Appendix A: Methods

A.1 Compute infrastructure

All computations were performed using `Python 3.9` on either CPU or one of the following GPUs: NVIDIA A30, NVIDIA RTX A5000.

A.2 Simulated Data

A.2.1 Data Simulation

Simulated data sets were designed with inspiration from sparse count data as we see in single-cell sequencing in order to get an understanding of what SAEs learn about the data structure and hidden variables. First, a set of hypotheses is defined to guide our the generation process:

- Gene regulation is determined by molecular regulators and gene programs, and thus the observed data $\mathcal{Y}$ should lie on a lower-dimensional manifold $\mathcal{X}$.
- Different cell types L have different patterns of active regulators/programs and different levels of overall expression $\mathcal{Y}$.
- Technical noise or other covariates $\mathcal{B}$ can cause shifts in $\mathcal{Y}$.

Simulated counts $\mathbf{y} \in \mathcal{Y} = \{Y_{i=1}, ..., Y_{i=N}\}^{\mathrm{T}}$ are generated through the following three steps:

$$\begin{aligned} \mathbf{x}' &= \mathbf{x} + \mathbf{b_c} \text{ with } \mathbf{x} \sim \mathcal{X} \\ \mathbf{x}'' &= \mathbf{x}' \, \mathbf{a_c} \\ \mathbf{y} &= \sum_{j=1}^{100} m_j \, \mathbf{x_j''} \end{aligned} \tag{4}$$

with $\mathcal{X} = (X_1, ..., X_{100})^{\mathrm{T}}$ presenting the ground truth multivariate latent variables. Noise vectors $\mathbf{b_c} = \mathcal{B}\mathbf{s_{c1}}$ and cell type activity vectors $\mathbf{a_c} = \mathbf{A}^{\mathrm{T}}\mathbf{s_{c2}}$ are products of one-hot selection column vectors $\mathbf{s_c}$ with noise distribution $\mathcal{B} = (B_1, ..., B_3)^{\mathrm{T}}$ and activity matrix $\mathbf{A} = (\mathrm{a}_{lj}) \in \mathbb{N}_0^{40\times 100}$, respectively. Matrix $\mathbf{M} = (\mathrm{m}_{ij}) \in \mathbb{N}_0^{N\times 100}$ presents the connectivity matrix between regulators/programs and genes. Random variables were sampled according to

$$\begin{aligned} X_j &\sim \mathrm{Pois}(\lambda = 1.1j) \\ B_g &\sim \mathcal{N}(\mu = j, \sigma = 0.1) \\ \mathbf{s_{c1}} &\sim \mathrm{Cat}(p = \frac{1}{3}, k = 3) \\ \mathrm{a}_{lj} &\sim \mathrm{Bin}(k = 1, p = 0.3) \\ \mathbf{s_{c2}} &\sim \mathrm{Cat}(p = \frac{1}{40}, k = 40) \\ \mathrm{m}_{ij} &\sim \mathrm{Bin}(k = 1, p = 0.1). \end{aligned} \tag{5}$$

For a "large" simulation with realistic dimensions, a data dimensionality of $N = 20000$ was chosen which is at the upper limit of the number of protein-coding genes in the human genome [54]. The latent dimensionality of $\mathcal{X}$ was set to 100. $L = 40$ dimensions for $\mathbf{A}$ represent different cell types and $G = 3$ variables in $\mathcal{B}$ simulate technical noise. Distribution parameters and the order of the generative process were chosen so that the simulated data $\mathcal{Y}$ would present similar structures and count distributions compared to real data (Supplementary Figure 1). 90000 train and 10000 validation data points were sampled. For simplicity, all of the variables of interest will be referred to as Y ($\{\mathcal{Y}, \mathbf{y}\}$), X ($\{\mathcal{X}, \mathbf{x}\}$), X' ($\mathbf{x}'$), X'' ($\mathbf{x}''$), A ($\mathbf{a_c}$), B ($\mathbf{b_c}$). Additionally, a "small" simulation set was created for a large-scale SAE sweep and the possibility to visually

verify superpositions. It features $|Y| = 5$, $|X| = 3$, $L = 1$, and no noise. Details can be found in Appendix A.2.1.

