# Visual Exploration of Feature Relationships in Sparse Autoencoders with Curated Concepts

**Anonymous Author(s)**
Affiliation
Address
email

## Abstract

Sparse autoencoders (SAEs) have emerged as a powerful tool for uncovering interpretable features in large language models (LLMs) through the sparse directions they learn. However, the sheer number of extracted directions makes comprehensive exploration intractable. While conventional embedding techniques such as UMAP can reveal global structure, they suffer from limitations including high-dimensional compression artifacts, overplotting, and misleading neighborhood distortions. In this work, we propose a focused exploration framework that prioritizes curated concepts and their corresponding SAE features over attempts to visualize all available features simultaneously. We present an interactive visualization system that combines topology-based visual encoding with dimensionality reduction to faithfully represent both local and global relationships among selected features. This hybrid approach enables users to investigate SAE behavior through targeted, interpretable subsets, facilitating deeper and more nuanced analysis of concept representation in latent space.

## 1 Introduction

Sparse autoencoders (SAEs) have emerged as a powerful technique for extracting interpretable features from large language models, decomposing superposed neural representations into disentangled components [1, 2, 3, 4, 5, 6]. Recent work has demonstrated remarkable scalability, extracting millions of interpretable features from state-of-the-art models [7, 8, 9, 10, 11, 12]. However, this success creates a paradox: the sheer number of learned sparse directions, often hundreds of thousands to millions, makes comprehensive exploration computationally and cognitively intractable. Moreover, recent studies [13, 14] also reveal that many learned SAE features are *polysemantic* (encoding multiple, unrelated concepts) or low-quality, raising questions about whether visualizing all features simultaneously is even desirable.

In this work, instead of attempting to visualize all the features in SAE [15, 16, 17] all at once, we advocate for a focused paradigm of carefully curated concept sets and only their corresponding SAE features. By concentrating on well-defined concepts, researchers can avoid confusion from the overwhelming number and potentially low-quality features and conduct targeted investigations of specific hypotheses about concept representation and feature relationships.

Specifically, we address three interconnected analytical objectives that are central to understanding the relationships between SAE features and revealing the conceptual structure. First, we investigate how well cosine similarity among features corresponds to semantic similarity for validating whether SAEs capture meaningful semantic structure [18]. Second, we examine how well human-curated concept hierarchies are reflected in the SAE feature space, assessing whether models learn organizational structures aligned with human understanding [19, 12]. Third, we analyze how these local and global

Submitted to 39th Conference on Neural Information Processing Systems (NeurIPS 2025). Do not distribute.

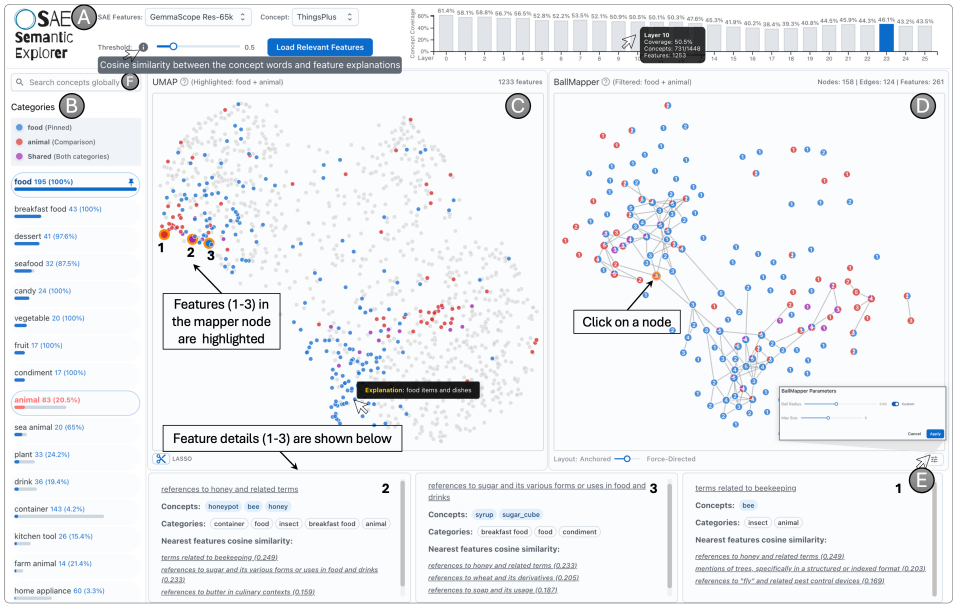


Figure 1: SAE-Explorer interface. **A. Data view.** Left: SAE features, a concept set (words with assigned categories), and a cosine similarity threshold for loading relevant features. Right: bar chart of discovered concepts per layer. **B. Category view.** For the selected layer (23), each row shows a category's feature count and its overlap with the pinned category *food*, enabling comparison with *animal*. **C. UMAP view.** Retrieved features for the selected layer, with *food* and *animal* highlighted. **D. Ball mapper view.** Topological structure of *food* and *animal* features represented as a graph. **E. Feature view.** Details of selected features via click or lasso. **F. Concept query.** Search interface for specific concepts.

patterns evolve across model layers, as feature representations undergo systematic changes through network depth [9, 10].

