Can sparse autoencoders make sense of gene expression latent variable models?

Viktoria Schuster[1,2]
[1] Eric and Wendy Schmidt Center,
Broad Institute of MIT and Harvard
[2] Department of Computer Science,
University of Copenhagen
vschuster@broadinstitute.org

Abstract

Sparse autoencoders (SAEs) have lately been used to uncover interpretable latent features in large language models. By projecting dense embeddings into a much higher-dimensional and sparse space, learned features become disentangled and easier to interpret. This work explores the potential of SAEs for decomposing embeddings in complex and high-dimensional biological data. Using simulated data, it outlines the efficacy, hyperparameter landscape, and limitations of SAEs when it comes to extracting ground truth generative variables from latent space. The application to embeddings from pretrained single-cell models shows that SAEs can find and steer key biological processes and even uncover subtle biological signals that might otherwise be missed. This work further introduces **scFeatureLens**, an automated interpretability approach for linking SAE features and biological concepts from gene sets to enable large-scale analysis and hypothesis generation in single-cell gene expression models.

1 Introduction

Neural networks have proven to be powerful tools for analyzing complex data, yet they often lack inherent interpretability from a human perspective [1]. While various approaches like disentanglement [2], adversarial training [2], and over-determined networks [3] have shown some success in improving model interpretability [2, 4], they fall short of providing a comprehensive understanding of all learned features within a model [5]. Recent research has revealed that features in neural networks are often learned in a state of superposition [6], where individual neurons encode multiple features (termed polysemanticity), and single features are distributed across multiple neurons. Simply speaking, each feature superposition is a linear combination of all dimensions in the latent space. In light of this complexity, sparse autoencoders (SAEs) [7] have emerged as a promising tool for interpreting entire neural network layers [8–11]. The application of SAEs to large language model layers has demonstrated remarkable success in reducing polysemanticity, effectively translating language model activations into singular, monosemantic features [8–11]. However, this research has primarily been limited to language models and transformer architectures. Given that superpositions are strongly influenced by data structure [6], there is a pressing need to extend this approach to different types of hidden streams and data domains.

Biology and health present a wealth of complex data and machine learning applications [12–17]. Single-cell gene expression (scRNAseq) data, for example, provide valuable insight into cellular functions and malfunctions within the human body. However, the high dimensionality and noise inherent in this data present

Machine Learning in Computational Biology 2025.

significant analytical challenges [18–20]. Several generative models have been suggested to model scRNAseq and multi-omics data and produce lower-dimensional representations for analysis [20–30]. Representation learning is of high interest in this field, as it is generally assumed that these high-dimensional biological processes are guided by lower-dimensional concepts such as regulatory programs.

This work investigates the limitations and potential applications of SAEs for models trained on high-dimensional and sparse single-cell gene expression data. It examines superpositions and SAE features derived from models trained on simulated data and applies SAEs to pre-trained models [31, 32]. This focus on pre-trained models is motivated by the wish to understand *what* models learn from the data in terms of concepts and to be able to compare them, rather than learning features from the data de novo. Code for reproducibility is available here. The core insights and contributions are:

- Distribution type and distance of hidden generative variables affect variable recovery.
- SAEs extract meaningful features from single-cell expression models that successfully steer cells into desired programs. Features can act either locally or globally.
- **scFeatureLens**: An analysis pipeline for interpreting single-cell expression models by automatically annotating SAE features with biological concepts derived from ontologies, available on GitHub.

2 Related work

The application of SAEs and dictionary learning in general has attracted a lot of attention in the field of natural language processing [8–11]. Recent research has demonstrated the efficacy of these methods in uncovering fine-grained features within language models, such as identifying hierarchical semantic structures [33], specific scriptures [9], and causal features of object identification [10]. Others have presented improvements in the tradeoff between sparsity and reconstruction, reduced the occurrence of dead neurons, and developed metrics for evaluating quality based on hypothesized features [11]. While much of the focus has been on language models, efforts to enhance interpretability have extended to other architectural domains. Bau et al. [34] developed a method for scoring convolutional activations based on pre-defined visual concepts, thereby enhancing our understanding of learned visual features.

In contrast to these advancements, the application of SAEs to the field of biology has been limited. Except for recent applications to protein language models [35, 36], dictionary learning has primarily been employed as a direct method for learning sparser representations from count data [37–39] or aligning representations more closely with specific biological concepts such as pathways [40]. More commonly, efforts to enhance the interpretability of biological representations have focused on disentanglement. Disentanglement is often applied to separate technical bias from biological signal through approaches such as adversarial training [41], sparsity-inducing priors [38], overcomplete autoencoders [14], or architectural modularity [31, 42].

3 Sparse autoencoders

In representation learning, data is generally assumed to exist on a lower-dimensional manifold due to dependencies between features [43]. Reducing the dimensionality into a latent representation through unsupervised learning can help reveal underlying structure. With a different constraint than dimensionality, data structure can also be revealed in a higher-dimensional setting by employing sparsity constraints on the latent representation [7]. This has lately been exploited to disentangle the polysemanticity of hidden layers in large language models [8–11]. Figure 1A shows a schematic of SAEs and superpositions.

Vanilla SAE: The simplest SAE maps an input $\mathbf{x} \in \mathbb{R}^d$ to a higher-dimensional hidden activation vector $\mathbf{z} \in \mathbb{R}^l_{\geq 0}$ and back, with an additional objective to promote sparsity in the activation space. The encoder is defined as

$$\mathbf{z} = \mathrm{ReLU}(\mathbf{W}_\phi(\mathbf{x}) + \mathbf{b}_\phi) \tag{1}$$

and the decoder as

$$\hat{\mathbf{x}} = \mathbf{W}_\theta(\mathbf{z}) + \mathbf{b}_\theta \tag{2}$$

with ϕ and θ indicating encoder and decoder parameter sets, respectively. The loss is given by

$$\mathcal{L} = \|\mathbf{x} - \hat{\mathbf{x}}\|_2^2 + \lambda\|\mathbf{z}\|_1 \tag{3}$$

where the first term is the mean squared error (MSE) loss for reconstruction. The second term is the sparsity penalty in the form of an L1 loss weighed by hyperparameter λ, which will be referred to as the L1 weight.

Other SAE setups: A widely used version of the SAE uses an additional pre-network bias $\mathbf{b}_{pre}$ term applied to $\mathbf{x}$ before encoding [9], which has shown to improve performance [6]. k-sparse autoencoders additionally use a different activation function (TopK) to directly control the number of active neurons (removing the need for the L1 loss) [44]. The latest advance in SAE research has been to reduce the number of dead hidden neurons by initializing encoder $\mathbf{W}_\phi$ and decoder $\mathbf{W}_\theta$ as transposes of each other and including dead neurons in an auxiliary loss [11].

4 Simulation experiments

As recent use cases of SAEs are mainly limited to the activations of large language models, this work presents an analysis of some common SAEs in a simulated setting with known underlying variables. The simulated data are inspired by sparse count data as we see in (single-cell) expression. Two datasets were created, a "small" one for hyperparameter sweeps with lower dimensionalities and a "large" one with realistic number of samples and dimensions in the observed variables Y (the "counts"). The simulation is based on a hierarchical generative process starting with hidden variables X representing core programs, cell-type specific factors A, and batch effects B with defined connectivity $\mathbf{M}$ of shape $(|Y|, |X|)$. The underlying hypotheses data simulation process are explained in detail in Appendix A.2.1 and depicted in Figure 1B. What follows is a discussion of what aspects of the data generation process can be recovered in superposition and SAE features, as well as performance differences of "Vanilla", "ReLU" [9], and "TopK" [11] SAE architectures.