The small simulation data set with $|Y| = 5$ and $|X| = 3$ was generated in two steps. First, the three-dimensional multivariate random variable X were sampled from Binomial distributions with probabilities $[0.5, 0.1, 0.9]$ multiplied with samples from Poisson variable A ($\lambda = 2$), resulting in latent variables X'. Secondly, X' was multiplied with $\mathbf{M}$ ($p = 0.1$) to produce observables Y. 10000 train and 2000 validation data points were sampled.

$$\begin{aligned} \mathbf{x'} &= \mathbf{x}\,\mathbf{a} \text{ with } \mathbf{x} \sim \mathcal{X} \\ \mathbf{y_i} &= \sum_{j=1}^{3} m_{i,j}\,\mathbf{x'_j} \end{aligned} \tag{6}$$

A.2.2 AE architectures and training

An autoencoder was trained that perfectly recovered latent variables X' (Supplementary Figure 3) of the small simulation data with latent dimension 4 equal to the number of generative variables, ReLU activation, Adam optimizer Kingma and Ba [55] (learning rate 10^{-4}), and MSE loss for 20000 epochs.
Autoencoder architectures for the large simulation were set up as either "narrow" or "wide" with mirrored encoder and decoder. d here is referred to as the latent dimensionality. A "narrow" encoder would be of structure $[\max(1000, 2d), \max(150, 2d), ..., \max(150, 2d)]$ unless the number of layers was only 2, in which case the hidden dimensionality would be $\max(150, 2d)$. A "wide" encoder would receive hidden dimensionalities sampled from equidistant points between the input dimension and d. Hyperparameters were determined through Optuna optimization Akiba et al. [56] based on the reconstruction loss with 50 trials and 100 epochs. The trials tested learning rates between 10^{-6} and 10^{-3}, weight decays $[0, 0.1, ..., 10^{-7}]$, dropout $[0, 0.1]$ and batch sizes between 32 and 512. Selected hyperparameters for each depth and width can be found in Table S2. Remaining parameters are shown in Table S3. All models were trained with Adam optimizer and early stopping for up to 10000 epochs.

A.2.3 Superpositions

Superpositions in latent representations were identified through linear regression. For the small simulations, superposition vectors and coefficients of determination (R^2) were computed through `sklearn`'s `LinearRegression`. For the sake of effiency on the large number of variables in the large simulation, linear regression was implemented using a single linear neural network layer trained for 100 epochs by optimizing the mean squared error with standard gradient descent optimization and a learning rate of 10^{-4}.

Given observed values y_i with mean $\bar{y}$ and predicted values $\hat{y}_i$, the coefficient of determination

$$R^2 = 1 - \frac{\sum_i (y_i - \hat{y}_i)^2}{\sum_i (y_i - \bar{y}_i)^2} \tag{7}$$

represents the fraction of variance explained by the model, with values ranging from 0 (no explanatory power) to 1 (perfect prediction).

A.2.4 SAE hyperparameter evaluation

Different SAE architectures were trained on varying hidden dimensionalities (latent size multiplied with a hidden factor), learning rates, and L1 weights for 500 epochs. All tested hyperparameters can be found in Table S4. In the case of TopK SAEs, the sparsity is controlled by k, which was tested as percentages of the hidden dimension. For each instance, the following metrics were computed:

- number of active hidden neurons (activity determined by activations of $> 10^{-10}$)

- number of redundant hidden neurons (neurons that fire with other neurons with a Pearson correlation ≥ 0.95)
- average number of neurons firing per sample
- average number of neurons corresponding to a given data variable (determined by Pearson correlation ≥ 0.95)
- highest Pearson correlation between a neuron and a given data variable

A.2.5 Structure identification

This analysis was done on one of the well-performing SAEs trained on representations from one of the best performing AEs in terms of validation loss and variable recovery. The AE featured 2 layers in the "wide" format with 5075 hidden neurons and a latent dimension of 150. The SAE featured a scaling factor of 100, an L1 weight of 0.001, and a learning rate of 10^{-5}. Cosine similarities between all 11849 active SAE features and all 20000 observables in Y were computed. Based on different percentiles of the cosine similarity matrix (as thresholds), connectivity matrices were computed between SAE features and Y and Binomial tests between all features and all variables in X'' w.r.t. Y were performed. The ground truth connectivity matrix was given by the data generation matrix $\mathbf{M}$. The best matching X'' for each feature was computed based on the maximum number of hits. The reported result is the maximum fraction of "genes" Y connected to X'' covered by the set of "genes" Y connected to the SAE features.

