CytoSAE: Interpretable Cell Embeddings for Hematology

Muhammed Furkan Dasdelen[1], Hyesu Lim[2,3], Michele Buck[4], Katharina S. Götze[4], Carsten Marr[1⋆], and Steffen Schneider[2,5⋆]

[1] Institute of AI for Health, Helmholtz Munich
[2] Institute of Computational Biology, Helmholtz Munich
[3] Korea Advanced Institute of Science & Technology
[4] Medical Department for Hematology and Oncology, Technical University Munich
[5] Munich Center for Machine Learning (MCML)

Abstract. Sparse autoencoders (SAEs) emerged as a promising tool for mechanistic interpretability of transformer-based foundation models. Recently, SAEs were also adopted for the visual domain, enabling the discovery of visual concepts and their patch-wise attribution to input images. While foundation models are increasingly applied to medical imaging, tools for interpreting their predictions remain limited. In this work, we propose CytoSAE, a sparse autoencoder trained on over 40,000 peripheral blood single-cell images. CytoSAE generalizes well to diverse and out-of-domain datasets-including bone marrow cytology. Here, it identifies morphologically relevant concepts which we validated with medical experts. Furthermore, we demonstrate scenarios in which CytoSAE can generate patient-specific and disease-specific concepts, enabling the detection of pathognomonic cells and localized cellular abnormalities at patch-level. We quantified the effect of concepts on a patient-level AML subtype classification task and show that CytoSAE concepts reach performance comparable to the state-of-the-art, while offering explainability on the sub-cellular level. Source code and model weights are available at https://github.com/dynamical-inference/cytosae.

Keywords: sparse autoencoders · explainability · disease classification.

1 Introduction

The advancement of foundation models has considerably popularized the application of machine learning models in various domains of medical imaging, including histopathology and cytology [22,6,14]. These models have demonstrated remarkable capabilities in tasks such as disease classification, cell segmentation, and feature extraction [14,6]. However, despite their high performance, they largely remain black-box systems. Limited transparency poses a challenge to

⋆ Co-corresponding: {carsten.marr,steffen.schneider}@helmholtz-munich.de.

reliable clinical adoption, where explainability is critical for ensuring trust in AI-driven diagnostics [22] and also becomes legally relevant [9].

To improve the explainability of deep learning models in medical imaging, various methods have been introduced [5], including Class Activation Maps (CAMs) [29] and attention-based approaches such as attention rollout [2,28]. Recently, weakly supervised multiple instance learning (MIL) methods have been applied in hematology, similar to their use in histology, enabling single-cell level explainability by identifying disease-related cells [12]. These techniques generate heatmaps or weighted attention scores that highlight relevant regions or key instances among their inputs. While these methods offer some level of insight into model predictions, they often lack fine-grained attribution and do not provide a structured understanding of the morphological concepts that contribute to a prediction. Sparse dictionary learning has emerged as a promising approach for enhancing neural network interpretability, recently in the form of sparse autoencoders (SAEs) [18,8,13,4,24,27,23,11]. SAEs are two-layer neural networks designed for unsupervised decomposition of high-dimensional representations into sparse, interpretable components. SAEs have been successfully applied in the context of language [4,16,19] and computer vision [17,26,10,7]. Although attempts have been made to extend their application to medical imaging, such as X-Ray [1] or pathology [15], their applicability to hematology remains to be demonstrated.

In this work, we introduce CytoSAE, a sparse autoencoder designed for morphological concept discovery in hematological imaging. CytoSAE is trained on embeddings from DinoBloom-B [14], a foundation model for hematology, and learns latent morphological concepts that generalize across diverse datasets, including peripheral blood smears and bone marrow cytology. We validated the discovered concepts with expert annotations. To demonstrate the utility of CytoSAE, we applied it to acute myeloid leukemia (AML) subtyping data, generating patient-level concept activations ("barcodes") and aggregating these concepts at the disease level. Our analysis shows that disease-specific concept distributions can identify morphological hallmarks of AML subtypes, providing a new level of interpretability for AI-driven hematological diagnostics.