4.1 What is learned in superposition?

Experimental set up: Autoencoders were trained with a variety of structures and training hyperparameters (Appendix A.2.2, Table S3) on observables Y of the large simulation data. Learned representations were extracted and used to compute superposition vectors and fits through linear regression (Appendix A.2.3).

Results: Observables are perfectly learned when validation loss is sufficiently low, and hidden variables can be partially recovered from latent representations (Figure 1C). Recovery of variables follows a distinct pattern: variables X'' directly upstream of Y are most accurately reconstructed, followed by A, B, X', and X. Regression fits R^2 did not scale linearly with the distance from Y, suggesting that the type of variables and their role in the data generation process influence recovery. Additionally, recovery of more distant variables seemed to decrease with larger (deep and wide) models despite lower validation loss (Figure S2).

4.2 How do different SAEs perform?

Experimental set up: A sweep of different SAE architectures and a wide range of hyperparameters (Table S4) was performed on embeddings from an autoencoder trained to perfectly recover observables and hidden variables X of the small simulation data (Appendix A.2.2). All SAEs were trained on the extracted representations and evaluation metrics were computed as described in Appendix A.2.4.

Results: Reconstruction and sparsity. Briefly summarized, reconstruction losses of Vanilla and ReLU SAEs were more robust compared to TopK (Figures S4A-B, S5). Sparsity (fraction of dead/active neurons) strongly increased for L1 weights above 10^{-3} and a k below $50\,\%$ (Figure S12), and strongly depended on the learning rate (Figure S6). As a result, the analysis was continued with the overall best-performing learning rate of 10^{-4}.

Recovery of X'. Figure S4 shows that a small hidden size can be detrimental to the performance and interpretability of TopK models. In terms of good recovery (high correlation between features and observables)

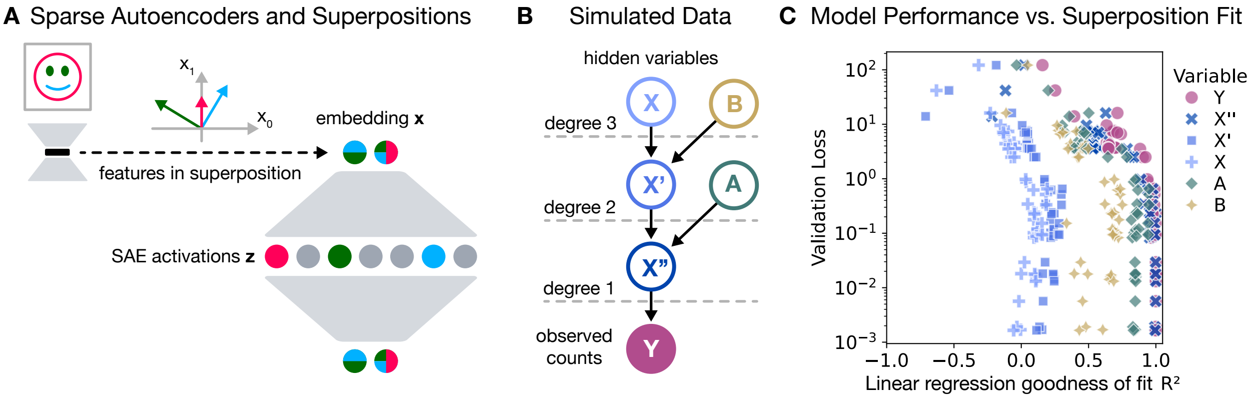

Figure 1: **Sparse Autoencoders, data simulation, and hidden variable recovery.** **A** Schematic of superpositions and SAEs. Given a sample generated from 3 features and encoded into a 2D latent space, there are more features than dimensions. The features have to be learned as linear combinations of the latent dimensions (meaning they are in superposition). These features can be disentangled by projecting them into a higher-dimensional space via SAEs. **B** Schematic of the data generation process. Filled and non-filled circles represent observed and hidden variables, respectively. Arrows indicate the dependencies between variables from parent to child. There are 3 levels to this generative process, indicated by the distance from observed counts Y (details in Appendix A.2.1). X' and X'' represent the hidden states altered by B and A, respectively. **C** AE performance (validation loss) plotted against superposition fit (R^2). Coefficients of determination R^2 describe how well a given variable can be retrieved from the latent embedding (see Appendix A.2.3 for details). Colors and marker styles match the variables from B.

with little redundancy for variables X and Y, the Vanilla SAEs showed the best tradeoff and TopK the worst (Figures S9,10). The best performing Vanilla models used L1 weights of 10^{-3} (10^{-4} for ReLU) with hidden dimensionalities of $5-50\times$ the size of the latent space (for best recovery and 1-5 neurons per variable). The number of features per variable scaled roughly exponentially for the k-sparse autoencoder (TopK) over the hidden dimension irrespective of k (Figure S12). For Vanilla and ReLU SAEs, there was no such scaling tendency and the L1 weight strongly determined the rate at which the number of neurons per variable grow, which is a disadvantage of these SAEs.

4.3 How well can data variables and structure be recovered?

Experimental set up: SAEs were trained on AE embeddings of the large simulation data from section 4.1 according to the results from the previous sweep (Appendix A.2.5). They were evaluated in terms of correlation between SAE neuron activations and data variables, and to what extent the structure of the generative connectivity matrix $\mathbf{M}$ is recovered by the SAE. Cosine similarities between observables and SAE neurons $(|Y|, |z|)$ were used to create pseudo connectivity matrices for different thresholds. These pseudo connectivity matrices were compared to $\mathbf{M}$ through Binomial tests (Appendix A.2.5).

Results: Variables. Recovery of a given variable from SAE features was measured as the correlation between that variable and SAE neuron activations. Observed variables Y and directly upstream hidden variables X'' could be nearly perfectly recovered, especially for larger hidden dimensionalities. The original generative random variables X, however, are not directly represented by individual SAE features. Comparing these results to baselines from PCA, ICA, and SVD (Table S5), there was no significant increase in superposition identifiability between SAE and baseline methods. The SAE's advantages, however, lie in the discovery of unknown features and providing a convenient way of extracting learned features that can be used for model steering.
Structure. In real-world applications, it may be difficult to identify generative variables due to the prevalence

of features corresponding to observables. We may, however, be able to identify concepts and structures in the data generation process in a different way. Figure S13 demonstrates an alternative approach comparing the structure of SAE features and observables with the data generation matrix $\mathbf{M}$. For each feature, the best matching X'' variable is determined. Their entries of pseudo connectivity matrix and $\mathbf{M}$ are used to calculate how many entries of Y match for each feature-X'' pair. On average, $75\,\%$ to $95\,\%$ of the entries in $\mathbf{M}$ could be recovered (with 20th to 70th percentiles of cosine similarity as thresholds, respectively).

