Showing SAE Latents Are Not Atomic Using Meta-SAEs

by **Bart Bussmann, Michael Pearce, Patrick Leask, Joseph Bloom, Lee Sharkey, Neel Nanda**

23rd Aug 2024 AI Alignment Forum

Bart, Michael and Patrick are joint first authors. Research conducted as part of MATS 6.0 in Lee Sharkey and Neel Nanda's streams. Thanks to Mckenna Fitzgerald and Robert Krzyzanowski for their feedback!

TL;DR:

- *Sparse Autoencoder (SAE) latents have been shown to typically be monosemantic (i.e. correspond to an interpretable property of the input). It is sometimes implicitly assumed that they are therefore atomic, i.e. simple, irreducible units that make up the model's computation.*
- *We provide evidence against this assumption by finding sparse, interpretable decompositions of SAE decoder directions into seemingly more atomic latents, e.g. Einstein -> science + famous + German + astronomy + energy + starts with E-*
- *We do this by training meta-SAEs, an SAE trained to reconstruct the decoder directions of a normal SAE.*
- *We argue that, conceptually, there's no reason to expect SAE latents to be atomic - when the model is thinking about Albert Einstein, it likely also thinks about Germanness, physicists, etc. Because Einstein always entails those things, the sparsest solution is to have the Albert Einstein latent also boost them.*
- *Key results*
  - *SAE latents can be decomposed into more atomic, interpretable meta-latents.*

    - *We show that when latents in a larger SAE have split out from latents in a smaller SAE, a meta SAE trained on the larger SAE often recovers this structure.*
    - *We demonstrate that meta-latents allow for more precise causal interventions on model behavior than SAE latents on a targeted knowledge editing task.*
- *We believe that the alternate, interpretable decomposition using MetaSAEs casts doubt on the implicit assumption that SAE latents are atomic. We show preliminary results that MetaSAE latents have significant ovelap with latents in a normal SAE of the same size but may relate differently to the larger SAEs used in MetaSAE training.*

We made a dashboard that lets you explore meta-SAE latents.

Terminology: Throughout this post we use "latents" to describe the concrete components of the SAE's dictionary, whereas "feature" refers to the abstract concepts, following Lieberum et al.

Introduction

Mechanistic interpretability (mech interp) attempts to understand neural networks by breaking down their computation into interpretable components. One of the key challenges of this line of research is the polysemanticity of neurons, meaning they respond to seemingly unrelated inputs. Sparse autoencoders (SAEs) have been proposed as a method for decomposing model activations into sparse linear sums of latents. Ideally, these latents should be monosemantic, i.e. respond to inputs that clearly share a similar meaning (implicitly, from the perspective of a human interpreter). That is, a human should be able to reason about the latents both in relation to the features to which they are associated, and also use the latents to better understand the model's overall behavior.

There is a popular notion, both implicitly in related work on SAEs within mech interp and explicitly by the use of the term "atom" in sparse dictionary learning as a whole, that SAE features are atomic or can be "true features". However, monosemanticity does not imply atomicity. Consider the example of shapes of different colors - the set of shapes is [circle, triangle, square], and the set of colors is [white, red, green, black], each of which is represented with a linear direction. 'Red triangle' represents a monosemantic feature, but not an atomic feature, as it can be decomposed into red and triangle. It has been

shown that sufficiently wide SAEs on toy models will learn 'red triangle', rather than representing 'red' and 'triangle' with separate latents.

Furthermore, whilst one may naively reason about SAE latents as bags of words with almost-random directions, there are hints of deeper structure, as argued by Wattenberg et al: UMAP plots (a distance-based dimensionality reduction method) group together conceptually similar latents, suggesting that they share components; and local PCA recovers a globally interpretable timeline direction.

Most notably, feature splitting makes clear that directions are not almost-random - When a latent in a small SAE "splits" into several latents in a larger SAE, the larger SAE latents have significant cosine sim with each other along with semantic connections. Arguably, such results already made it clear that SAE features are not atomic, but we found the results of our investigation sufficiently surprising, that we hope it is valuable to carefully explore and document this phenomena.

We introduce meta-SAEs, where we train an SAE on the decoder weights of an SAE, effectively decomposing the SAE latents into new, sparse, monosemantic latents. For example, we find a decomposition of a latent relating to Albert Einstein into meta-latents for Physics, German, Famous people, and others. Similarly, we find a decomposition of a Paris-related latent into meta-latents for French city names, capital cities, Romance languages, and words ending in the *-us* sound.

In this post we make the following contributions:

- We show that meta-SAEs are a useful tool for exploring and understanding SAE latents through a series of case studies, and provide a dashboard for this exploration.
- We show that when latents in a larger SAE have split out from latents in a smaller SAE, a meta SAE trained on the larger SAE often recovers this structure.
- We demonstrate that meta-latents are useful for performing causal interventions to edit factual knowledge associations in language models on a dataset of city attributes. For example, we find a combination of meta-latents that let us steer Tokyo to speak French and use the Euro, but to remain in Japan.
- We investigate baselines for breaking down larger SAE latents, like taking the latents in a smaller SAE with the highest cosine sim, and show that these are also interpretable, suggesting meta-SAEs are not the only path to these insights.

Whilst our results suggest that SAE latents are not atomic, we do not claim that SAEs are not useful. Rather, we believe that meta-SAEs provide another frame of reference for interpreting the model. In the natural sciences there are multiple levels of abstraction for understanding systems, such as cells and organisms, and atoms and molecules; with each different level being useful in different contexts.

We also note several limitations. Meta-SAEs give a lossy decomposition, i.e. there is an error term, and meta-features may not be intrinsically lower level than SAE features- Albert Einstein is arguably more fine-grained than man, for example, and may be a more appropriate abstraction in certain contexts. We also do not claim that meta-SAEs have found the 'true' atoms of neural computation, and it would not surprise us if they are similarly atomic to normal SAEs of the same width.

Our results shed new light on the atomicity of SAE latents, and suggest a path to exploring feature geometry in SAEs. We also think that meta-latents provide a novel approach for fine-grained model editing or steering using SAE latents.