2 CytoSAE: Concept discovery for hematology imaging

CytoSAE is a two-layer network trained with a reconstruction objective. A token from the input model is mapped to a high-dimensional latent space using a linear layer and non-linearity, then mapped back to the input space with a linear layer,

$$\begin{aligned} z &= f(x) = \mathrm{ReLU}(W_{\mathrm{enc}}(x - b_{\mathrm{dec}}) + b_{\mathrm{enc}}), \\ \hat{x} &= g(z) = W_{\mathrm{dec}} z + b_{\mathrm{dec}} \end{aligned} \tag{1}$$

where $x \in \mathbb{R}^{d_m}$ is the token embedding from DinoBloom, $W_{\mathrm{enc}} \in \mathbb{R}^{d_{\mathrm{SAE}} \times d_m}$ and $W_{\mathrm{dec}} \in \mathbb{R}^{d_m \times d_{\mathrm{SAE}}}$ are the encoder and decoder weight matrices, and $b_{\mathrm{enc}} \in \mathbb{R}^{d_{\mathrm{SAE}}}$, $b_{\mathrm{dec}} \in \mathbb{R}^{d_m}$ are the encoder and decoder biases, respectively. We optimized

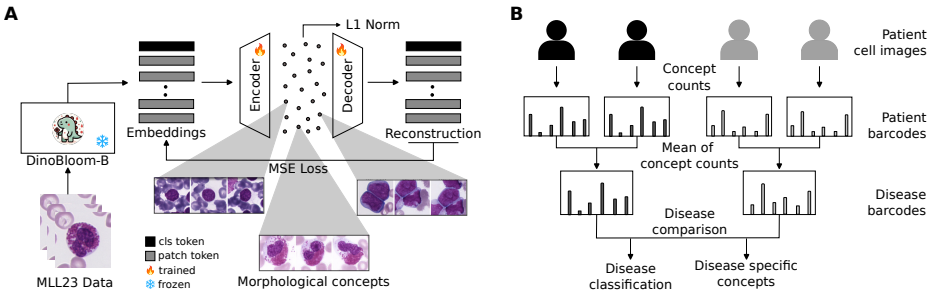

Fig. 1. Method overview. (A) CytoSAE is trained on over 40k peripheral blood single-cell images using embeddings from DinoBloom-B, disentangling compact embedding information into meaningful morphological concepts. We employ MSE loss for embedding reconstruction and apply L1 normalization to SAE latents. **(B)** Patient-level information is aggregated by counting the number of activations for each concept within the patient's single-cell images, forming a "barcode". This analysis is extended to the disease level to identify disease-specific morphological concepts.

the mean squared error (MSE) for token reconstruction while enforcing sparsity in its latent representation through an L1 norm regularizer [4],

$$\min_{W_{\text{enc}},b_{\text{enc}},W_{\text{dec}},b_{\text{dec}}} \sum_{i=1}^{K} \left(\|x_i - \hat{x}_i\|_2^2 + \lambda \|f(x_i)\|_1 \right), \tag{2}$$

for K tokens $x_1, \ldots, x_K$ extracted from the backbone. We centered the input by subtracting the decoder bias b_{dec} initialized with the geometric median of the training tokens. We used ghost gradient resampling [4] to limit unused latents.