5 Case Study: Extracting and annotating meaningful features from single-cell models

Next, SAEs were applied to representations from models pre-trained on single-cell RNAseq and multi-omics data. SAE hyperparameters were evaluated on a model trained on three different datasets from Schuster et al. [31]. Meaningfulness of extracted features and how they can be used for steering samples towards biological programs is demonstrated in a manual evaluation. A major contribution of this work is an automated analysis pipeline for practical large-scale interpretability analysis demonstrated on multiDGD [31] and the latest version of Geneformer [32]. While we do not include VAE-based models in this comparison, the approach would be the same except for having to choose between predicted encoder means and reparametrized embeddings. In this case study, Gene Ontology (GO) terms [45, 46], which provide functional information about sets of genes, represent examples of biological concepts.

5.1 SAE training

Experimental set up: SAE hyperparameters were evaluated on a small sweep for extracted representations from multiDGD instances trained on human bone marrow [47], mouse gastrulation [48], and human brain data [49] (Appendix A.3.1). Results scaled well compared to the preceding simulation experiments. Hyperparameters and training are described in Appendix A.3.3. The final SAE hidden dimension was chosen to be 10000 neurons in favor of redundant features over a lack of sensitivity. Another SAE was trained on representations of the human bone marrow data extracted from Geneformer for the automated pipeline. See Appendices A.3.1 and A.3.3 for embedding extraction, training details and compute estimates.

Results: In the SAE trained on multiDGD embeddings from the human bone marrow data, 5318 remained as "live" SAE neurons with 185.7 firing on average per cell. Since the representations are highly structured with respect to cell type, average activations of cell types naturally create unique patterns (Figure S15). Significant differences in activations with respect to cell types revealed two major SAE feature categories: "local" and "global" (categorization and significance measure in Appendix A.3.6). Local features are characterized by higher activations for a single cell type compared to all other cell types. Among the 5318 live neurons, there were 4410 global and 908 local features. Training the SAE on different random seeds revealed robust results in the number of live neurons and feature types (Table S6). Monocytes and cells along the red blood cell differentiation trajectory accounted for most of the local features (not related to numbers of cells in the data, Figure S17).

5.2 Manual feature analysis

Experimental set up: Evaluating what biological potential functions a feature has is difficult. In this work, concepts of biological function of a given feature was approximated by GO terms. In order to create gene sets associated with a given SAE feature, Differential Gene Expression (DGE) analysis was performed on either "perturbed-vs-normal" or "high-vs-low" sample subsets. Perturbed subsets were created by selecting a cell type along the global feature trajectory, computing sample activations, maximizing the feature of interest (also called "steering"), and predicting the perturbed representations. "High-vs-low" subsets were created by selecting the 95th and 5th percentile activations of sample representations per feature (excluding 0 if done in a specific cell type). DGE analysis was then performed based on the single-cell model's predicted expression values according to Appendix A.3.4.

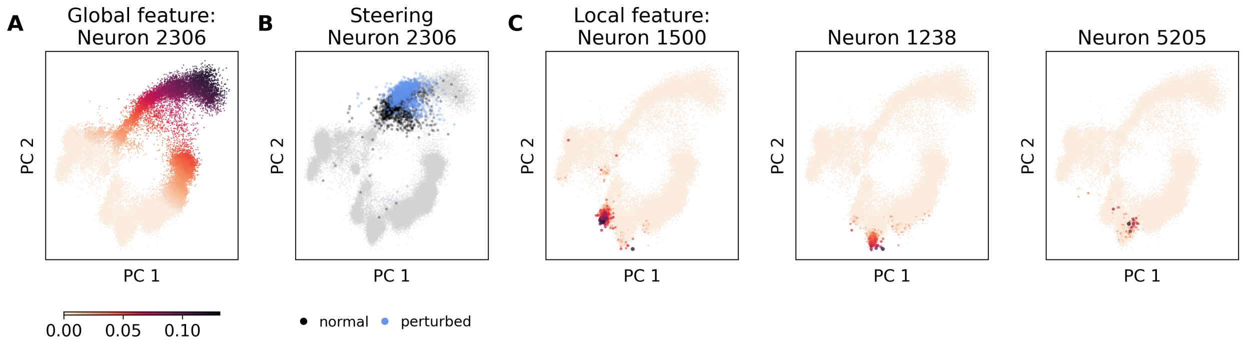

Figure 2: **SAE features in multiDGD' bone marrow representation space.** Visualized as PCA plots of the extracted 20-dimensional representations. **A** Representations colored by activations of neuron 2306. Lightest color represents zero values, darker colors present higher activations. **B** Representations from SAE feature steering/perturbation experiments on Proerythroblast representations (black, "normal"). Representations predicted by the SAE after maximizing feature 2306 are shown in blue ("perturbed"). **C** Local features. From left to right: Representations colored and size-scaled by activations of neurons 1238, 5205, and 1500.

Results: Global features. Red blood cell differentiation is a prominent biological process in this dataset. Based on the rule set described in Appendix A.3.2, neuron 2306 was identified as the best aligning feature (Figure S20). Activations are shown in Figure 2A. Although feature 2306 was most prevalent along the axis of red blood cell differentiation, moderate activations were also found in NK and some CD8+ T cells. Steering was performed by maximizing feature 2306 in HSCs, Proerythroblasts, NK, and CD8+ T cells (Figure 2B). While each analysis resulted in different gene sets and GO terms, the identified processes are highly specific and show a strong functional overlap (Table S8). Results highlight ion homeostasis and gas transport, which are crucial processes in erythropoiesis and cytotoxicity. This global feature presents an important higher-level and more general concept in cellular processes of the bone marrow.

Local features. Among local features, B cells presented multiple of the top 20 features regarding mean activation. This analysis investigates one of the most significant local features for each of the three different types of B cells present in the data: Transitional, Naive CD20+, and B1 B cells. Activations are shown in Figure 2C. DGE analysis ("high-vs-low") and GO term enrichment analysis within each cell type revealed distinctive molecular signatures of each feature. Feature 1500 (Transitional B cells) was characterized by GO terms related to the response to interferon beta. Interferon beta is a critical regulator during early transitional B cell development, playing a role in differentiation towards a regulatory phenotype vs. an inflammatory phenotype [50]. Feature 1238 (Naive CD20+ B cells) showed enrichment in histone H3R26 citrullination, an indicator of cellular aging [51]. Another sign of cell aging is increased closed chromatin. Cells with high activations of feature 1238 had significantly more closed chromatin. The 95th percentile had an average chromatin openness of 0.03322 ± 0.00124 SEM compared to the 5th percentile with a mean of 0.04393 ± 0.00002 (based on "high": 35 samples, "low": 3483 samples, 129921 columns). Feature 5205 (B1 B cells) presented enriched GO terms predominantly centered around molecular functions associated with pattern recognition receptor activities. Specifically, the terms highlighted activation of the innate immune system, referencing key receptors such as toll-like receptor 4, haptoglobin, and RAGE receptor. The activation profile of these cells suggests a trajectory towards increased immune cell activity and potential cytotoxicity, paralleling observations from previous results on T cells.