Defining Meta-SAEs

We use sparse autoencoders as in Towards Monosemanticity and Sparse Autoencoders Find Highly Interpretable Directions. In our setup, the feature activations are computed as:

$$f_i(x) = ReLU(W^{enc}_{i,\cdot} \cdot (x - b^{dec}) + b^{enc}_i)$$

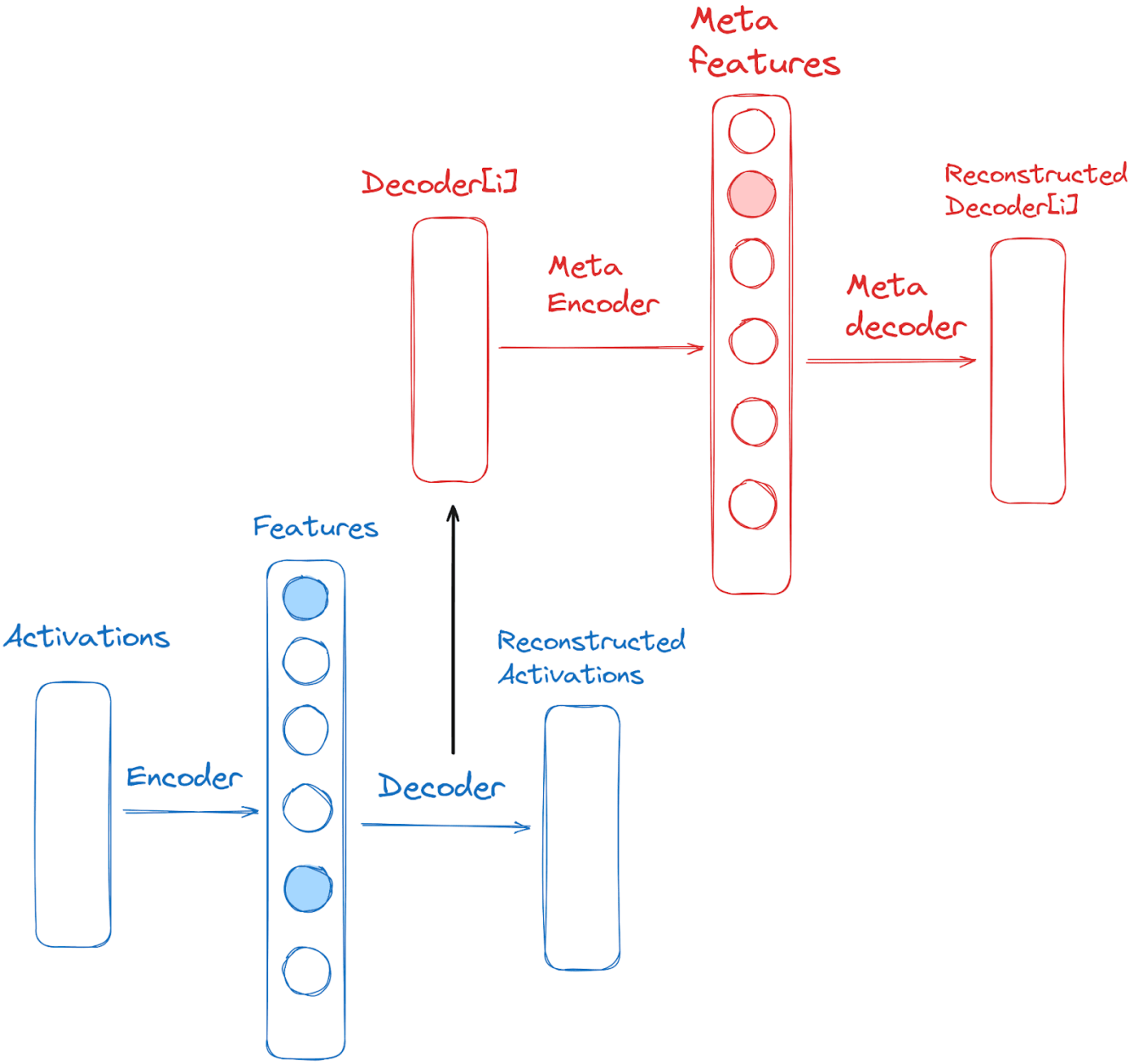
Based on these feature activations, the input is then reconstructed as

$$\hat{x} = b^{dec} + \sum_{i=1}^{F} f_i(x) W^{dec}_{\cdot,i}$$

The encoder and decoder matrices and biases are trained with a loss function that combines an L2 penalty on the reconstruction loss and an L1 penalty on the feature activations:

$$\mathcal{L} = \mathbb{E}_x[||x - \hat{x}||_2^2 + \lambda \sum_{i=1}^{F} f_i(x)]$$

After the SAE is trained on the model activations, we train a meta-SAE. A meta-SAE is trained on batches of the decoder directions W^{dec} rather than model activation X.



*Meta-SAEs are trained on the decoder weights of an SAE, but do **not** depend on the input of the original SAE.*

The meta-SAE is trained on a standard SAE with dictionary size 49152 (providing 49152 input samples for the meta-SAE) trained on the gpt2-small residual stream before layer 8 and is one of the same SAEs as was used in Stitching SAEs of different sizes°.

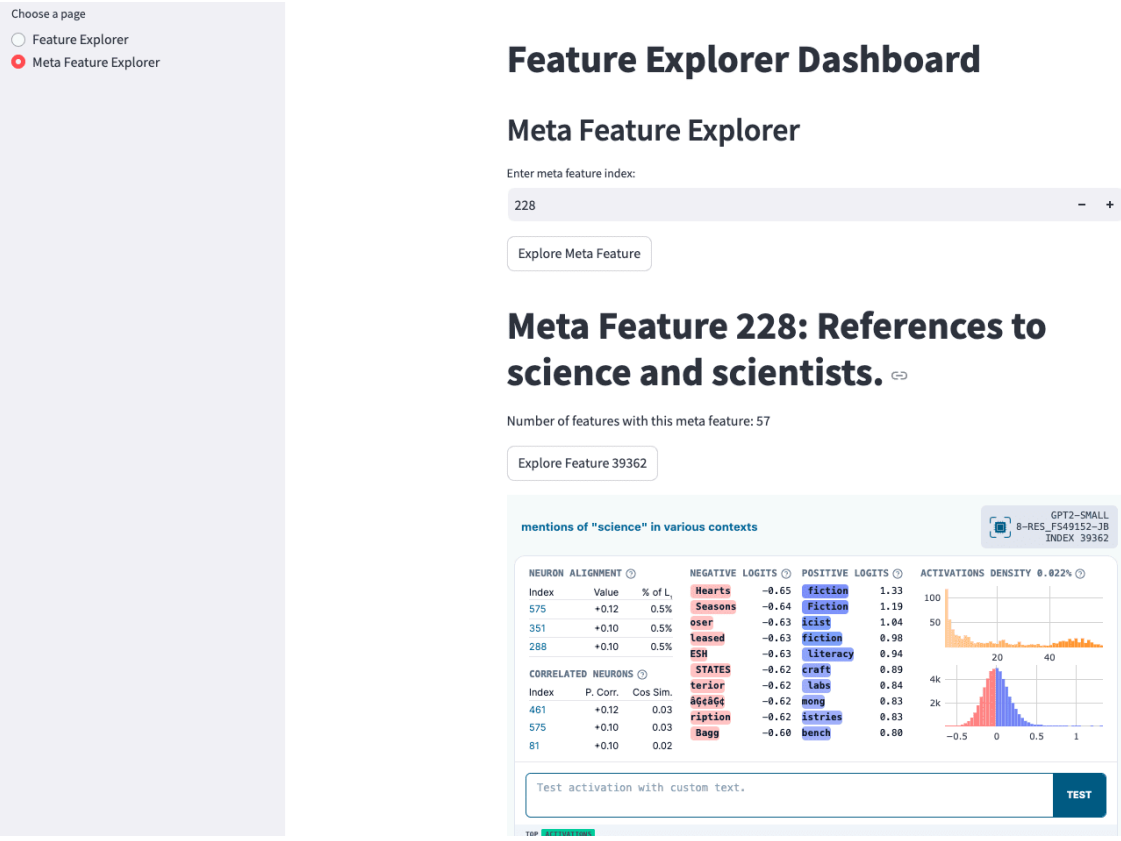
For the meta-SAE, we use the BatchTopK° activation function (k=4), as it generally reconstructs better than standard ReLU and TopK architectures and provides the benefit of allowing a flexible amount of meta-latents per SAE latent. The meta-SAE has a dictionary size of 2304 and is trained for 2000 batches of size 16384 (more than 660

epochs due to the tiny data set). These hyperparameters were selected based on a combination of reconstruction performance and interpretability after some limited experimentation, but (as with standard SAEs) hyperparameter selection and evaluation remain open problems.