Training and validation. We obtained tokens from DinoBloom-B [14], a hematology foundation model that partitions images into 14×14 sized patches. It processes 257 tokens, including a CLS token, each with a dimension of 768. For the SAE, we expanded this dimension by a factor of 64 and arrived at a 49,152 dimensional latent space. Among various options for backbone model representations to decompose, we used the residual stream output from the second last attention layer. After feeding cell images into DinoBloom-B, we extracted the residual layer output and passed it to the SAE. During model development, we performed an ablation study on the hyperparameters and chose the value highlighted with *: expansion factor [16, 32, 64*, 128], b_{dec} initialization method [mean, geometric median*], residual output layer [2, 5, 11*, 12], ghost gradient resampling [True*, False], L1 coefficient [8×10^{-4}, $8 \times 10^{-5*}$, 8×10^{-6}], and learning rate [4×10^{-3}, $4 \times 10^{-4*}$, 4×10^{-5}] with a constant warmup schedule (500 warmup steps). We used MSE, L1, and L0 losses to measure training performance. Each experiment was repeated three times with different random seeds.

We trained CytoSAE on the MLL23 dataset [25], which includes 41,906 peripheral blood single-cell images from 18 different cell types (Fig. 1A). The dataset includes blood cells at various maturation stages and abnormalities from both the myeloid and lymphoid lineages. After training, we validated the CytoSAE model on various datasets, encompassing both peripheral blood and bone marrow cytology.

Analysis and evaluations. We evaluated CytoSAE on MLL23 and four additional datasets. Acevedo [3] contains 17,092 single-cell images labeled into 11 classes. Matek19 [21] consists of 18,365 expert-labeled single-cell images from peripheral blood smears, classified into 15 classes. BMC [20] includes 171,373 expert-annotated cells from bone marrow smears. AML Hehr [12] contains patient-wise arranged single-cell images from 189 patients, covering four genetic AML subtypes and a healthy control group. For the analysis of our sparse encoder, we define the following key concepts:

- *Reference images* are a set of maximally activating images per SAE latent. Image-level latent activations are computed by summing the number of activating patches. We select the top-k images with the highest scores.
- *Activated frequency* is the fraction of training images for which the latent is activated. A high frequency suggests a common or nonspecific concept.
- *Mean activation value* is the average activation strength across activated images. Higher values indicate higher model confidence.
- *Label entropy* measures how many unique labels exist among activated images. Low entropy shows class specificity while high entropy shows diversity.

To identify morphological concepts, we first collected maximally activating reference images for each SAE latent from all five datasets and collaborated with an expert cytomorphologist with approx. 15 years of experience for validation. We randomly sampled 50 latents from high-activation clusters (above a mean threshold of -3) using the k-means clustering algorithm with $k = 10$, selecting 5 latents from each cluster. The expert reviewed and annotated whether these latents represented meaningful concepts.

Patch-wise analysis enables a subcellular understanding of morphological concepts and allows for validation across different datasets. To derive a global interpretation of an image for a given SAE latent, we aggregated patch-level activations into an image-level activation by counting the number of patches that activate the corresponding latent [17]. Specifically, we first binarize the SAE latent activation at the j-th patch of image i for the s-th latent using a threshold τ ($a_{i,j}[s]$). From this statistic, we derive image level (a_i), patient-level (a_p) and disease-level (a_d) features:

$$a_{i,j}[s] = I(h_{i,j}[s] > \tau), \tag{3}$$

$$a_i[s] = \sum_{j=1}^{n_i} a_{i,j}[s],\ a_p[s] = \frac{1}{N_p} \sum_{i \in I_p} a_i[s],\ a_d[s] = \frac{1}{N_d} \sum_{p \in P_d} a_p[s], \tag{4}$$