Although existing visualization of SAEs uses techniques such as UMAP [20], since they perform extreme compression from thousands of dimensions into two, they can cause severe structural distortions. These distortions often misrepresent neighborhood relationships [21], which are critical for interpretability and understanding semantic similarities between features. For example, similar features may appear separated, or different features may appear artificially clustered.

To address this critical limitation of neighborhood misrepresentation in traditional embeddings, we introduce a novel topology-based visual encoding inspired by topological data analysis. Building on the Ball Mapper algorithm [22], our approach provides a network representation that better preserves both local and global structural properties. Unlike projection methods that force features into continuous 2D space, our topological representation maintains discrete feature clusters while accurately representing their interconnections, enabling more faithful communication of similarity relationships among SAE features. To support flexible exploration and the interactive nature of exploratory data analysis, we present a human-centered visual analytics system, SAE-Explorer, which facilitates qualitative analysis of feature relationships that are local, global, or across layers via multiple coordinated views. The system code, datasets, and video demonstration are available at https://anonymous.4open.science/r/SAEExploration-A57B.

## 2 SAE-Explorer: A Visualization Tool for Exploring Curated Concepts

Driven by analytical goals, as shown in Figure 1, we design a visual analytics system for inspecting SAE features that enable human-in-the-loop exploration.

**Data description and preprocessing.** The input data includes SAE features (high-dimensional vectors) across layers, LLM auto-interpreted [23] textual explanations per feature, and a concept dataset with concept words and their categories (e.g., the category *animal* contains *dog*, *cat*, etc., and a concept may belong to multiple categories). For each layer, we retrieve relevant features for each concept via the Neuropedia API [24] that relies on the cosine similarity between concept and

feature explanations in a sentence embedding space. By default, features with similarity above $0.5$ are visualized, though users can adjust this threshold interactively.

**Embedding View Based on Ball Mapper.** To better capture local structure and mitigate misleading distortions in traditional dimensionality reduction, we introduce an embedding view based on ball mapper. Ball mapper [22] encodes the topology of a high-dimensional point cloud as a graph, where nodes are local neighborhoods (balls) of points and edges indicate overlaps between them. Given a user-defined radius $\epsilon$, the method selects a subset of points (e.g., via a greedy procedure) such that the entire dataset is covered by balls centered at these points. Each ball defines a node, and two nodes are connected if their balls share points. By default, $\epsilon$ is determined by computing all pairwise cosine distances and selecting the elbow point. To reduce visual clutter, we employ an adaptive variant: for a specified maximum node size, the ball radius is iteratively reduced by a factor $\eta$ (default $0.9$) until the constraint is met. In our system, the maximum node size defaults to $5$, though users may adjust this and other parameters.

**Interface Design.** The interface integrates multiple linked views. As shown in Figure 1, view A allows users to select SAE feature data, concept data, and a similarity threshold to retrieve concept-relevant features per layer. The resulting distribution of discovered concepts is displayed in a bar chart. Clicking a bar reveals the corresponding layer data in views B, C, and D.

For a selected layer, view B lists category information, including the number of relevant features associated with each category. View C shows a UMAP projection of the features, supporting zoom, pan, hover, lasso, and selection, with the three nearest features highlighted on click. View D presents the ball mapper view, where users can drag, adjust parameters, and select nodes or edges. The ball mapper graph supports a smooth transition between two layouts: an anchored layout, in which nodes are positioned by the average location of their UMAP points for alignment, and a force-directed layout for a more aesthetically pleasing visualization.

View E displays detailed information about selected features, including explanations, relevant concepts and categories, and nearby features within the same layer. Explanations are linked to the Neuropedia feature detail page [24]. View F provides a search interface for retrieving features associated with specific concepts.

To aid navigation, tooltips are available throughout the interface. Selecting a category in view B highlights its corresponding features in both the UMAP and ball mapper views. To further explore category relationships, users can pin a category, after which other categories are sorted by shared features. When a comparison category is selected, features are consistently color-coded across UMAP and ball mapper to reveal overlaps and differences.