5.3 scFeatureLens: Automated SAE analysis demonstrated on multiDGD and Geneformer

Manual analyses, while useful for validation, are limited in their scalability. Deriving biological semantic concepts in an automated fashion is highly desirable and a key contribution of this work. The pipeline presented here can be adapted to any database using gene sets to characterize semantic concepts. The

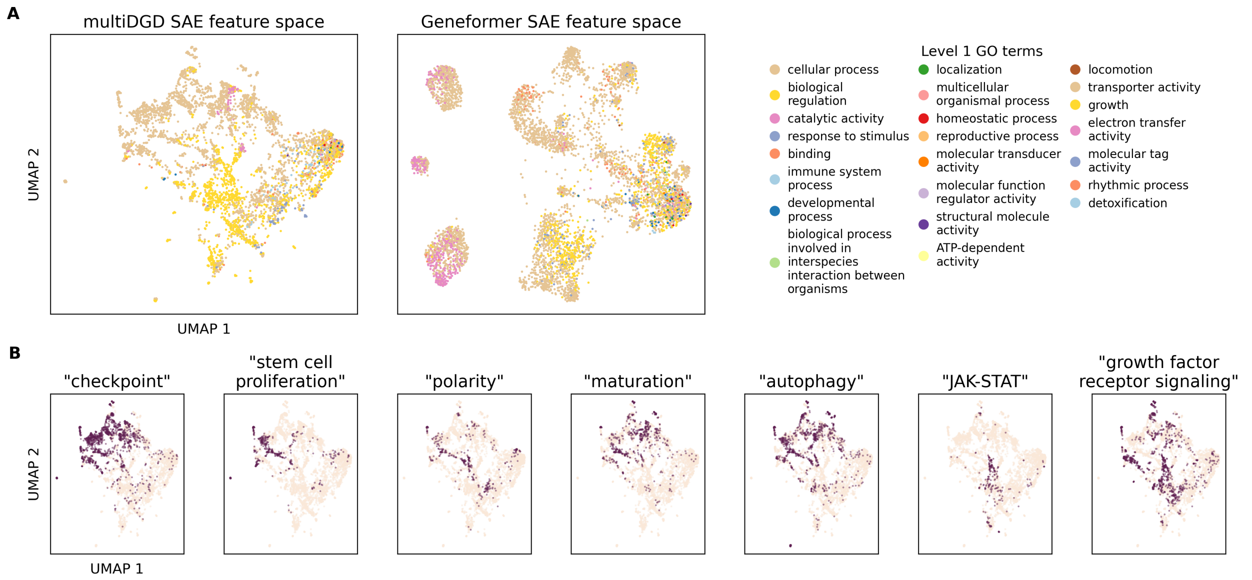

Figure 3: **SAE feature space.** Feature spaces are visualized as UMAPs of the GO-feature matrices described in Appendix A.3.7. Locations of the manually analyzed features are shown in Figure S24. **A** Feature maps of SAEs from multiDGD (left) and Geneformer (right) representations are colored by the most common 1st level of GO terms associated with the feature (legend on the right). **B** Probing multiDGD SAE features by broad semantic concepts (plot titles) included in GO terms. Dark points indicate features with at least one GO term containing the concept.

automated analysis in this work is performed both on the previously introduced SAE features trained on multiDGD representations of human bonemarrow data and an equivalent SAE trained on Geneformer [32] representations from the same data.

Experimental set up: The basis of the automated analysis is a concept-by-gene matrix summarizing the gene sets associated with each concept. “high-vs-low” sample sets were created for each active feature with the 99th percentile of the feature activations as the “high” set and a sample of maximum 1000 cells from those with zero values as “low”. This was followed by DGE analysis on the predicted expression counts and a simple GO term analysis inspired by Mi et al. [52] (details in Appendix A.3.7). The analysis is parallelized over GO terms for efficiency with a compute time of ~ 30 seconds per feature on the used Hardware. GO terms with p-values below 0.01 were recorded for each feature. Feature spaces are visualized as UMAPs [53] of the binary matrix of the matches between unique GO terms and features (distance 1.0, 10 neighbors, seed 0, spread 10).

Results: The analysis on multiDGD’s SAE returned GO terms for 4374 (82.25 %) of the active features, with overall 1875 unique biological process and 624 molecular function GO terms. Individual GO terms appeared between once and over 2500 times. Terms that appeared very often are broader, high-level GO terms associated with immune response and signaling pathways (Table S7). Many of the features active in a small fraction of cells did not cluster with cell types and would go completely unnoticed in traditional analysis of the dense latent space, making this pipeline very valuable. Figures 3A and S22 show the SAE features’ concept space. This space organizes features with respect to GO terms, largely separating into cellular processes and biological regulation at the highest level of the Gene Ontology. It can be probed for specific biological components and concepts, which is demonstrated in Figures 3B and S25. Examples of meaningful overlaps include a large overlap of features associated with protein localization and checkpoint signaling. Within this area there are processes that collectively contribute to stem cell homeostasis, fate determination and maintenance, such as stem cell proliferation, cell polarity, maturation, and autophagy.

The JAK-STAT signaling pathway takes a central role in this feature space. It appears at the intersection of features annotated with concepts from growth factor signaling, NK cell activation, antiviral response, and death - to name a few key functions.
The SAE trained on Geneformer representations resulted in 7073 active features ($33\,\%$ more than the SAE trained on multiDGD) out of which 5290 were annotated with GO terms. Interestingly, there are only 409 local features. This potential lack of local separability may be due to the curse of dimensionality (Geneformer has a latent dimensionality of 896 vs multiDGD's 20) and the complex latent distribution of multiDGD. See visualizations of embedding and more feature space plots in Figures S26-30. Feature spaces are difficult to compare. Visually, many observations made previously in terms of overlapping concepts seem to be consistent, although more specific GO terms do not cluster well (Figure S31). Additionally, all 2499 unique GO terms identified in the multiDGD SAE were also recovered from the Geneformer embeddings, with 97 additional GO terms found in this larger, pre-trained model. The most common GO terms center less around immunity, which is not very suprising given that Geneformer was trained on a large and more varied dataset (Table S7). Concept count distributions between the two models' SAEs varied with Spearman and Pearson correlations of 0.46 and 0.43, respectively (Figures S29,32). multiDGD's SAE shows an average of 95.5 ± 1.7 SEM GO terms per active feature with a range from $1 - 482$. The SAE trained on Geneformer embeddings has a lower range of GO terms per feature from $1 - 254$ with an average of 48.7 ± 1.0 SEM. A more fine-grained analysis of the feature spaces through optimal bipartite matching based on the shared GO terms reveals a low similarity of 0.16 (Appendix A.3.8 for methodology and interpretation). These results suggest that there may be a shared broad semantic structure that is learned by different models on the same kind of data, but individual features seem to potentially serve very different purposes and the functional focus of the embeddings are largely influenced by the scope of data the model was trained on.

