-SAEs are SAEs trained to reconstruct the decoder directions $\mathbf{W}_i^{\text{dec}}$ of a standard SAE using a dictionary of meta-latents, rather than reconstructing network activations. That is, we treat the latents $\mathbf{W}_i^{\text{dec}}$ as the training data for our meta-SAE. We find that meta-latents and meta-SAE decompositions are interpretable, with the meta-latents being monosemantic. Table 1 provides some examples of meta-SAE decompositions.

The latents of meta-SAEs have similar decoder directions to those found in SAEs of comparable size trained directly on the same network activations. This observation further supports the hypothesis that larger SAE latents are not entirely new features but may be compositions of features already learned, albeit less precisely, by smaller models.

We use the BatchTopK SAE (Bussmann et al., 2024) to train our meta-SAEs. We train our meta-SAE on the decoder directions of the GPT-2 SAE with dictionary size 49152. The meta-SAE has a dictionary size of 2304 meta-latents, with on average 4 of meta-latents active per SAE latent. Due to small number of training samples for the meta-SAE (49152), the meta-SAE is trained for 2000 epochs. We use the Adam optimizer with learning rate $1e-4$ and a batch size of 4096. After training, the meta-SAE explains 55.47% of the variance of the decoder directions of the SAE.

5.1 EVALUATING META-SAE DECOMPOSITIONS

We follow the lead of (Bills et al., 2023; Bricken et al., 2023; Cunningham et al., 2023; Rajamanoharan et al., 2024b) in evaluating neural network and SAE latents using automated interpretability with LLMs.

First, we generate explanations of SAE latents by presenting GPT-4o-mini with a list of input sequences that activate an SAE latent to varying degrees, and prompting it to generate a natural language explanation of the feature consistent with the activations. Second, we collect all of the SAE latents on which a meta-SAE latent is active, and prompt again with the explanation and a number of top activating examples of each of the SAE latents, asking the model to provide an explanation of the common behavior of the SAE latents, which becomes the meta-SAE latent explanation.

We evaluate the meta-SAE latent explanations in a zero-shot multiple-choice-question setting. For a given latent we prompt GPT-4o-mini with the explanations of the meta-SAE latents that are active on that latent, and ask it to choose which of 5 SAE latent explanations most relate to the explanations of the meta-SAE latents. One of these SAE latent explanations is of the correct latent, with the remaining 4 explanations corresponding to random latents from the SAE. On a random sample of 1,000 SAE latents, GPT-4o-mini chose the correct answer of the five options 73% of the time.

5.2 COMPARISON TO SMALLER SAE LATENTS

In Section 4, we hypothesised that latents in larger SAEs can be described as the composition of latents in smaller SAEs. We find that meta-SAEs learn similar latents to similar sized SAEs trained on the original reconstruction problem. Plots of the maximum cosine similarity between meta-SAE latents and latents from SAEs of different sizes are shown in Figure 6. We validate this by replacing meta-SAE decoder directions with the most similar SAE decoder direction, and retrained the encoder. This results in only a small decrease in meta-SAE reconstruction performance (Appendix Figure 20). This suggests that larger SAE latents are indeed composed of latents from smaller SAEs.

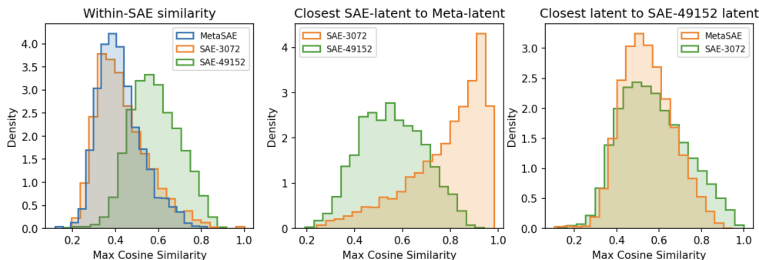


Figure 6: Cosine similarity between SAE latents and meta-SAE latents. Note the high maximum cosine similarity between latents from a meta-SAE with 2304 latents, and a standard SAE with 3072 latents.

6 CONCLUSION

Our findings challenge the idea that SAEs can discover a canonical set of features. Through SAE stitching, we demonstrated that smaller SAEs are incomplete, missing information that novel features in larger SAEs capture. Moreover, our meta-SAE experiments showed that, due to the sparsity penalty, latents in larger SAEs are often not atomic but compositions of interpretable meta-latents. These findings suggest that there is no single SAE width at which it learns a unique and complete dictionary of atomic features that can be used to explain the behavior of the model.

These results imply that rather than converging on a unique, complete, and irreducible set of features, SAEs of different sizes offer varying granularities and compositions of features. This indicates that the choice of SAE size should be guided by the specific interpretability task at hand, accepting that no single SAE configuration provides a universal solution. However, our methods neither identify canonical units of analysis, nor the size of dictionary to use for a given task. Furthermore, our work only studies two LLMs and does not include very large SAEs, such as in Templeton (2024). We also acknowledge that the use of SAEs in mechanistic interpretability is nascent, and whilst early results are encouraging, associating the learned latents of SAEs with interpretable concepts is still an open problem. In conclusion, our research suggests that alternative methods are required for identifying canonical units, and that SAE practitioners should embrace a pragmatic approach towards choosing dictionary size when using SAEs on interpretability tasks such as probing, unlearning and steering.

REFERENCES

- Evan Anders, Clement Neo, Jason Hoelscher-Obermaier, and Jessica N. Howard. Sparse autoencoders find composed features in small toy models. <https://www.lesswrong.com/posts/a5wwqza2cY3W7L9cj/sparse-autoencoders-find-composed-features-in-small-toy>, 2024.
- Anonymous. Evaluating sparse autoencoders on concept removal tasks, 2024. URL <https://github.com/anonymous664422>. Upcoming work, unofficially published.
- Yamini Bansal, Preetum Nakkiran, and Boaz Barak. Revisiting model stitching to compare neural representations. *Advances in neural information processing systems*, 34:225–236, 2021.
- Steven Bills, Nick Cammarata, Dan Mossing, Henk Tillman, Leo Gao, Gabriel Goh, Ilya Sutskever, Jan Leike, Jeff Wu, and William Saunders. Language models can explain neurons in language models. URL <https://openaipublic.blob.core.windows.net/neuron-explainer/paper/index.html>. (Date accessed: 14.05. 2023), 2, 2023.
- Trenton Bricken, Adly Templeton, Joshua Batson, Brian Chen, Adam Jermy, Tom Conerly, Nick Turner, Cem Anil, Carson Denison, Amanda Askell, et al. Towards monosemanticity: Decomposing language models with dictionary learning. *Transformer Circuits Thread*, 2, 2023.
- Bart Bussmann, Patrick Leask, and Neel Nanda. Batchtopk sparse autoencoders. *arXiv preprint arXiv:2412.06410*, 2024.