## 3 Results

In this section, we present the insights obtained through the proposed visualization framework in understanding the relationships between the learned SAE features and human-interpretable concepts.

**Concept Sets**. We considered two hierarchical concept sets for our analysis: (1) `THINGSplus` [25], a human-curated set containing $1448$ concepts organized into $53$ categories (e.g., mammals, food, etc), where each category contains a list of relevant concepts (e.g., *panda*, *cat*, *breadsticks*); and (ii) `Subjects`, a field-of-study oriented set curated via an LLM. This set includes a total of $1683$ concepts across eight disciplines (e.g., math and physics) and the relevant concepts within each discipline (e.g., *algebra*, *geometry*, *calculus* for math). While most SAE visualization frameworks [8] attempt to provide the user with a view of the learned features, they often lack the ability to connect these features to specific concepts that a practitioner cares about. Instead, we address this gap by allowing users to load a concept set and retrieve features relevant to the concepts of interest thus enabling a more targeted exploration. To demonstrate the utility, we consider the open-source 65k SAE features for each layer of the `gemma-2-2b` model [26] provided by Gemma Scope [27] from the residual stream. Neuronpedia [24] provides a comprehensive set of features and textual explanations obtained via the Auto-Interp framework [28] for each feature across $26$ layers of the model.

**Evolution of Concept Relationships Across Layers**. Through our visualization framework, we make a number of interesting observations regarding how concept relationships evolve across layers. We observed that for the `ThingsPlus` concept set, the representations in UMAP are all clustered together in the early layers, while they gradually separate into two distinct clusters in the middle

layers and finally merge again in the later layers. This pattern suggests that the model initially learns general features, then refines them into more specific concepts, before finally integrating these concepts into a set of cohesive representations. These results are consistent with the observations made in [29] regarding the evolution of factual knowledge in LLMs. However, we also note that the ball mapper view reveals a more nuanced picture, with multiple smaller clusters emerging in the middle layers, indicating that the model is learning a diverse set of features that may not be fully captured by UMAP's global structure. Moreover, we also observe that for the `Subjects` concept set, this behavior is less pronounced, with the UMAP representations remaining relatively clustered throughout the layers. We hypothesize that this is due to the more abstract nature of the concepts in this set, which may not lend themselves to clear separations in the feature space. In addition, we made another interesting observation regarding the local neighborhoods identified via ball mapper. We found that local semantic proximity often persists across layers, consistent with previous findings [30]. For example, in Figure 2(A, bottom), querying *Fox* consistently returns features associated with *Fox News*, with *Wolf* appearing as the nearest neighbor in the same mapper node.

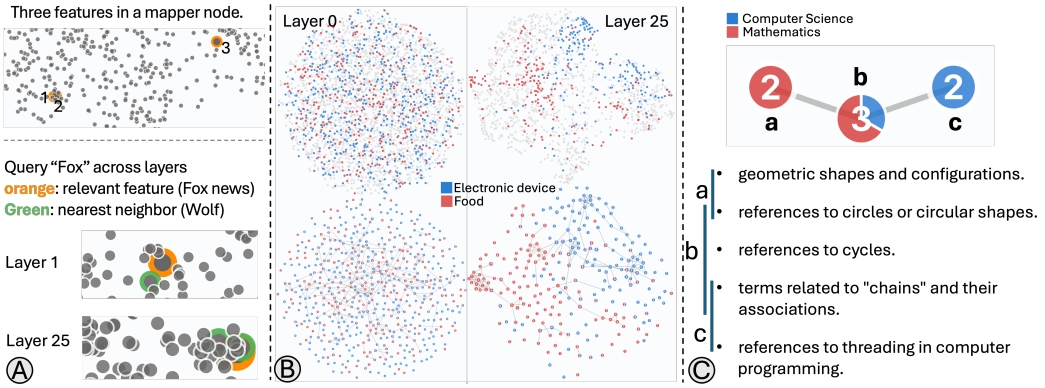


Figure 2: **A.** Top: a mapper node containing features 1–3, all related to a *music album*; bottom: querying *fox* across layers consistently yields *Fox News*, with *wolf* as the nearest neighbor. **B.** UMAP and ball mapper views at layers 0 and 25, highlighting features of *food* and *electronic devices*. **C.** A ball mapper path of subject concepts illustrating a transition from *Mathematics* to *Computer Science*.

**Complementary Local and Global Analysis.** Since our framework integrates both UMAP and ball mapper views, it supports local and global analysis of the feature space: ball mapper complements UMAP and helps reveal distortions introduced by dimensionality reduction. The complementary nature of the two views is illustrated in Figure 2(A, top), where features 1–3 within a ball mapper node all relate to *music album*. While Features 1 and 2 are true nearest neighbors, this proximity is distorted in UMAP. As noted earlier, our visualization also shows that local semantic proximity often persists across layers.