6 Conclusion

This work explored the potential of sparse autoencoders (SAEs) to interpret latent representations in biological tabular data. Through data simulation with ground-truth generative variables, it provided valuable insights into the behavior and capabilities of SAE architectures. SAEs were found to effectively recover hidden variables if they have been learned in superposition, with performance improving as hidden dimensionality and model width increase. The presence of hidden variables in superposition depends, however, on their position in the data generation process, the impact they have on the observables, and likely also their type of distribution. Variables with an indirect effect on the observed data and little structure in the generative process could practically not be recovered. SAEs further do not pose an advantage in the recovery of known or hypothesized features compared to simple baselines. However, the connectivity of SAE features and observables can unearth valuable insight into the data generation structure.
Despite their limitations, the application of SAEs to single-cell expression models demonstrated that they present practical value in a real-world biological context. Identifying and steering features manually uncovered specific biological processes, validating the relevance of the SAE-derived features. Local features helped identify small cell type subpopulations previously not distinguishable in the latent representations. The automated annotation pipeline employs well-established methods such as DGE and enrichment analysis. Its novelty and utility stem from direct integration with the disentangled SAE features extracted from scRNAseq embeddings. This provides a novel, powerful, and scalable framework for improved interpretability. It is available as a tool on GitHub. While this case study was limited to Gene Ontology (GO) terms which are incomplete and biased towards well-studied genes, the improvement in interpretability is immense and can have a significant impact on single-cell analysis. Additionally, the pipeline can be applied to any gene expression embedding and can be used with different databases providing semantic context from gene sets. Altogether, this work presents an important step towards more interpretable models in biology, but much more research is needed in this field. Future work could explore more metrics for evaluating the biological meaningfulness in and differences between embeddings, and methods to help overcome the limitations in recovering variables that are difficult to decompose.

References

[1] Cynthia Rudin. Stop explaining black box machine learning models for high stakes decisions and use interpretable models instead. *Nat Mach Intell*, 1(5):206–215, May 2019. ISSN 2522-5839. doi: 10.1038/s42256-019-0048-x. Number: 5 Publisher: Nature Publishing Group.

[2] Tilman Räuker, Anson Ho, Stephen Casper, and Dylan Hadfield-Menell. Toward Transparent AI: A Survey on Interpreting the Inner Structures of Deep Neural Networks, August 2023. arXiv:2207.13243 [cs].

[3] Adityanarayanan Radhakrishnan, Mikhail Belkin, and Caroline Uhler. Wide and deep neural networks achieve consistency for classification. *Proceedings of the National Academy of Sciences*, 120(14): e2208779120, April 2023. doi: 10.1073/pnas.2208779120. Publisher: Proceedings of the National Academy of Sciences.

[4] Ričards Marcinkevičs and Julia E. Vogt. Interpretability and Explainability: A Machine Learning Zoo Mini-tour, March 2023. arXiv:2012.01805 [cs].

[5] Cynthia Rudin, Chaofan Chen, Zhi Chen, Haiyang Huang, Lesia Semenova, and Chudi Zhong. Interpretable machine learning: Fundamental principles and 10 grand challenges. *Statistics Surveys*, 16 (none):1–85, January 2022. ISSN 1935-7516. doi: 10.1214/21-SS133. Publisher: Amer. Statist. Assoc., the Bernoulli Soc., the Inst. Math. Statist., and the Statist. Soc. Canada.

[6] Nelson Elhage, Tristan Hume, Catherine Olsson, Nicholas Schiefer, Tom Henighan, Shauna Kravec, Zac Hatfield-Dodds, Robert Lasenby, Dawn Drain, Carol Chen, Roger Grosse, Sam McCandlish, Jared Kaplan, Dario Amodei, Martin Wattenberg, and Christopher Olah. Toy Models of Superposition, September 2022. arXiv:2209.10652 [cs].

[7] Bruno A. Olshausen and David J. Field. Sparse coding with an overcomplete basis set: A strategy employed by V1? *Vision Research*, 37(23):3311–3325, December 1997. ISSN 0042-6989. doi: 10.1016/S0042-6989(97)00169-7.

[8] Lee Sharkey, Dan Braun, and beren. [Interim research report] Taking features out of superposition with sparse autoencoders. December 2022.

[9] Trenton Bricken, Adly Templeton, Joshua Batson, Brian Chen, Adam Jermyn, Tom Conerly, Nick Turner, Cem Anil, Carson Denison, Amanda Askell, Robert Lasenby, Yifan Wu, Shauna Kravec, Nicholas Schiefer, Tim Maxwell, Nicholas Joseph, Zac Hatfield-Dodds, Alex Tamkin, Karina Nguyen, Brayden McLean, Josiah E Burke, Tristan Hume, Shan Carter, Tom Henighan, and Christopher Olah. Towards Monosemanticity: Decomposing Language Models With Dictionary Learning. *Transformer Circuits Thread*, 2023.

[10] Robert Huben, Hoagy Cunningham, Logan Riggs Smith, Aidan Ewart, and Lee Sharkey. Sparse Autoencoders Find Highly Interpretable Features in Language Models. October 2023.

[11] 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, June 2024.

[12] Joeky T. Senders, Patrick C. Staples, Aditya V. Karhade, Mark M. Zaki, William B. Gormley, Marike L. D. Broekman, Timothy R. Smith, and Omar Arnaout. Machine Learning and Neurosurgical Outcome Prediction: A Systematic Review. *World Neurosurgery*, 109:476–486.e1, January 2018. ISSN 1878-8750. doi: 10.1016/j.wneu.2017.09.149.

[13] Emilly M. Lima, Antônio H. Ribeiro, Gabriela M. M. Paixão, Manoel Horta Ribeiro, Marcelo M. Pinto-Filho, Paulo R. Gomes, Derick M. Oliveira, Ester C. Sabino, Bruce B. Duncan, Luana Giatti, Sandhi M. Barreto, Wagner Meira Jr, Thomas B. Schön, and Antonio Luiz P. Ribeiro. Deep neural network-estimated electrocardiographic age as a mortality predictor. *Nat Commun*, 12(1):5117, August 2021. ISSN 2041-1723. doi: 10.1038/s41467-021-25351-7. Number: 1 Publisher: Nature Publishing Group.

[14] Xinyi Zhang, Xiao Wang, G. V. Shivashankar, and Caroline Uhler. Graph-based autoencoder integrates spatial transcriptomics with chromatin images and identifies joint biomarkers for Alzheimer's disease. *Nat Commun*, 13(1):7480, December 2022. ISSN 2041-1723. doi: 10.1038/s41467-022-35233-1. Number: 1 Publisher: Nature Publishing Group.

[15] Chiara Corti, Marisa Cobanaj, Edward C. Dee, Carmen Criscitiello, Sara M. Tolaney, Leo A. Celi, and Giuseppe Curigliano. Artificial intelligence in cancer research and precision medicine: Applications, limitations and priorities to drive transformation in the delivery of equitable and unbiased care. *Cancer Treatment Reviews*, 112:102498, January 2023. ISSN 0305-7372. doi: 10.1016/j.ctrv.2022.102498.

[16] Frank W. Pun, Ivan V. Ozerov, and Alex Zhavoronkov. AI-powered therapeutic target discovery. *Trends in Pharmacological Sciences*, 44(9):561–572, September 2023. ISSN 0165-6147. doi: 10.1016/j.tips.2023.06.010. Publisher: Elsevier.

[17] Theogene Habineza, Antônio H. Ribeiro, Daniel Gedon, Joachim A. Behar, Antonio Luiz P. Ribeiro, and Thomas B. Schön. End-to-end risk prediction of atrial fibrillation from the 12-Lead ECG by deep neural networks. *Journal of Electrocardiology*, 81:193–200, November 2023. ISSN 0022-0736. doi: 10.1016/j.jelectrocard.2023.09.011.