The weights for the meta-SAE are available here, and the weights for the SAE are available here.

Meta-latents form interpretable decompositions of SAE latents

We can often make sense of the meta-latents from the set of SAE latents they activate on, which can conveniently be explored in the meta-SAE Explorer we built. Many meta-latents appear interpretable and monosemantic, representing concepts contained within the focal SAE latent.



Click the links into our meta-SAE feature explorer! Thanks to Neuronpedia for the amazing SAE feature integration.

For example, a latent that activates on references to Einstein has meta-latents for science, famous people, cosmic concepts, Germany, references to electricity and energy,

and words starting with a capital E — all relevant to Einstein. The physics terms however are more focused on electricity and energy, as opposed to Einstein's main research areas —relativity and quantum mechanics—which are rarer concepts.

By exploring the five SAE latents that most strongly activate each meta-latent, we can build a graph of SAE latents that have something in common with Einstein, clustered by what they have in common:

Graph of a subset of latents that share meta-latents with the Einstein latent. Thicker lines indicate stronger connections (i.e. higher meta-latent activation).

Here are some other interesting (slightly cherry-picked) SAE latents and their decomposition into meta-latents:

- SAE Latent 38079: References to rugby and rugby-related topics
  - Meta-Latent 2150: References to sports activities
  - Meta-Latent 1982: Words starting with "R"
  - Meta-Latent 1142: References to Ireland
  - Meta-Latent 1067: References to sports leagues
  - Meta-Latent 1024: Terms related to activities or processes

- SAE Latent 5315: Phrases related to democratic principles and social equality

    - Meta-Latent 1974: Conjunctions of phrases related to emotions
    - Meta-Latent 2038: Cultural identity and politics
    - Meta-Latent 1840: Themes of personal development
    - Meta-Latent 1803: Regulatory and policy related themes

- SAE Latent 18157: References to the Android operating system
    - Meta-Latent 625: References to mobile phones
    - Meta-Latent 2020: Mentions of operating systems
    - Meta-Latent 985: References of California cities

Not all meta-latents are easily interpretable however. For example, the most frequent meta-latent activates on 2% of SAE latents but doesn't appear to have a clear meaning. It might represent an average direction for parts not well-explained by the other meta-latents.

Are Meta-Latents different from SAE Latents?

Naively, both meta-latents and SAE latents are trying to learn interpretable properties of the input, so we may not expect much of a difference between which features are represented. For example the meta-latents into which we decompose Einstein, such as Germany and Physics, relate to features we would anticipate being important for an SAE to learn.

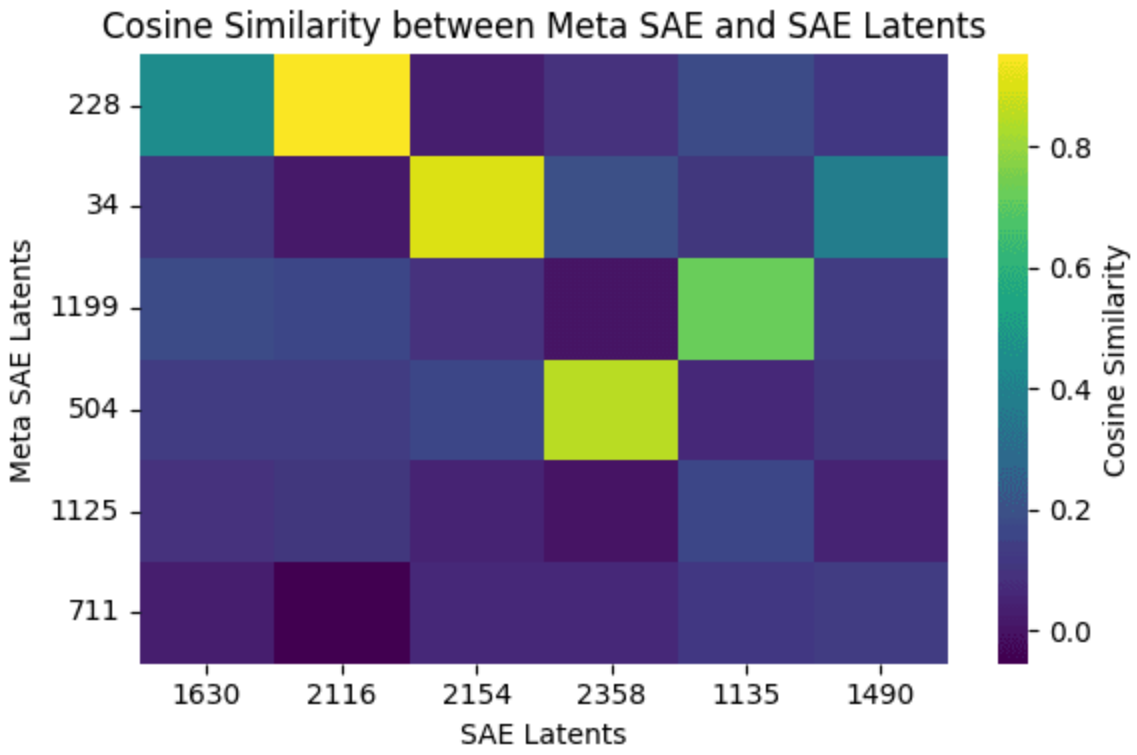
The table below shows the 5 meta-latents we find for Einstein, compared with the 5 latents in SAE-49152 and SAE-3072 with the highest cosine similarity to the Einstein latent (excluding the Einstein latent itself). All of the columns largely consist of latents that are clearly related to Einstein. However, the SAE-49152 latents are sometimes overly specific, for example one latent activates on references to Edison. Edison clearly has many things in common with Einstein, but is of the same class of things as Einstein, rather than potentially being a property of Einstein.

The latents from SAE-3072 give a similar decomposition as the meta-latents, often finding similar concepts relevant to Einstein, such as physics, scientist, famous, German, and astronomical. Compared to the meta-latents, however, the latents may be more specific. For example, the SAE latent° for astronomical terms activates mostly on tokens for the Moon, Earth, planets, and asteroids. The similar meta-latent activates across a

variety of space-related features including those for galaxies, planets, black holes, star wars, astronauts, etc.