At a global level, Figure 2(B) highlights category features of *electronic device* and *food*. In early layers, features appear mixed and scattered across both views, whereas later layers reveal clearer intra-category separation and stronger inter-category concentration. Overlaps between categories also emerge: for example, Figure 1(B) shows food-related features intersecting with *breakfast* and *dessert*. In Figure 1(D), the ball mapper view reveals a node containing *sugar*, *honey*, and *bee*, effectively marking a boundary between food and animal.

Finally, our framework enables exploration of transitions between related concepts. In Figure 2(C), a path in the ball mapper graph (i.e., a sequence of interconnected local neighborhoods) progresses from *Computer Science* to *Mathematics*, illustrating how the model captures relationships between fields of study. This is consistent with the previous observation that mapper graph captures the evolution of linguistic phenomenon along a path in the embedding space [31].

Overall, although we present only a few illustrative cases, our framework offers a powerful and much-needed tool for practitioners and researchers to scalably explore and understand the relationships between user-defined concepts and the knowledge encoded in an SAE at each layer. We believe this enhanced exploration will yield deeper insights and a more comprehensive understanding of model behavior, enabling targeted adaptations for downstream tasks such as intervention [32], editing [33], and unlearning [34].

# References


[1] Lee Sharkey, Dan Braun, and Beren Millidge. Taking features out of superposition with sparse autoencoders. *AI Alignment Forum*, 2022.

[2] Nelson Elhage, Tristan Hume, Catherine Olsson, Nicholas Schiefer, Tom Henighan, Shauna Kravec, Zac Hatfield-Dodds, Robert Lasenby, Dawn Drain, Carol Chen, et al. Toy models of superposition. *Transformer Circuits Thread*, 2022.

[3] Trenton Bricken, Adly Templeton, Joshua Batson, Brian Chen, Adam Jermyn, Tom Conerly, Nicholas L Turner, Cem Anil, Carson Denison, Amanda Askell, et al. Towards monosemanticity: Decomposing language models with dictionary learning. *Transformer Circuits Thread*, 2023.

[4] Hoagy Cunningham, Aidan Ewart, Logan Riggs, Robert Huben, and Lee Sharkey. Sparse autoencoders find highly interpretable features in language models. *arXiv preprint arXiv:2309.08600*, 2023.

[5] Francesco Tonolini, Ashley Lyons, Piergiorgio Caramazza, Daniele Faccio, and Roderick Murray-Smith. Sparse-coding variational autoencoders. *Neural Computation*, 36(12):2571–2601, 2024.

[6] Yuxuan Chen et al. Interpretability as compression: Reconsidering sae explanations of neural activations with mdl-saes. *arXiv preprint arXiv:2410.11179*, 2024.

[7] Adly Templeton, Tom Conerly, Jonathan Marcus, Jack Lindsey, Trenton Bricken, Brian Chen, Adam Pearce, Craig Citro, Emmanuel Ameisen, Andy Jones, et al. Scaling monosemanticity: Extracting interpretable features from claude 3 sonnet. *Transformer Circuits Thread*, 2024.

[8] Leo Gao, Tom Dupré la Tour, Henk Tillman, Gabriel Goh, Rajan Troll, Alec Radford, Ilya Sutskever, Jan Leike, and Jeffrey Wu. Scaling and evaluating sparse autoencoders. *arXiv preprint arXiv:2406.04093*, 2024.

[9] Daniel Balcells, Benjamin Lerner, Michael Oesterle, Ediz Ucar, and Stefan Heimersheim. Evolution of sae features across layers in llms. *arXiv preprint arXiv:2410.08869*, 2024.

[10] Tim Lawson, Lucy Farnik, Conor Houghton, and Laurence Aitchison. Residual stream analysis with multi-layer saes. *arXiv preprint arXiv:2409.04185*, 2024.

[11] Senthooran Rajamanoharan et al. Towards principled evaluations of sparse autoencoders for interpretability and control. *arXiv preprint arXiv:2405.08366*, 2024.

[12] Tatsuro Inaba, Kentaro Inui, Yusuke Miyao, Yohei Oseki, Benjamin Heinzerling, and Yu Takagi. How llms learn: Tracing internal representations with sparse autoencoders. *arXiv preprint arXiv:2503.06394*, 2025. https://arxiv.org/abs/2503.06394.