[18] Peter V. Kharchenko. The triumphs and limitations of computational methods for scRNA-seq. *Nat Methods*, 18(7):723–732, July 2021. ISSN 1548-7105. doi: 10.1038/s41592-021-01171-x. Number: 7 Publisher: Nature Publishing Group.

[19] David Lähnemann, Johannes Köster, Ewa Szczurek, Davis J. McCarthy, Stephanie C. Hicks, Mark D. Robinson, Catalina A. Vallejos, Kieran R. Campbell, Niko Beerenwinkel, Ahmed Mahfouz, Luca Pinello, Pavel Skums, Alexandros Stamatakis, Camille Stephan-Otto Attolini, Samuel Aparicio, Jasmijn Baaijens, Marleen Balvert, Buys de Barbanson, Antonio Cappuccio, Giacomo Corleone, Bas E. Dutilh, Maria Florescu, Victor Guryev, Rens Holmer, Katharina Jahn, Thamar Jessurun Lobo, Emma M. Keizer, Indu Khatri, Szymon M. Kielbasa, Jan O. Korbel, Alexey M. Kozlov, Tzu-Hao Kuo, Boudewijn P.F. Lelieveldt, Ion I. Mandoiu, John C. Marioni, Tobias Marschall, Felix Mölder, Amir Niknejad, Lukasz Raczkowski, Marcel Reinders, Jeroen de Ridder, Antoine-Emmanuel Saliba, Antonios Somarakis, Oliver Stegle, Fabian J. Theis, Huan Yang, Alex Zelikovsky, Alice C. McHardy, Benjamin J. Raphael, Sohrab P. Shah, and Alexander Schönhuth. Eleven grand challenges in single-cell data science. *Genome Biology*, 21(1):31, February 2020. ISSN 1474-760X. doi: 10.1186/s13059-020-1926-6.

[20] Lukas Heumos, Anna C. Schaar, Christopher Lance, Anastasia Litinetskaya, Felix Drost, Luke Zappia, Malte D. Lücken, Daniel C. Strobl, Juan Henao, Fabiola Curion, Herbert B. Schiller, and Fabian J. Theis. Best practices for single-cell analysis across modalities. *Nat Rev Genet*, 24(8):550–572, August 2023. ISSN 1471-0064. doi: 10.1038/s41576-023-00586-w. Number: 8 Publisher: Nature Publishing Group.

[21] Ricard Argelaguet, Anna S. E. Cuomo, Oliver Stegle, and John C. Marioni. Computational principles and challenges in single-cell data integration. *Nat Biotechnol*, 39(10):1202–1215, October 2021. ISSN 1546-1696. doi: 10.1038/s41587-021-00895-7. Number: 10 Publisher: Nature Publishing Group.

[22] Chenling Xu, Romain Lopez, Edouard Mehlman, Jeffrey Regier, Michael I Jordan, and Nir Yosef. Probabilistic harmonization and annotation of single-cell transcriptomics data with deep generative models. *Molecular Systems Biology*, 17(1):e9620, January 2021. ISSN 1744-4292. doi: 10.15252/msb.20209620. Publisher: John Wiley & Sons, Ltd.

[23] Romain Lopez, Jeffrey Regier, Michael B. Cole, Michael I. Jordan, and Nir Yosef. Deep generative modeling for single-cell transcriptomics. *Nat Methods*, 15(12):1053–1058, December 2018. ISSN 1548-7105. doi: 10.1038/s41592-018-0229-2. Number: 12 Publisher: Nature Publishing Group.

[24] Tal Ashuach, Mariano I. Gabitto, Rohan V. Koodli, Giuseppe-Antonio Saldi, Michael I. Jordan, and Nir Yosef. MultiVI: deep generative model for the integration of multimodal data. *Nat Methods*, pages

1–10, June 2023. ISSN 1548-7105. doi: 10.1038/s41592-023-01909-9. Publisher: Nature Publishing Group.

[25] Yingxin Lin, Tung-Yu Wu, Sheng Wan, Jean Y. H. Yang, Wing H. Wong, and Y. X. Rachel Wang. scJoint integrates atlas-scale single-cell RNA-seq and ATAC-seq data with transfer learning. *Nat Biotechnol*, 40(5):703–710, May 2022. ISSN 1546-1696. doi: 10.1038/s41587-021-01161-6. Number: 5 Publisher: Nature Publishing Group.

[26] Stefan G Stark, Joanna Ficek, Francesco Locatello, Ximena Bonilla, Stéphane Chevrier, Franziska Singer, Tumor Profiler Consortium, Gunnar Rätsch, and Kjong-Van Lehmann. SCIM: universal single-cell matching with unpaired feature sets. *Bioinformatics*, 36(Supplement_2):i919–i927, December 2020. ISSN 1367-4803. doi: 10.1093/bioinformatics/btaa843.

[27] Karren Dai Yang, Anastasiya Belyaeva, Saradha Venkatachalapathy, Karthik Damodaran, Abigail Katcoff, Adityanarayanan Radhakrishnan, G. V. Shivashankar, and Caroline Uhler. Multi-domain translation between single-cell imaging and sequencing data using autoencoders. *Nat Commun*, 12(1): 31, January 2021. ISSN 2041-1723. doi: 10.1038/s41467-020-20249-2. Number: 1 Publisher: Nature Publishing Group.

[28] Chunman Zuo and Luonan Chen. Deep-joint-learning analysis model of single cell transcriptome and open chromatin accessibility data. *Briefings in Bioinformatics*, 22(4):bbaa287, July 2021. ISSN 1477-4054. doi: 10.1093/bib/bbaa287.

[29] Chunman Zuo, Hao Dai, and Luonan Chen. Deep cross-omics cycle attention model for joint analysis of single-cell multi-omics data. *Bioinformatics*, 37(22):4091–4099, November 2021. ISSN 1367-4803. doi: 10.1093/bioinformatics/btab403.

[30] Kodai Minoura, Ko Abe, Hyunha Nam, Hiroyoshi Nishikawa, and Teppei Shimamura. A mixture-of-experts deep generative model for integrated analysis of single-cell multiomics data. *Cell Reports Methods*, 1(5):100071, September 2021. ISSN 2667-2375. doi: 10.1016/j.crmeth.2021.100071.

[31] Viktoria Schuster, Emma Dann, Anders Krogh, and Sarah A. Teichmann. multiDGD: A versatile deep generative model for multi-omics data, August 2023. Pages: 2023.08.23.554420 Section: New Results.

[32] Christina V. Theodoris, Ling Xiao, Anant Chopra, Mark D. Chaffin, Zeina R. Al Sayed, Matthew C. Hill, Helene Mantineo, Elizabeth M. Brydon, Zexian Zeng, X. Shirley Liu, and Patrick T. Ellinor. Transfer learning enables predictions in network biology. *Nature*, 618(7965):616–624, June 2023. ISSN 1476-4687. doi: 10.1038/s41586-023-06139-9. Publisher: Nature Publishing Group.

[33] Zeyu Yun, Yubei Chen, Bruno A. Olshausen, and Yann LeCun. Transformer visualization via dictionary learning: contextualized embedding as a linear superposition of transformer factors, March 2021.

[34] David Bau, Bolei Zhou, Aditya Khosla, Aude Oliva, and Antonio Torralba. Network Dissection: Quantifying Interpretability of Deep Visual Representations. pages 6541–6549, 2017.