Decomposition of Latent 11329: references to Einstein					
Meta-latents & Acts.		**Top latents by cosine similarity to Einstein latent, SAE-49152**		**Top latents by cosine similarity to Einstein late SAE-3072**	
Meta-latent 228: references to science and scientists	0.31	SAE-latent 43445: mentions of “physics”	0.50	SAE-latent 1630°: references to economics, math, or physics	0
Meta-latent 34: prominent figures	0.30	SAE-latent 23058: famous scientists and philosophers (Hegel, Newton, etc)	0.49	SAE-latent 2116°: references to science and scientists	0
Meta-latent 1199: cosmic and astronomical terms	0.25	SAE-latent 39865: mentions of “astronomer”	0.47	SAE-latent 2154°: prominent figures	0
Meta-latent 504: German names, locations, and words	0.21	SAE-latent 6230: references to Edison	0.47	SAE-latent 2358°: mentions of Germany or Germans	0
Meta-latent 1125: terms related to electricity and energy	0.20	SAE-latent 37285: famous writers and philosophers (Melville, Vonnegut, Locke)	0.45	SAE-latent 1135°: astronomical terms, esp. about the Moon, asteroids, spacecraft	0
Meta-latent 711: words starting with a capital E.	0.19	SAE-latent 6230: mentions of “scientist”	0.43	SAE-latent 1490°: mentions of Wikileaks, Airbnb and other orgs.	0

The cosine similarity between each of the meta-latent decoder directions (y-axis) and the SAE latent decoder directions is plotted in the heatmap below.



We see a similar pattern when comparing meta-latents and SAE-3072 latents for randomly selected SAE-49152 latents, with both sets giving reasonable decompositions.

Decomposition of Latent 42: phrases with “parted” such as “parted ways” or “parted company”			
Meta-latents & Acts.		**Top latents by cosine similarity to “parted” latent, SAE-3072 [list°]**	
Meta-latent 266: words related to ending or stopping	0.35	SAE latent 1743°: mentions of “broke” or “break”	0
Meta-latent 409: adverbial phrases like “square off”, “ramp down”, “iron out”	0.27	SAE latent 392°: terms related to detaching or extracting	0
Meta-latent 1853: words related to part, portion, piece.	0.27	SAE latent 1689°: terms related to fleeing or escaping	0
Meta-latent 1858: words related to crossings or boundaries, probably related to predicting words related to “way” like “path” or “road”	0.23	SAE latent 1183°: mentions of “cross”, high positive logits for “roads”	0

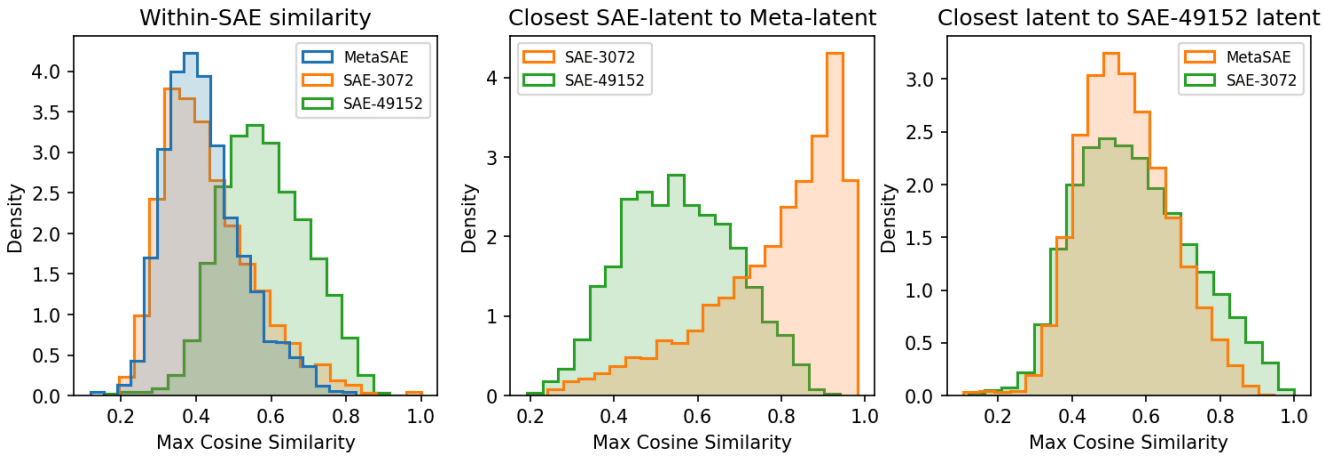
Meta-latent 2004: terms related to collaborations	0.18	SAE latent 2821°: verbs that can be followed by “up” or “down”, also positive logits for “oneself” [unclear]	0

Decomposition of Latent 0: descriptions of the form “sheet of ____”, “wall of ____”, “lin of ____”, etc.

Meta-latents & Acts.		**Top latents by cosine similarity to “sheet latent, SAE-3072 [list°]**	
Meta-latent 161: “of” phrases specifically for collections, such as “team of volunteers” and “list of places”.	0.35	SAE latent 1206°: “of” phrases for collections, such as “network of” or “group of”	0
Meta-latent 1999: physical attributes and descriptions	0.27	SAE latent 1571°: descriptions of being immersed, such as “drenched in ____” or “dripping with ____”	0
Meta-latent 732: “of” phrases for quantities	0.27	SAE latent 2100°: “[noun] of” phrases	0
Meta-latent 1355: phrases with “of” and “with”	0.23	SAE latent 2461°: prominent “[noun] of” organizations, such as “Board of Education” and “House of Commons”	0
Meta-latent 926: User prompts with “to”	0.18	SAE latent 2760°: “[number] of” phrases	0

We can compare the similarities within and between SAEs, in particular focusing on the similarity of the closest feature for a given one. The first plot below shows meta-SAEs generally have meta-features that are more orthogonal to each other than the SAE they are trained on (SAE-49152). However, this difference is explained by the size of the