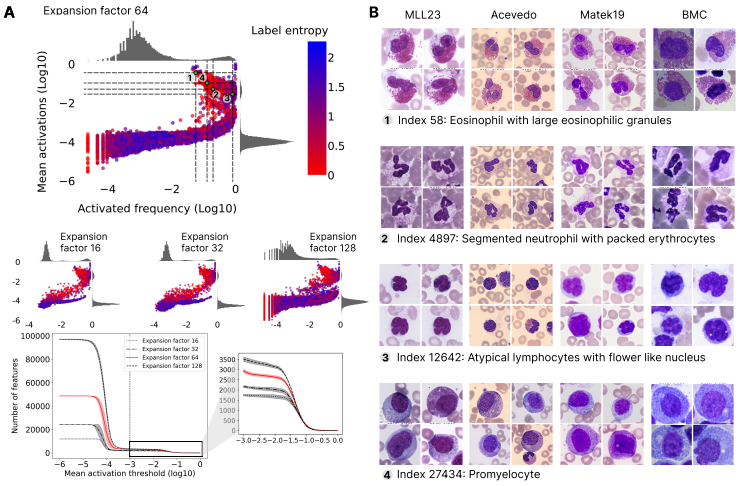

Fig. 2. CytoSAE discovers morphologically relevant concepts across datasets. (A) SAE latents with varying expansion factors (x-axis: $\log_{10}$ of activation frequency, y-axis: $\log_{10}$ of mean activation), color-coded by label entropy. Histograms show latent distributions along corresponding axes. Bottom left: Comparison of the number of latents exceeding a given threshold across models with different SAE dimensions. **(B)** Reference images from four datasets corresponding to selected latents.

where $I(\cdot)$ is an indicator function, I_p represents the set of images associated with patient p, and P_d represents the set of patients diagnosed with disease d. This hierarchical aggregation allows the identification of shared morphological features at the patient and disease levels while accounting for variations in image and patient count.

Disease Classification. Using the AML Hehr dataset, we generated patient barcodes a_p (Eq. 4). To assess the predictive power of barcodes and encoded concepts, we applied linear probing to quantify the intermediate representations captured by the SAE. Specifically, we attempted to classify patient disease status using only barcodes. For linear probing, we used logistic regression with L2 regularization. The regularization coefficient was set to $\frac{c \times n}{100}$, where n is the number of training samples and c corresponds to the number of classes [14].

3 Results

CytoSAE discovers morphologically relevant concepts. First, we evaluated the effect of different expansion factors on the latent distribution. Specifically, we analyzed the mean latent activations and their activation frequencies on the MLL23 dataset. The latent distribution is depicted in Fig. 2A, revealing

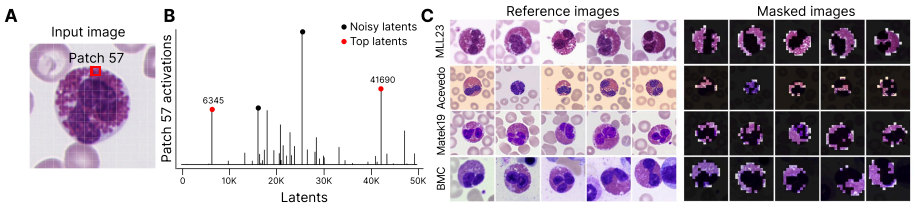

Fig. 3. Patch-level morphological concepts are consistent across diverse datasets. (A) We highlighted patch 57 of an example image, containing eosinophilic granules, and fed its corresponding token into the SAE. **(B)** Activated latents of token 57, along with the top two latents after filtering out those are noisy. **(C)** Reference images with their corresponding patches that activate the selected latents.

two distinct clusters of activation patterns. We focused on latents where the model was confident about the underlying concept, i.e., those with high mean activation. To quantify this, we counted the number of latents exceeding different mean activation thresholds. Higher-dimensional SAEs produced more latents above any given threshold (Fig. 2A). However, while the difference was substantial at lower thresholds, the gap became less pronounced at higher thresholds (e.g., at threshold -3, the number of activated latents was 1748±56, 2170±61, 2936±114, and 3518±145 for expansion factors 16, 32, 64, and 128, respectively). Notably, the number of latents in the highly activated cluster—those of primary interest—remained relatively stable (Fig. 2A, bottom left). Next, we randomly selected latents from the high mean activation cluster and visualized the recovered concepts across the four evaluation datasets (Fig. 2B). The extracted concepts were consistent across datasets and qualitatively, were not affected by domain shift. Expert manual labeling of these concepts (see Sec. 2) confirmed that CytoSAE successfully captured various morphologically relevant features.