A.3 Single-cell case study

A.3.1 Single-cell representations

Representations were extracted from three pre-trained multiDGD models [31] trained on single-cell multi-omics data from human bone marrow [47], mouse gastrulation [48], and human brain [49]. The following table of dataset sizes was taken from Schuster et al. [31].

Supplementary Table 1: Summary of single-cell multi-omics data used.

Dataset	Species	Number of cells	RNA dimensionality	ATAC dimensionality
Bone marrow	Human	69249	13431	116490
Brain	Human	3534	15172	95677
Gastrulation	Mouse	56861	11792	69862

The same train-validation-test splits were used as in Schuster et al. [31]. The latent space of the model is small with only 20 dimensions. The paper highlighted the structure of the latent space, especially with regard to the clear trajectory of differentiation from stem cells to red blood cells (erythrocytes) [31]. The pre-trained model and data were downloaded as instructed by Schuster et al. [31]. Furthermore, embeddings for the human bone marrow data were extracted from Geneformer [32] following their instructions to extract embeddings by passing the scRNAseq data through the most recent version of Geneformer "gf-20L-95M-i4096". The extracted embeddings had a dimensionality of 896.

A.3.2 Identifying a feature for red blood cell differentiation

Red blood cell differentiation: This rule set was created to identify potential features of red blood cell differentiation:

1. The average activation must be higher in the red blood cell line than in other cell types.
2. Average activations must consistently increase from the stem cells to the final differentiation stage of red blood cells.

Applying this rule set provided 44 neurons as potential features. These neurons were inspected visually in terms of cell-wise activations and tested to see which ones would result in the largest shift in latent space towards differentiated cells when maximizing the neuron's activations in stem cells (Supplementary Figure 20). See the next section for details on perturbations. This returned neuron 2306 as the most promising candidate feature.

A.3.3 SAE training

A small hyperparameter search was performed on the multiDGD embeddings to see if the simulation results translated well to real world settings. Both Vanilla and Bricken SAEs were tested, but not TopK since this method was not robust in previous experiments and has the disadvantage of having to estimate the number of active neurons beforehand. Hyperparameters tested were hidden scaling factors $[20, 100, 200, 500]$, $L1$ weights $[1, 0.1, 0.01, 10^{-3}, 10^{-4}]$, and learning rates $[10^{-4}, 10^{-5}]$ with a batch size of 128 for 1000 epochs with early stopping (patience 50).

A learning rate of 10^{-4} gave best results and was most robust, which aligns with simulation results. When training long enough, reconstruction loss generally decreased with the scaling factor. Learning rate 10^{-4} presented the lowest reconstruction losses and a much less drastic difference between scaling factors than smaller learning rates. Lower $L1$ weights lead to steeper increases in the number of active neurons against the scaling factor (again aligning with simulation results). Lower number of active neurons ($25th$ percentile) and good reconstruction loss ($5th$ percentile) can be achieved with learning rates of 10^{-4} and a $L1$ weights of 0.001 or 0.0001 (slight differences for datasets, shift by one log step). There were no trends or large differences between Vanilla and Bricken SAEs.

For analysis, Vanilla SAEs were trained for 500 epochs with Adam optimizer Kingma and Ba [55], a learning rate of 10^{-4}, batch size 128, hidden activation dimension 10000 (500-fold increase for multiDGD) and an L1 weight of 10^{-3} (see loss curves in Supplementary Figure 14 for multiDGD human bone marrow) for multiDGD's and Geneformer's embeddings from the human bone marrow with random seeds $[0, 42, 9307]$. Compute requirements are low, with training taking 20 minutes for the 56k training samples.

A.3.4 DGE analysis

Sample groups were investigated in terms of relevant changes to gene expression through differential gene expression analysis (DGE). In the case of the "perturbed-vs-normal" paired samples, this was done with negative binomial generalized linear models as is common in biological data analysis [57, 58]. The resulting p-values and fold changes from the models are reported. For the unpaired "high-vs-low" comparison, t-tests were performed between the groups for each gene and calculated the fold change based on mean expression. Corrected p-values were computed based on multi-test correction with Benjamini/Hochberg correction for non-negative values [59] for all experiments.