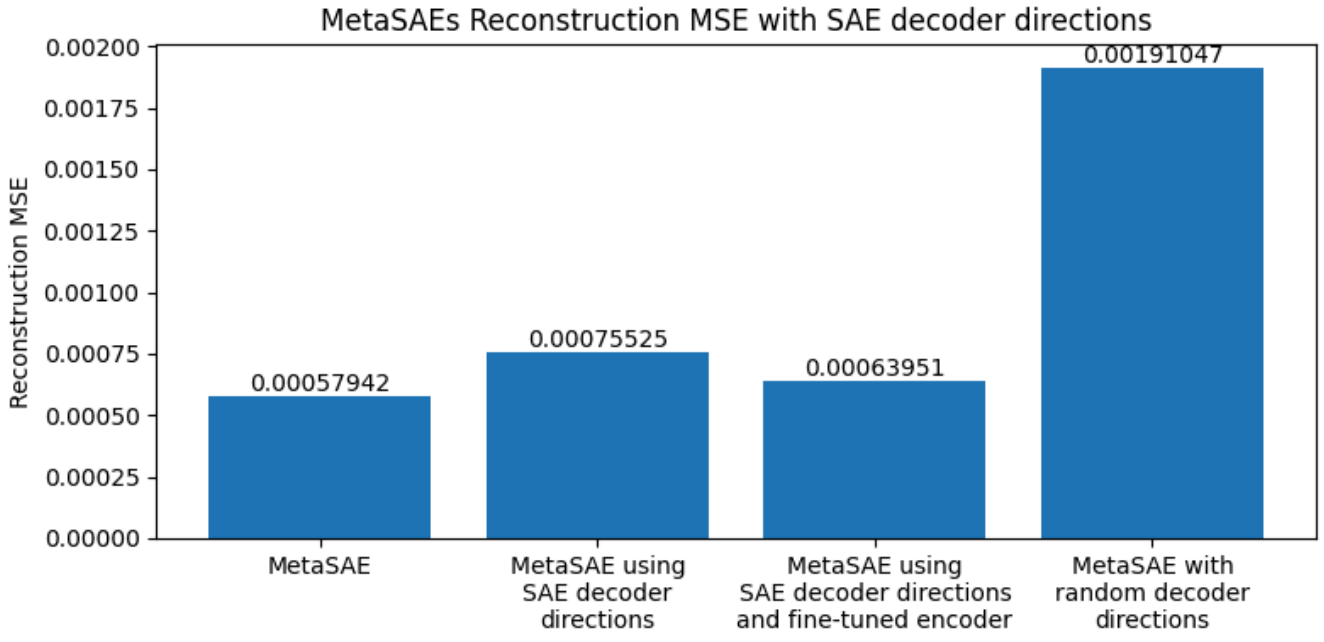
meta-SAE since SAE-3072 has a similar distribution of max cosine similarities. In the second and the third plot, we find that for many meta-latents there is not a very similar (cosine similarity > 0.7) latent in SAE-49152, but there often is one in SAE-3072.



Distributions of maximum cosine similarities between SAE latents and meta-SAE latents

We evaluated the similarity of meta-SAE latents with SAE latents by comparing the reconstruction performance of the meta-SAE with variants of the meta-SAE.

- The first variant replaces the decoder weights of the meta-SAE latents with the decoder weights of the SAE latent with maximum cosine similarity from 4 SAEs from SAE-768 to SAE-6144.
- The second variant uses the same decoder as the first variant, but fine-tunes the encoder for 5,000 epochs.
- The last variant uses random directions in the decoder as a baseline.



We see that the reconstruction performance when using the SAE decoder directions and fine-tuning the encoder is only slightly worse than the performance using original meta-SAE.

Although the SAE-3072 finds similar latents, we find that the meta-SAE reveals different relationships between the SAE-49k latents compared to the relationships induced by cosine similarities with the SAE-3072. To demonstrate this, we construct two adjacency matrices for SAE-49k: one based on shared meta-latents, and another using cosine similarity with latents of SAE-3072 as follows:

- In the meta-latent adjacency graph, each SAE-49k latent is connected to another if they share a meta-feature.
- In the cosine similarity adjacent graph, each SAE-49k latent is connected to another if they both have cosine similarity greater than a threshold to the same SAE-3k latent. The threshold is set such that both graphs contain the same number of edges.

We determine a cosine similarity threshold to ensure the same number of edges in both graphs. Looking at the ratio of shared edges with the size of the total amount of edges in each graph, we find that 28.92% of the vertices were shared. This indicates that while there is some similarity, the meta-SAE approach uncovers substantially different latent relationships than those found through cosine similarity with a smaller SAE.

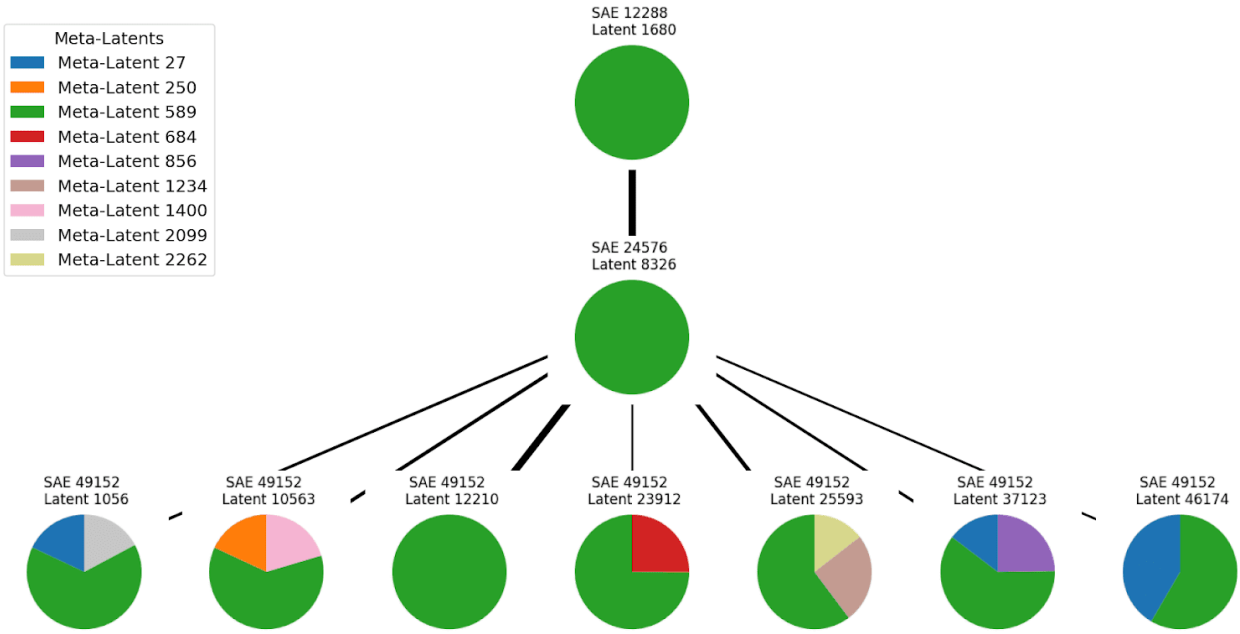
Together, these results suggest that while there are similarities between meta-latents and latents from smaller SAEs, there are also differences in the relationships they capture. We currently don't have a complete understanding of how the original data distribution, latents, and meta-latents relate to each other. Training a meta-SAE likely captures some patterns similar to those found by training a smaller SAE on the original data. However, the hierarchical approach of meta-SAEs may also introduce meaningful distinctions, whose implications for interpretability require further investigation.

It's plausible that we can get similar results to meta-SAEs by decomposing larger SAE latents into smaller SAE latents using sparse linear regression or inference time optimization (ITO)[°]. Meta-SAEs are cheaper to train on large SAEs, especially when many levels of granularity are desired, but researchers may also have a range of SAE sizes already available that can be used instead. We think then that meta-SAEs are a valid approach for direct evaluation of SAE latent atomicity, but may not be required in practice where smaller SAEs are available.

Using Meta-SAEs to Interpret Split Features

In Towards Monosemanticity, the authors observed the phenomenon of feature splitting, where latents in smaller SAEs split into multiple related latents in larger SAEs. We recently showed° that across different sizes of SAEs, some latents represent the same information as latents in smaller SAEs but in a sparsified way. Potentially we can understand better what is happening here by looking at the meta-latents of the latents at different levels of granularity.