Patch-wise attribution allows analysis on sub-cellular level. Next, we leveraged CytoSAE for patch-level attributions of the discovered concepts. In Fig. 3, we show the activated latents for an example patch containing eosinophilic granules and selected the top two latents after filtering out those that were always-on or noisy (threshold of 0.1). Next, we retrieved reference images that also activated these selected latents. Using patch-level latent activations from the retrieved images, we generated segmentation maps that highlighted the same concept present in the given input patch. Specifically, for a given image x_i and an SAE latent index s, we visualized the concept by multiplying each patch $x_{i,j}$ with its corresponding latent activation value $h_{i,j}[s]$. Notably, all retrieved images–including those from different datasets and bone marrow cytology–consistently highlighted eosinophilic granules, demonstrating that the latents successfully capture meaningful morphological concepts.

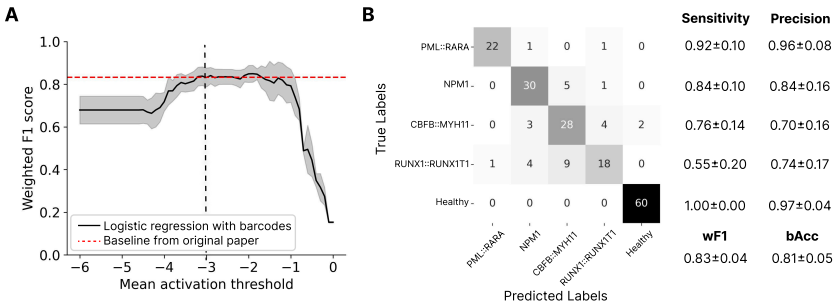

Fig. 4. Patient-wise concepts can predict disease. (A) We trained a logistic regression classifier on patient barcodes, varying the number of latents; latents above the threshold were used in classification, with others set to zero. **(B)** Classification performance of logistic regression using patient barcodes (wF1=0.83) is comparable to a fully deep learning approach (wF1 $\approx$ 0.83)[12].

Patient-wise analysis discovers disease-specific morphologies. For patient-wise analysis, we used AML peripheral blood single-cell image data [12]. This dataset includes four subtypes of AML: (i) APL with *PML::RARA* fusion, (ii) AML with *NPM1* mutation, (iii) AML with *CBFB::MYH11* fusion (without NPM1 mutation), and (iv) AML with *RUNX1::RUNX1T1* fusion, along with healthy stem cell donors. For each patient, we counted how many times a latent (morphological concept) was activated across single-cell images and generated a patient-specific barcode. By averaging activation counts across patients with the same disease, we created disease barcodes (Fig. 1B, Eq. 4). This approach enabled both patient-wise comparisons and disease classification while identifying morphological concepts uniquely expressed in individual patients. To evaluate the discriminative power of these morphological concepts, we performed disease classification using patient barcodes while varying the number of latents included in the classification (Fig. 4A). Although adjusting the threshold from -6 to 0 ($\log_{10}$) reduced the number of latents used in classification (Fig. 2A), classification performance remained stable up to a certain threshold and using only meaningful concepts resulted in higher classification performance (Fig. 4A). Our findings confirm that classification performance is only affected when meaningful concepts are removed. Multi-class classification at a threshold of -3 ($\log_{10}$) achieved a weighted F1-score of 0.832 $\pm$ 0.044 which is similar to the performance of fully deep learning baseline [12]. By comparing disease barcodes, we identified the top-100 latents that were differentially expressed between the *CBFB::MYH11* and *PML::RARA* subtypes (50 for each disease). Experts annotated these concepts and assessed whether they represented disease-specific morphological features recognizable by a cytomorphologist. Among the top features, 5 out of 10 and 32 out of 50 were identified as *CBFB::MYH11*-specific, while 4 out of 10 and 10 out of 50 were classified as *PML::RARA*-specific. *CBFB::MYH11*-specific concepts predominantly high-