[13] Joshua Engels, Eric J. Michaud, Isaac Liao, Wes Gurnee, and Max Tegmark. Not all language model features are linear. In *ICLR (poster)*, 2025. Also available as arXiv:2405.14860.

[14] Senthooran Rajamanoharan, Arthur Conmy, Lewis Smith, Tom Lieberum, Vikrant Varma, János Kramár, Rohin Sharkey, Neel Nanda, and Chris Riggs. Improving sparse decomposition of language model activations with gated sparse autoencoders. In *NeurIPS*, 2024. Poster presentation.

[15] SAELens Contributors. SAELens: An open-source library for training and analyzing sparse autoencoders, 2024. Comprehensive framework for SAE research with visualization capabilities.

[16] Neuronpedia Contributors. Neuronpedia: An open platform for interpretability research. https://neuronpedia.org/, 2024. Interactive platform for exploring SAE features across language models.

[17] Prisma Contributors. Prisma: An open source toolkit for mechanistic interpretability in vision and video, 2024. Extension of SAE methods to vision transformers.

[18] Samuel Marks et al. Sparse feature circuits: Discovering and editing interpretable causal graphs in language models. In *ICML*. PMLR, 2024.

[19] Angie Boggust, Hyemin Bang, Hendrik Strobelt, and Arvind Satyanarayan. Abstraction alignment: Comparing model-learned and human-encoded conceptual relationships. *arXiv preprint arXiv:2407.12543*, 2024. https://arxiv.org/abs/2407.12543.

[20] Leland McInnes, John Healy, and James Melville. Umap: Uniform manifold approximation and projection for dimension reduction. *arXiv preprint arXiv:1802.03426*, 2018.

[21] Shusen Liu, Bei Wang, P-T Bremer, and Valerio Pascucci. Distortion-guided structure-driven interactive exploration of high-dimensional data. In *Computer Graphics Forum*, volume 33, pages 101–110. Wiley Online Library, 2014.

[22] Paweł Dłotko. Ball mapper: A shape summary for topological data analysis. *arXiv preprint arXiv:1901.07410*, 2019.

[23] Steven Bills et al. Open source automated interpretability for sparse autoencoder features. *EleutherAI Blog*, 2024.

[24] Johnny Lin. Neuronpedia: Interactive reference and tooling for analyzing neural networks, 2023. Software available from neuronpedia.org.

[25] Laura M Stoinski, Jonas Perkuhn, and Martin N Hebart. Thingsplus: New norms and metadata for the things database of 1854 object concepts and 26,107 natural object images. *Behavior Research Methods*, 56(3):1583–1603, 2024.

[26] Gemma Team. Gemma 2: Improving open language models at a practical size. *arXiv preprint arXiv:2408.00118*, August 2024.

[27] Tom Lieberum, Senthooran Rajamanoharan, Arthur Conmy, Lewis Smith, Nicolas Sonnerat, Vikrant Varma, János Kramár, Anca Dragan, Rohin Shah, and Neel Nanda. Gemma scope: Open sparse autoencoders everywhere all at once on gemma 2. *arXiv preprint arXiv:2408.05147*, 2024.

[28] Gonçalo Santos Paulo, Alex Troy Mallen, Caden Juang, and Nora Belrose. Automatically interpreting millions of features in large language models. In *42nd International Conference on Machine Learning*, 2025.

[29] Kevin Meng, David Bau, Alex Andonian, and Yonatan Belinkov. Locating and editing factual associations in gpt. *Advances in Neural Information Processing Systems*, 35, 2022. arXiv:2202.05262.

[30] Nikita Balagansky, Ian Maksimov, and Daniil Gavrilov. Mechanistic permutability: Match features across layers. *arXiv preprint arXiv:2410.07656*, 2024.

[31] Archit Rathore, Yichu Zhou, Vivek Srikumar, and Bei Wang. TopoBERT: Exploring the topology of fine-tuned word representations. *Information Visualization*, 22(3):186–208, 2023.

[32] Sviatoslav Chalnev, Matthew Siu, and Arthur Conmy. Improving steering vectors by targeting sparse autoencoder features. *arXiv preprint arXiv:2411.02193*, 2024.

[33] Song Wang, Yaochen Zhu, Haochen Liu, Zaiyi Zheng, Chen Chen, and Jundong Li. Knowledge editing for large language models: A survey. *ACM Computing Surveys*, 57(3):1–37, 2024.

[34] Aashiq Muhamed, Jacopo Bonato, Mona T Diab, and Virginia Smith. Saes can improve unlearning: Dynamic sparse autoencoder guardrails for precision unlearning in llms. In *ICML 2025 Workshop on Reliable and Responsible Foundation Models*, 2025.