In order to do so, we trained a single meta-SAE on the decoders of 8 SAEs with different dictionary sizes (ranging from 768 to 98304) trained on layer 8 of the residual of gpt2-small. Then we identify split latents based on cosine similarity > 0.7.



The ratio of meta-latent activations in SAE latents in SAEs of different sizes

If we take a look at this example of a latent splitting into 7 different latents. Latent 8326° in SAE size 24576 activates on the word "talk". It only has one meta-latent active, namely meta-latent 589, which activates on features related to talking/speaking. But then, in SAE size 49152, it splits in the 7 different latents, with the following meta-latents[1]:

- Latent 37123°: "talk" as a noun
  - Meta-latent 589: talking/speaking
  - Meta-latent 856: verbs used as nouns

    - Meta-latent 27: “conversation”/”discussion”/”interview”

- Latent 23912°: “Talk” or “talk” as the first word of a sentence
    - Meta-latent 589: talking/speaking
    - Meta-latent 684: verbs in imperative (start of sentence)

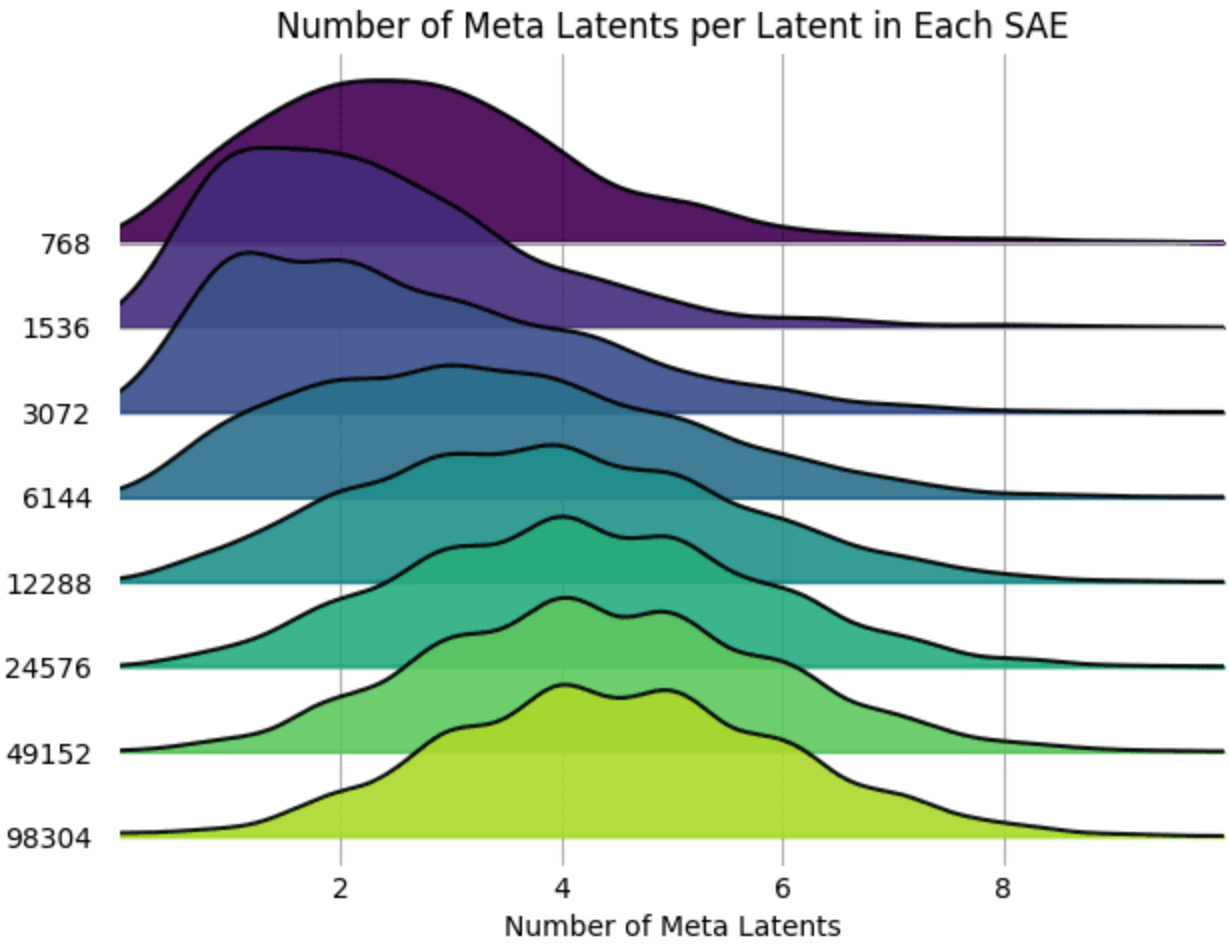
- Latent 10563°: “talk” in the future (e.g. “let’s talk about”)
    - Meta-latent 589: talking/speaking
    - Meta-latent 250: “let’s” / “I want to”
    - Meta-latent 1400: regarding a topic

- Latent 12210°: “talk”
    - Meta-latent 589: talking/speaking

- Latent 25593°: “talking” as participial adjective (talking point/talking head)
    - Meta-latent 589: talking/speaking
    - Meta-latent 1234: verbs as as participial adjective
    - Meta-latent 2262: profane/explicit language

- Latent 46174°: “talking/chatting”
    - Meta-latent 589: talking/speaking
    - Meta-latent 27: “conversation”/”discussion”/”interview”

We observe that a relatively broad latent, with only one meta-latent, splits into seven more specific latents by adding on meta-latents. We suspect that this is largely due to the fact that the latents close to the latents in the smaller SAEs appear more often in the meta-SAE training because they are present across multiple SAE sizes. Therefore, it makes sense for this meta-SAE to assign a meta-latent for these latents. Indeed, we observe that latents in the smaller SAEs activate less meta-latents on average.



Distribution of meta-latents active in latens from SAEs of different sizes

Causally Intervening and Making Targeted Edits with Meta-Latents

The Ravel benchmark introduces a steering problem for interpretability tools that includes the problem of changing attributes of cities, such as their country, continent, and language. We use this example in an informal case-study into whether meta-SAEs are useful for steering model behavior.

We evaluate whether we can steer on SAE latents and meta-SAE latents to steer GPT-2's factual knowledge about cities in the categories of country, language, continent, and currency. To simplify the experiments, we chose major cities where both their name, and all of these attributes are single tokens; as such we test on Paris, Tokyo, and Cairo. GPT2-small is poor at generating text containing these city attributes, so instead we use the logprobs directly. We do not evaluate the impact on model performance in general, only on the logits of the attributes.

First, we evaluate the unmodified model, which assigns the highest logits to the correct answers.

Prompt	Answer	Logprobs
Tokyo is a city in the country of Japan	Japan	-11.147
	France	-16.933
	Egypt	-21.714
The primary language spoken in Tokyo is Japanese	Japanese	-9.563
	French	-16.401
	Arabic	-16.496
Tokyo is a city on the continent of Asia	Asia	-12.407
	Europe	-14.848
	Africa	-17.022
The currency used in Tokyo is the Yen	Yen	-11.197
	Euro	-14.965
	Pound	-15.475

We then steer (Turner Templeton Nanda°) using the SAE latents. In the GPT2-49152 SAE, both Tokyo° and Paris° are represented as individual latents, which means that steering on these latents essentially corresponds to fully replacing the Tokyo-latent with the Paris-latent, i.e. we clamp the Tokyo-latent to zero and clamp the Paris-latent to its activation on the Paris related inputs at all token positions. We see that steering on these latents results in all the attributes of Tokyo being replaced with those of Paris.