[35] Elana Simon and James Zou. InterPLM: Discovering Interpretable Features in Protein Language Models via Sparse Autoencoders, November 2024. Pages: 2024.11.14.623630 Section: New Results.

[36] Etowah Adams, Liam Bai, Minji Lee, Yiyang Yu, and Mohammed AlQuraishi. From Mechanistic Interpretability to Mechanistic Biology: Training, Evaluating, and Interpreting Sparse Autoencoders on Protein Language Models, February 2025. URL `https://www.biorxiv.org/content/10.1101/2025.02.06.636901v1`. Pages: 2025.02.06.636901 Section: New Results.

[37] Mona Rams and Tim O.F. Conrad. Dictionary learning allows model-free pseudotime estimation of transcriptomic data. *BMC Genomics*, 23(1):56, January 2022. ISSN 1471-2164. doi: 10.1186/s12864-021-08276-9.

[38] Romain Lopez, Natasa Tagasovska, Stephen Ra, Kyunghyun Cho, Jonathan Pritchard, and Aviv Regev. Learning Causal Representations of Single Cells via Sparse Mechanism Shift Modeling. In *Proceedings*

of the Second Conference on Causal Learning and Reasoning, pages 662–691. PMLR, August 2023. ISSN: 2640-3498.

[39] Yuhan Hao, Tim Stuart, Madeline H. Kowalski, Saket Choudhary, Paul Hoffman, Austin Hartman, Avi Srivastava, Gesmira Molla, Shaista Madad, Carlos Fernandez-Granda, and Rahul Satija. Dictionary learning for integrative, multimodal and scalable single-cell analysis. *Nat Biotechnol*, 42(2):293–304, February 2024. ISSN 1546-1696. doi: 10.1038/s41587-023-01767-y. Publisher: Nature Publishing Group.

[40] Ioulia Karagiannaki, Krystallia Gourlia, Vincenzo Lagani, Yannis Pantazis, and Ioannis Tsamardinos. Learning biologically-interpretable latent representations for gene expression data. *Mach Learn*, 112 (11):4257–4287, November 2023. ISSN 1573-0565. doi: 10.1007/s10994-022-06158-z.

[41] Tiantian Guo, Yang Chen, Minglei Shi, Xiangyu Li, and Michael Q Zhang. Integration of single cell data by disentangled representation learning. *Nucleic Acids Research*, 50(2):e8, January 2022. ISSN 0305-1048. doi: 10.1093/nar/gkab978.

[42] Zoe Piran, Niv Cohen, Yedid Hoshen, and Mor Nitzan. Disentanglement of single-cell data with biolord. *Nat Biotechnol*, pages 1–6, January 2024. ISSN 1546-1696. doi: 10.1038/s41587-023-02079-x. Publisher: Nature Publishing Group.

[43] Yoshua Bengio, Aaron Courville, and Pascal Vincent. Representation Learning: A Review and New Perspectives. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 35(8):1798–1828, August 2013. ISSN 1939-3539. doi: 10.1109/TPAMI.2013.50. Conference Name: IEEE Transactions on Pattern Analysis and Machine Intelligence.

[44] Alireza Makhzani and Brendan Frey. k-Sparse Autoencoders, March 2014. arXiv:1312.5663 [cs].

[45] Michael Ashburner, Catherine A. Ball, Judith A. Blake, David Botstein, Heather Butler, J. Michael Cherry, Allan P. Davis, Kara Dolinski, Selina S. Dwight, Janan T. Eppig, Midori A. Harris, David P. Hill, Laurie Issel-Tarver, Andrew Kasarskis, Suzanna Lewis, John C. Matese, Joel E. Richardson, Martin Ringwald, Gerald M. Rubin, and Gavin Sherlock. Gene Ontology: tool for the unification of biology. *Nat Genet*, 25(1):25–29, May 2000. ISSN 1546-1718. doi: 10.1038/75556. Publisher: Nature Publishing Group.

[46] The Gene Ontology Consortium, Suzi A Aleksander, James Balhoff, Seth Carbon, J Michael Cherry, Harold J Drabkin, Dustin Ebert, Marc Feuermann, Pascale Gaudet, Nomi L Harris, David P Hill, Raymond Lee, Huaiyu Mi, Sierra Moxon, Christopher J Mungall, Anushya Muruganugan, Tremayne Mushayahama, Paul W Sternberg, Paul D Thomas, Kimberly Van Auken, Jolene Ramsey, Deborah A Siegele, Rex L Chisholm, Petra Fey, Maria Cristina Aspromonte, Maria Victoria Nugnes, Federica Quaglia, Silvio Tosatto, Michelle Giglio, Suvarna Nadendla, Giulia Antonazzo, Helen Attrill, Gil dos Santos, Steven Marygold, Victor Strelets, Christopher J Tabone, Jim Thurmond, Pinglei Zhou, Saadullah H Ahmed, Praoparn Asanitthong, Diana Luna Buitrago, Meltem N Erdol, Matthew C Gage, Mohamed Ali Kadhum, Kan Yan Chloe Li, Miao Long, Aleksandra Michalak, Angeline Pesala, Armalya Pritazahra, Shirin C C Saverimuttu, Renzhi Su, Kate E Thurlow, Ruth C Lovering, Colin Logie, Snezhana Oliferenko, Judith Blake, Karen Christie, Lori Corbani, Mary E Dolan, Harold J Drabkin, David P Hill, Li Ni, Dmitry Sitnikov, Cynthia Smith, Alayne Cuzick, James Seager, Laurel Cooper, Justin Elser, Pankaj Jaiswal, Parul Gupta, Pankaj Jaiswal, Sushma Naithani, Manuel Lera-Ramirez, Kim Rutherford, Valerie Wood, Jeffrey L De Pons, Melinda R Dwinell, G Thomas Hayman, Mary L Kaldunski, Anne E Kwitek, Stanley J F Laulederkind, Marek A Tutaj, Mahima Vedi, Shur-Jen Wang, Peter D'Eustachio, Lucila Aimo, Kristian Axelsen, Alan Bridge, Nevila Hyka-Nouspikel, Anne Morgat, Suzi A Aleksander, J Michael Cherry, Stacia R Engel, Kalpana Karra, Stuart R Miyasato, Robert S Nash, Marek S Skrzypek, Shuai Weng, Edith D Wong, Erika Bakker, Tanya Z Berardini, Leonore Reiser, Andrea Auchincloss, Kristian Axelsen, Ghislaine Argoud-Puy, Marie-Claude Blatter, Emmanuel Boutet, Lionel Breuza, Alan Bridge, Cristina Casals-Casas, Elisabeth Coudert, Anne Estreicher, Maria Livia Famiglietti, Marc Feuermann, Arnaud Gos, Nadine Gruaz-Gumowski, Chantal Hulo, Nevila

Hyka-Nouspikel, Florence Jungo, Philippe Le Mercier, Damien Lieberherr, Patrick Masson, Anne Morgat, Ivo Pedruzzi, Lucille Pourcel, Sylvain Poux, Catherine Rivoire, Shyamala Sundaram, Alex Bateman, Emily Bowler-Barnett, Hema Bye-A-Jee, Paul Denny, Alexandr Ignatchenko, Rizwan Ishtiaq, Antonia Lock, Yvonne Lussi, Michele Magrane, Maria J Martin, Sandra Orchard, Pedro Raposo, Elena Speretta, Nidhi Tyagi, Kate Warner, Rossana Zaru, Alexander D Diehl, Raymond Lee, Juancarlos Chan, Stavros Diamantakis, Daniela Raciti, Magdalena Zarowiecki, Malcolm Fisher, Christina James-Zorn, Virgilio Ponferrada, Aaron Zorn, Sridhar Ramachandran, Leyla Ruzicka, and Monte Westerfield. The Gene Ontology knowledgebase in 2023. *Genetics*, 224(1):iyad031, May 2023. ISSN 1943-2631. doi: 10.1093/genetics/iyad031.