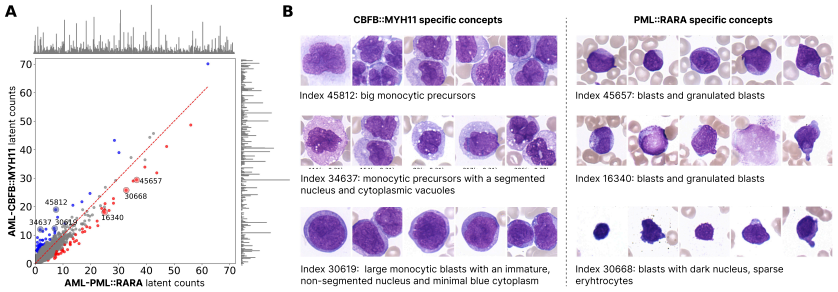

Fig. 5. Disease-wise comparison reveals disease-specific morphological features. (A) Mean count of latents across patients. Blue: top 50 latents that are more frequently activated in the *CBFB::MYH11* subtype; red: top 50 latents for *PML::RARA* subtype. Disease barcodes are positioned along the corresponding axes.**(B)** Representative images of disease-specific concepts.

lighted large monocytic precursors with segmented nuclei and cytoplasmic vacuoles (Fig. 5B). In contrast, *PML::RARA*-specific concepts were characterized by granulated blasts and morphological features suggestive of anemia (Fig. 5B).

Ablation and variation studies. With our final set of parameters, 2936±114 concepts are discovered at a threshold of -3. Varying the expansion factor changes the number of concepts as this threshold in a range of 1748±56 (expansion 16) to 3518±145 (expansion 128), remaining within the same order of magnitude. Increasing the sparsity regularizer λ by 10x results in too few concepts (128±83) at low L0, while lowering by 10x results in a two orders of magnitude increase in the L0, motivating our current choice. Increasing the learning rate too much resulted in unstable training, while lowering to 4e-5 did not meaningfully impact the results, yielding 2755±108 concepts. We then varied the number of layers: At layer 2 and 7, we discovered 85±4 and 1253±34 concepts while in the last layer 12, 15161±3477 concepts crossed the threshold. This indicates that the layers between 7–11 are the most interesting choices for finding a good balance in the number of concepts. Ghost gradient regularization did not substantially influence the number of concepts (2907±133) but decreased the fraction of dead features from 0.92±0.03 to 0.57±0.03.

4 Conclusion

We introduced CytoSAE for morphological concept discovery in hematological cytology. Our findings demonstrate that CytoSAE successfully learned interpretable morphological features across different datasets, including peripheral blood smears and bone marrow cytology, while maintaining generalizability to out-of-domain datasets. This capability is crucial in medical AI, where real-world applications often require robustness to dataset shifts caused by varia-

tions in staining, background artifacts, and imaging protocols. We showed patch-level, image-level, patient-level, and disease-level analysis and demonstrated interpretable disease-level subtype classification.

While CytoSAE provides a powerful approach for morphological feature discovery, some limitations remain. Manual expert validation was conducted at the dataset level, but we did not evaluate reference images within each individual patient to avoid human selection bias in concept discovery. Future work could explore patient-specific concept retrieval, though care must be taken to minimize subjectivity in the selection process. Future studies could incorporate semi-supervised learning techniques to refine the discovered features further.

Author contributions. Conceptualization: StS, CM; Methodology: MFD, HL, StS; Software: MFD, HL; Investigation: MFD; Data Curation: MB, KSG, MFD; Writing–Original Draft: HL, MFD; Writing–Editing: StS, HL, CM.

Acknowledgments. C.M. acknowledges funding from the European Research Council (ERC) under the European Union's Horizon 2020 research and innovation program (Grant Agreement No. 866411 & 101113551) and support from the Hightech Agenda Bayern.