Prompt	Answer	Logprobs
Tokyo is a city in the country of France	Japan	-19.304
	France	-9.986
	Egypt	-17.692

The primary language spoken in Tokyo is French	Japanese	-14.137
	French	-9.859
	Arabic	-13.126
Tokyo is a city on the continent of Europe	Asia	-14.607
	Europe	-13.308
	Africa	-13.959
The currency used in Tokyo is the Euro	Yen	-13.241
	Euro	-12.354
	Pound	-13.002

We can decompose these city SAE latents into meta latents using our meta-SAE. The Tokyo latent decomposes into 4 latents, whereas Paris decomposes into 5. These allow us more fine-grained control of city attributes than we could manage with SAE latents. The city meta-latents with a short description are provided below (human labeled rather than auto-interp as in the dashboard).

Meta Latent	In Paris	In Tokyo	Description
281	✅	✅	City references
389	❌	✅	Names starting with T
572	❌	✅	References to Japan
756	❌	✅	-i*o suffixes
1512	✅	❌	-us substrings
1728	✅	❌	French city names and regions
1809	✅	❌	Features of Romantic language words e.g. accents
1927	✅	❌	Capital cities

In order to steer the SAE latents using meta-latents, we would like to remove some Tokyo attributes from Tokyo and add in some Paris attributes instead. To do this, we

reconstruct the SAE latent whilst clamping the activations of the unique meta-latents of Tokyo to zero, and the activations of the unique meta-latents of Paris to their activation value on Paris, and then normalizing the reconstruction.

For example, one combination of Tokyo and Paris meta-latents results in changing the language of Tokyo to French and the currency to the Euro, whilst not affecting the geographic attributes (though in some cases the result is marginal). The meta-latents removed were 281, 389, 572; and meta-latent 1809 was added.

Prompt	**Answer**	**Logprobs**
Tokyo is a city in the country of Japan	Japan	-12.287
	France	-13.274
	Egypt	-19.015
The primary language spoken in Tokyo is French	Japanese	-12.210
	French	-10.870
	Arabic	-13.947
Tokyo is a city on the continent of Asia	Asia	-13.135
	Europe	-14.380
	Africa	-15.983
The currency used in Tokyo is the Euro	Yen	-13.270
	Euro	-11.496
	Pound	-13.863

Not all combinations of attributes can be edited, however. The table below displays the combinations of attributes that we managed to edit from Tokyo to Paris. These were found by enumerating all combinations of meta-latents for both cities.

Country	**Language**	**Continent**	**Currency**	**Start city meta-latents removed**	**Target city meta-latents added**
❌	❌	❌	❌		

❌	❌	❌	✅		1512, 1728, 1809 1927
❌	✅	❌	✅	281, 389, 572	1809
✅	✅	❌	✅	281, 389, 572	1512
✅	✅	✅	✅	281, 389, 572	1728

The meta-latents used to steer some of these combinations makes sense:

- To steer Tokyo to speak French, a meta latent (1809) that composes SAE latents for European languages is used.
- To steer Tokyo into Europe, a meta latent (1728) that composes SAE latents for European cities and countries is used.

However, a latent for '-us' suffixes can be used to steer Tokyo into France. Paris is the only major capital city with the ending '-us' (there's also Vilnius and Damascus), but this explanation feels unsatisfactory, particularly given that 281, which is the shared major city latent between Paris and Tokyo, is not present in the reconstruction.

Discussion

In this post we introduced meta-SAEs as a method for decomposing SAE latents into a set of new monosemantic latents, and now discuss the significance of these results to the SAE paradigm.

Our results suggest that SAEs may not find the basic units of LLM computation, and instead find common compositions of those units. For example, an Einstein latent is defined by a combination of German, physicist, celebrity, etc. that happens to co-occur commonly in the dataset, as well as the presence of the Einstein token. The SAE latents do not provide the axes of the compositional representation space in which these latents exist. Identifying this space seems vital to compactly describing the structure of the model's computation, a major goal of mechanistic interpretability, rather than describing the dataset in terms of latent cooccurrence.

Meta-SAEs appear to provide some insight into finding these axes of the compositional space of SAE latents. However, just as SAEs learn common compositions of dataset

features, meta-SAEs may learn common compositions of these compositions; certainly in the limit, a meta-SAE with the same dictionary size as an SAE will learn the same latents. Therefore there are no guarantees that meta-latents reflect the true axes of the underlying representational space used by the network. In particular, we note that meta-SAEs are lossy - Einstein is greater than the sum of his meta-latents, and this residual may represent something unique about Einstein, or it may also be decomposable.

Our results highlight the importance of distinguishing between 'monosemanticity' and 'semantic atomicity'. We think that meta-SAEs plausibly learn latents that are 'more atomic' than those learned by SAEs, but this does not imply that we think they are 'maximally atomic'. Nor does it imply that we think that we can find more and more atomic latents by training towers of ever-more-'meta' meta-SAEs. We're reminded of two models of conceptual structure in cognitive philosophy (Laurence and Margolis, 1999). In particular, the 'Containment model' of conceptual structure holds that "*one concept is a structured complex of other concepts just in case*[2] *it literally has those other concepts as proper parts*". We sympathize with the Containment model of conceptual structure when we say that "some latents can be more atomic than others". By contrast, the 'Inferential model' of conceptual structure holds that "*one concept is a structured complex of other concepts just in case it stands in a privileged relation to these other concepts, generally, by way of an inferential disposition*". If the Inferential model of conceptual structure is a valid lens for understanding SAE and meta-SAE latents, it might imply we should think about them as nodes in a cyclic graph of relations to other latents, rather than as a strict conceptual hierarchy. These are early thoughts, and we do not take a particular stance with regard to which model of conceptual structure is most applicable in our context. However, we agree with Smith (2024)° that we as a field will make faster progress if we "think more carefully about the assumptions behind our framings of interpretability work". Previous work in cognitive- and neuro-philosophy will likely be useful tool sets for unearthing these latent assumptions.

More practically, meta-SAEs provide us with new tools for exploring feature splitting in different sized SAEs, allowing us to enumerate SAE latents by the meta-latents of which they are composed. We also find that meta-SAEs offer greater flexibility than SAEs on a specific city attribute steering task.

There are a number of limitations to the research in this post:

- This research was conducted on a single meta-SAE, trained on a single SAE, at a single point of a single small model. We have begun exploratory research on the

Gemma Scope SAEs, early results are encouraging but significantly more research is required.

- Currently, we do not have a good grasp of whether the meta-latents learned are substantially different from the latents learned in smaller SAEs. While our initial results suggest some similarities and differences, more rigorous evaluation is needed. For example, Scaling and Evaluating SAEs evaluates their SAEs using a probing dataset metric and a feature explanation task, which we want to apply to meta-SAEs.
- Our steering experiments used a simplified version of the Ravel benchmark, and we tested only a small number of city pairs. Further validation of this approach is required, as well as refining the approach taken for identifying the steering meta-latents.