[47] Malte Luecken, Daniel Burkhardt, Robrecht Cannoodt, Christopher Lance, Aditi Agrawal, Hananeh Aliee, Ann Chen, Louise Deconinck, Angela Detweiler, Alejandro Granados, Shelly Huynh, Laura Isacco, Yang Kim, Dominik Klein, BONY DE KUMAR, Sunil Kuppasani, Heiko Lickert, Aaron McGeever, Joaquin Melgarejo, Honey Mekonen, Maurizio Morri, Michaela Müller, Norma Neff, Sheryl Paul, Bastian Rieck, Kaylie Schneider, Scott Steelman, Michael Sterr, Daniel Treacy, Alexander Tong, Alexandra-Chloe Villani, Guilin Wang, Jia Yan, Ce Zhang, Angela Pisco, Smita Krishnaswamy, Fabian Theis, and Jonathan M Bloom. A sandbox for prediction and integration of dna, rna, and proteins in single cells. In J. Vanschoren and S. Yeung, editors, *Proceedings of the Neural Information Processing Systems Track on Datasets and Benchmarks*, volume 1, 2021.

[48] Ricard Argelaguet, Tim Lohoff, Jingyu Gavin Li, Asif Nakhuda, Deborah Drage, Felix Krueger, Lars Velten, Stephen J. Clark, and Wolf Reik. Decoding gene regulation in the mouse embryo using single-cell multi-omics, November 2022. URL `https://www.biorxiv.org/content/10.1101/2022.06.15.496239v2`. Pages: 2022.06.15.496239 Section: New Results.

[49] Alexandro E. Trevino, Fabian Müller, Jimena Andersen, Laksshman Sundaram, Arwa Kathiria, Anna Shcherbina, Kyle Farh, Howard Y. Chang, Anca M. Pașca, Anshul Kundaje, Sergiu P. Pașca, and William J. Greenleaf. Chromatin and gene-regulatory dynamics of the developing human cerebral cortex at single-cell resolution. *Cell*, 184(19):5053–5069.e23, September 2021. ISSN 0092-8674, 1097-4172. doi: 10.1016/j.cell.2021.07.039. URL `https://www.cell.com/cell/abstract/S0092-8674(21)00942-9`. Publisher: Elsevier.

[50] Ryan D. Schubert, Yang Hu, Gaurav Kumar, Spencer Szeto, Peter Abraham, Johannes Winderl, Joel M. Guthridge, Gabriel Pardo, Jeffrey Dunn, Lawrence Steinman, and Robert C. Axtell. Interferon-beta treatment requires B cells for efficacy in neuro-autoimmunity. *Journal of immunology (Baltimore, Md. : 1950)*, 194(5):2110–2116, March 2015. ISSN 0022-1767. doi: 10.4049/jimmunol.1402029.

[51] Dongwei Zhu, Yue Zhang, and Shengjun Wang. Histone citrullination: a new target for tumors. *Molecular Cancer*, 20(1):90, June 2021. ISSN 1476-4598. doi: 10.1186/s12943-021-01373-z.

[52] Huaiyu Mi, Anushya Muruganujan, John T. Casagrande, and Paul D. Thomas. Large-scale gene function analysis with PANTHER Classification System. *Nature protocols*, 8(8):1551–1566, August 2013. ISSN 1754-2189. doi: 10.1038/nprot.2013.092.

[53] Leland McInnes, John Healy, and James Melville. UMAP: Uniform Manifold Approximation and Projection for Dimension Reduction, September 2020. arXiv:1802.03426 [stat].

[54] Paulo Amaral, Silvia Carbonell-Sala, Francisco M. De La Vega, Tiago Faial, Adam Frankish, Thomas Gingeras, Roderic Guigo, Jennifer L. Harrow, Artemis G. Hatzigeorgiou, Rory Johnson, Terence D. Murphy, Mihaela Pertea, Kim D. Pruitt, Shashikant Pujar, Hazuki Takahashi, Igor Ulitsky, Ales Varabyou, Christine A. Wells, Mark Yandell, Piero Carninci, and Steven L. Salzberg. The status of the human gene catalogue. *Nature*, 622(7981):41–47, October 2023. ISSN 1476-4687. doi: 10.1038/s41586-023-06490-x. Publisher: Nature Publishing Group.

[55] Diederik P. Kingma and Jimmy Ba. Adam: A method for stochastic optimization, 2014. cite arxiv:1412.6980. Published as a conference paper at the 3rd International Conference for Learning Representations, San Diego, 2015.

[56] Takuya Akiba, Shotaro Sano, Toshihiko Yanase, Takeru Ohta, and Masanori Koyama. Optuna: A next-generation hyperparameter optimization framework. In *Proceedings of the 25th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, 2019.

[57] Simon Anders and Wolfgang Huber. Differential expression analysis for sequence count data. *Genome Biology*, 11(10):R106, October 2010. ISSN 1474-760X. doi: 10.1186/gb-2010-11-10-r106.

[58] Michael I. Love, Wolfgang Huber, and Simon Anders. Moderated estimation of fold change and dispersion for RNA-seq data with DESeq2. *Genome Biol*, 15(12):550, December 2014. ISSN 1474-760X. doi: 10.1186/s13059-014-0550-8.

[59] Yoav Benjamini and Yosef Hochberg. Controlling the false discovery rate: a practical and powerful approach to multiple testing. *J. Roy. Statist. Soc. Ser. B*, 57(1):289–300, 1995. ISSN 0035-9246.

Code availability

Code is made available for reproducibility in this GitHub repository. The presented tool **scFeatureLens** is available in this GitHub collection.

Data availability

All data and models used in this work are publicly available and cited.

Conflict of Interest

I declare no conflict of interest.

Acknowledgements

I would like to acknowledge great discussions with coworkers at the Center of Health Data Science (University of Copenhagen) and the Eric and Wendy Schmidt Center at the Broad Institute. I thank the reviewers for their time and effort to help improve the quality and communication of the work. I especially want to thank my mentor Anders Krogh for his tremendous support. I also want to thank Uthsav Chitra and Kristoffer Stensbo-Smidt for their feedback and advice, and Jonas Sindlinger for being a wonderful rubber duck.

Impact Statement

This paper presents work whose goal it is to advance the development and application of mechanistic interpretability for the fields of biology and medicine. There are many potential positive impacts for society related to improving disease understanding and treatment. A potential negative impact with interpretability of biological models is the exploitation of knowledge about differences related to gender, ethnicity, socioeconomic background, and genetics. I believe, however, that the open source development of interpretability techniques will lead to both discovery and removal of such biases in biological models.