Disclosure of Interests. The authors have no competing interests to declare that are relevant to the content of this article.

References

1. Abdulaal, A., Fry, H., Montaña-Brown, N., Ijishakin, A., Gao, J., Hyland, S., Alexander, D.C., Castro, D.C.: An x-ray is worth 15 features: Sparse autoencoders for interpretable radiology report generation. arXiv preprint arXiv:2410.03334 (2024)
2. Abnar, S., Zuidema, W.: Quantifying attention flow in transformers. arXiv preprint arXiv:2005.00928 (2020)
3. Acevedo, A., Merino, A., Alférez, S., Molina, Á., Boldú, L., Rodellar, J.: A dataset of microscopic peripheral blood cell images for development of automatic recognition systems. Data in brief **30**, 105474 (2020)
4. 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., Olah, C.: Towards monosemanticity: Decomposing language models with dictionary learning. Transformer Circuits Thread (2023), https://transformer-circuits.pub/2023/monosemantic-features/index.html
5. Chen, H., Gomez, C., Huang, C.M., Unberath, M.: Explainable medical imaging ai needs human-centered design: guidelines and evidence from a systematic review. NPJ digital medicine **5**(1), 156 (2022)
6. Chen, R.J., Ding, T., Lu, M.Y., Williamson, D.F., Jaume, G., Song, A.H., Chen, B., Zhang, A., Shao, D., Shaban, M., et al.: Towards a general-purpose foundation model for computational pathology. Nature Medicine **30**(3), 850–862 (2024)

7. Cywiński, B., Deja, K.: Saeuron: Interpretable concept unlearning in diffusion models with sparse autoencoders. arXiv preprint arXiv:2501.18052 (2025)
8. Elhage, N., Nanda, N., Olsson, C., Henighan, T., Joseph, N., Mann, B., Askell, A., Bai, Y., Chen, A., Conerly, T., DasSarma, N., Drain, D., Ganguli, D., Hatfield-Dodds, Z., Hernandez, D., Jones, A., Kernion, J., Lovitt, L., Ndousse, K., Amodei, D., Brown, T., Clark, J., Kaplan, J., McCandlish, S., Olah, C.: A mathematical framework for transformer circuits. Transformer Circuits Thread (2021), https://transformer-circuits.pub/2021/framework/index.html
9. European Parliament and Council: in Regulation of the European Parliament and of the Council laying down harmonised rules on Artificial Intelligence (Artificial Intelligence Act) (2024), (ed European Parliament and Council)
10. Fry, H.: Towards multimodal interpretability: Learning sparse interpretable features in vision transformers (Apr 30 2024), https://www.lesswrong.com/posts/iYFuZo9BMvr6GgMs5/case-study-interpreting-manipulating-and-controlling-clip
11. Gao, L., la Tour, T.D., Tillman, H., Goh, G., Troll, R., Radford, A., Sutskever, I., Leike, J., Wu, J.: Scaling and evaluating sparse autoencoders. arXiv preprint arXiv:2406.04093 (2024)
12. Hehr, M., Sadafi, A., Matek, C., Lienemann, P., Pohlkamp, C., Haferlach, T., Spiekermann, K., Marr, C.: Explainable ai identifies diagnostic cells of genetic aml subtypes. PLOS Digital Health **2**(3), e0000187 (2023)
13. Huben, R., Cunningham, H., Smith, L.R., Ewart, A., Sharkey, L.: Sparse autoencoders find highly interpretable features in language models. In: The Twelfth International Conference on Learning Representations (2023)
14. Koch, V., Wagner, S.J., Kazeminia, S., Sancar, E., Hehr, M., Schnabel, J.A., Peng, T., Marr, C.: Dinobloom: a foundation model for generalizable cell embeddings in hematology. In: International Conference on Medical Image Computing and Computer-Assisted Intervention. pp. 520–530. Springer (2024)
15. Le, N.M., Patel, N., Shen, C., Martin, B., Eng, A., Shah, C., Grullon, S., Juyal, D.: Learning biologically relevant features in a pathology foundation model using sparse autoencoders. In: Advancements In Medical Foundation Models: Explainability, Robustness, Security, and Beyond (2024)
16. Lieberum, T., Rajamanoharan, S., Conmy, A., Smith, L., Sonnerat, N., Varma, V., Kramár, J., Dragan, A., Shah, R., Nanda, N.: Gemma scope: Open sparse autoencoders everywhere all at once on gemma 2. arXiv preprint arXiv:2408.05147 (2024)
17. Lim, H., Choi, J., Choo, J., Schneider, S.: Sparse autoencoders reveal selective remapping of visual concepts during adaptation. arXiv preprint arXiv:2412.05276 (2024)
18. Makhzani, A., Frey, B.: K-sparse autoencoders. arXiv preprint arXiv:1312.5663 (2013)
19. Marks, S., Rager, C., Michaud, E.J., Belinkov, Y., Bau, D., Mueller, A.: Sparse feature circuits: Discovering and editing interpretable causal graphs in language models. arXiv preprint arXiv:2403.19647 (2024)
20. Matek, C., Krappe, S., Münzenmayer, C., Haferlach, T., Marr, C.: An expert-annotated dataset of bone marrow cytology in hematologic malignancies (2021). https://doi.org/10.7937/TCIA.AXH3-T579, data set
21. Matek, C., Schwarz, S., Marr, C., Spiekermann, K.: A single-cell morphological dataset of leukocytes from aml patients and non-malignant controls (2019). https://doi.org/10.7937/tcia.2019.36f5o9ld, data set