---

1. ^ We use Neuronpedia Quick Lists rather than the dashboard for these latents, as these experiments use a different meta-SAE (trained on 8 different SAEs sizes rather than just the 49k).
2. ^ Note that in philosophical texts, “just in case” means “if and only if”/”precisely when”

Mentioned in

117 Research directions Open Phil wants to fund in technical AI safety
98 Matryoshka Sparse Autoencoders
82 SAEBench: A Comprehensive Benchmark for Sparse Autoencoders
73 [Paper] A is for Absorption: Studying Feature Splitting and Absorption in Sparse Autoencoders
49 Toy Models of Feature Absorption in SAEs

Load More (5/9)

10 comments, sorted by top scoring

[-] **Charlie Steiner** 1y Ω 6 ▼ 9 ▲ ✕ 1 ✓

The fact that latents are often related to their neighbors definitely seems to support your thesis, but it's not clear to me that you couldn't train a smaller, somewhat-lossy meta-SAE even on an idealized SAE, so long as the data distribution had rare events or rare properties you could thow away cheaply.

You could also play a similar game showing that latents in a larger SAE are "merely" compositions of latents in a smaller SAE.

So basically, I was left wanting a more mathematical perspective of what kinds of properties you're hoping for SAEs (or meta-SAEs) and their latents to have.

It would be interesting to meditate in the question "What kind of training procedure could you use to get a meta-SAE directly?" And I think answering this relies in part on mathematical specification of what you want.

• • •

When you showed the decomposition of 'einstein', I also kinda wanted to see what the closest latents were in the object-level SAE to the components of 'einstein' in the meta-SAE.

[-] **Lee Sharkey** 1y Ω 8 ▼ 11 ▲ ✕ 0 ✓

> It would be interesting to meditate in the question "What kind of training procedure could you use to get a meta-SAE directly?" And I think answering this relies in part on mathematical specification of what you want.

At Apollo we're currently working on something that we think will achieve this. Hopefully will have an idea and a few early results (toy models only) to share soon.

4

[-] **Neel Nanda** 1y Ω 2 ▼ 2 ▲ ✕ 0 ✓

> but it's not clear to me that you couldn't train a smaller, somewhat-lossy meta-SAE even on an idealized SAE, so long as the data distribution had rare events or rare properties you could thow away cheaply.

IMO am "idealized" SAE just has no structure relating features, so nothing for a meta SAE to find. I'm not sure this is possible or desirable, to be clear! But I think that's what idealized units of analysis should look like

> You could also play a similar game showing that latents in a larger SAE are "merely" compositions of latents in a smaller SAE.

I agree, we do this briefly later in the post, I believe. I see our contribution more as showing that this kind of thing is possible, than that meta SAEs are objectively the best tool for it

[-] **Rohin Shah** 11mo Ω 5 ▼ 7 ▲ ✕ 0 ✓

Suppose you trained a regular SAE in the normal way with a dictionary size of 2304. Do you expect the latents to be systematically different from the ones in your meta-SAE?

For example, here's one systematic difference. The regular SAE is optimized to reconstruct activations uniformly sampled from your token dataset. The meta-SAE is optimized to reconstruct decoder vectors, which in turn were optimized to reconstruct activations from the token dataset -- however, different decoder vectors have different frequencies of firing in the token dataset, so uniform over decoder vectors

!= uniform over token dataset. This means that, relative to the regular SAE, the meta-SAE will tend to have less precise / granular latents for concepts that occur frequently in the token dataset, and more precise / granular latents for concepts that occur rarely in the token dataset (but are frequent enough that they are represented in the set of decoder vectors).

It's not totally clear which of these is "better" or more "fundamental", though if you're trying to optimize reconstructed loss, you should expect the regular SAE to do better based on this systematic difference.

(You could of course change the training for the meta-SAE to decrease this systematic difference, e.g. by sampling from the decoder vectors in proportion to their average magnitude over the token dataset, instead of sampling uniformly.)

[-] **Neel Nanda** 11mo Ω 5 ▼ 7 ▲ ✕ 0 ✓

Interesting thought! I expect there's systematic differences, though it's not quite obvious how. Your example seems pretty plausible to me. Meta SAEs are also more incentived to learn features which tend to split a lot, I think, as then they're useful for more predicting many latents. Though ones that don't split may be useful as they entirely explain a latent that's otherwise hard to explain.

Anyway, we haven't checked yet, but I expect many of the results in this post would look similar for eg sparse linear regression over a smaller SAEs decoder. Re why meta SAEs are interesting at all, they're much cheaper to train than a smaller SAE, and BatchTopK gives you more control over the L0 than you could easily get with sparse linear regression, which are some mild advantages, but you may have a small SAE lying around anyway. I see the interesting point of this post more as "SAE latents are not atomic, as shown by one method, but probably other methods would work well too"

[-] **Alexander Gietelink Oldenziel** 1y ▼ 2 ▲ ✕ 0 ✓

I'm curious if these observations are related at all to the work by Mendel, Hanni and Vaintrob on SAE features°, more discussion here°.

[-] **Neel Nanda** 1y ▼ 2 ▲ ✕ 0 ✓

Is the first post the one you meant to link, or did you mean the followup post from Jake? The first post is on toy models of AND and XORs, which I don't see as being super relevant. But I think Jake's argument that there's clear structure that naive hypotheses neglect seems clearly legit

[-] **Seonglae Cho** 6mo ▼ 1 ▲ ✕ 0 ✓

What happens when we learn Meta-SAE's decoder weights again? Meta-Meta-SAE? 🤔

I can only expect greater lossy decomposition

[-] **Ian Johnson** 1y Ω 0 1 0

Are the datasets used to train the meta-SAEs the same as the datasets to train the original SAEs? If atomicity in a subdomain were a goal, would training a meta-SAE with a domain-specific dataset be interesting?

It seems like being able to target how atomic certain kinds of features are would be useful. Especially if you are focused on something, like identifying functionality/structure rather than knowledge. A specific example would be training on a large code dataset along with code QA. Would we find more atomic "bug features" like in scaling monosemanticity?

[-] **Neel Nanda** 1y 3 0

The dataset for a meta SAE is literally the decoder directions of the original SAE, there are no tokens involved

Moderation Log

More from Bart Bussmann

42	Learning Multi-Level Features with Matry…	Bart Bussmann, Patrick Leask, Ne…	8mo	6
61	BatchTopK: A Simple Improvement for To…	Bart Bussmann, Patrick Leask, Ne…	1y	0
53	Calendar feature geometry in GPT-2 layer …	Patrick Leask, Bart Bussmann, Ne…	1y	0

View more

Curated and popular this week

230	The Problem ★	Rob Bensinger, tanagrabeast, yams, So8res, Eliezer …	21h	97
178	Emotions Make Sense ★	DaystarEld	5d	31

192 How anticipatory cover-ups go wrong Kaj_Sotala 3d 12