22. Moor, M., Banerjee, O., Abad, Z.S.H., Krumholz, H.M., Leskovec, J., Topol, E.J., Rajpurkar, P.: Foundation models for generalist medical artificial intelligence. Nature **616**(7956), 259–265 (2023)
23. Rajamanoharan, S., Conmy, A., Smith, L., Lieberum, T., Varma, V., Kramár, J., Shah, R., Nanda, N.: Improving dictionary learning with gated sparse autoencoders. arXiv preprint arXiv:2404.16014 (2024)
24. Sharkey, L., Chughtai, B., Batson, J., Lindsey, J., Wu, J., Bushnaq, L., Goldowsky-Dill, N., Heimersheim, S., Ortega, A., Bloom, J., et al.: Open problems in mechanistic interpretability. arXiv preprint arXiv:2501.16496 (2025)
25. Shetab Boushehri, S., Gruber, A., Kazeminia, S., Matek, c., Spiekermann, K., Pohlkamp, C., Haferlach, T., Marr, C.: A large expert-annotated single-cell peripheral blood dataset for hematological disease diagnostics. medRxiv pp. 2025–02 (2025)
26. Stevens, S., Chao, W.L., Berger-Wolf, T., Su, Y.: Sparse autoencoders for scientifically rigorous interpretation of vision models. arXiv preprint arXiv:2502.06755 (2025)
27. Templeton, A.: Scaling monosemanticity: Extracting interpretable features from claude 3 sonnet. Anthropic (2024)
28. Wagner, S.J., Reisenbüchler, D., West, N.P., Niehues, J.M., Zhu, J., Foersch, S., Veldhuizen, G.P., Quirke, P., Grabsch, H.I., van den Brandt, P.A., et al.: Transformer-based biomarker prediction from colorectal cancer histology: A large-scale multicentric study. Cancer Cell **41**(9), 1650–1661 (2023)
29. Zhou, B., Khosla, A., Lapedriza, A., Oliva, A., Torralba, A.: Learning deep features for discriminative localization. In: Proceedings of the IEEE conference on computer vision and pattern recognition. pp. 2921–2929 (2016)