A.3.5 Manual GO term enrichment analysis

In order to identify biological processes related to the differentially expressed genes, genes were filtered by adjusted p-values (threshold 10^{-10}) and in the case of CD8+ T cells also fold change (10-fold and inverse) to get as highly specific processes as possible. Biological processes related to the resulting gene sets were identified through GO term analysis with default parameters at https://geneontology.org/docs/go-enrichment-analysis/ [45, 46].

A.3.6 Feature characterization

SAE features were distinguished into local and global features based on whether they were only active in a single cell type or similarly active in multiple cell types. This was assessed by calculating the significance measures of activations per feature over cells from a specific cell type vs all other cells. Features with

significantly higher activations in only one cell type were labeled as local. Significance was determined based on a two-tailed test with confidence interval $95\,\%$ ($z = 1.96$) as

$$\alpha = |\mu_j - \mu_i| - 1.96(\frac{\sigma_j}{\sqrt{N_j}} + \frac{\sigma_i}{\sqrt{N_i}}) \tag{8}$$

with means μ, standard deviations σ, and number of observations N for two cell type distributions i and j. The null hypothesis is rejected if $\alpha \geq 0.05$. This significance measure is used to determine relevant differences between samples for SAE feature activations and in one analysis also chromatin accessibility (openness).

A.3.7 Automated GO term analysis

DGE analysis was performed on the predicted expression counts for feature-specific "high-vs-low" sample sets as described in A.3.4. Next, a GO term analysis was performed according to Mi et al. [52] with a binomial test and a Mann-Whitney U (MWU) test for all GO terms with 20 to 500 reference genes available in our 13431 genes. The MWU test were performed with the ranked fold changes (smallest rank 1). The metric was calculated as

$$\begin{aligned} U = \min \Big(& U_1 = n_1 n_2 \frac{n_1(n_1+1)}{2} - R_1, \\ & U_2 = n_1 n_2 \frac{n_2(n_2+1)}{2} - R_2 \Big) \end{aligned} \tag{9}$$

with n_1 and n_2 presenting the number of genes in the GO term gene set and the remaining genes, respectively. R_1 and R_2 correspondingly present the average ranks of these groups. Z-scores, p-values, and effect sized of the test are reported. The binomial test was conducted on the most relevant genes from the DGE analysis based on two thresholds. Firstly, the number of genes identified for an adjusted p-value threshold of 10^{-5} and a fold change of at least 2 (or below 0.5) were computed. If this returned zero genes, the p-value threshold was increased to 0.05 and the fold change excluded. Afterward, the p-value, number of expected genes, fold enrichment and false discovery rate for k hits (relevant genes that are also found in the GO term gene set), n_s samples in the study (the relevant genes returned by DGE analysis), and p_c as the probability of randomly finding one of the GO term genes ($p_c = n_c/n$ with n as the total number of genes and n_c as the number of genes associated with the GO term) were computed.

A.3.8 Optimal bipartite matching

Optimal bipartite matching computes the Jaccard distance between all features from DGD and Geneformer SAEs and then finds the optimal matching via the Hungarian algorithm. Overall matrix similarity is computed as the average Jaccard distance of the matched pairs. Values can be between 0 (no similarity) and 1 (perfect similarity).

Appendix B: Supplementary Materials

Tables

Supplementary Table 2: Autoencoder hyperparameter configurations

Model Configuration	Dropout	Learning Rate	Weight Decay	Batch Size
20-2-narrow	0.0	1e-4	1e-3	128
20-2-wide	0.0	1e-5	1e-3	128
20-4-narrow	0.0	1e-4	0.0	128
20-4-wide	0.0	1e-5	1e-5	128
20-6-narrow	0.0	1e-5	1e-5	128
20-6-wide	0.0	1e-6	1e-5	512
100-2-narrow	0.0	1e-4	1e-5	128
100-2-wide	0.0	1e-5	1e-5	128
100-4-narrow	0.0	1e-4	1e-7	128
100-4-wide	0.0	1e-5	1e-5	128
150-2-narrow	0.0	1e-4	1e-5	128
150-2-wide	0.0	1e-5	1e-5	128
150-4-narrow	0.0	1e-4	1e-7	128
150-4-wide	0.0	1e-5	1e-5	128
150-6-narrow	0.0	1e-4	1e-5	512
150-6-wide	0.0	1e-5	0.1	256
1000-2-narrow	0.0	1e-5	1e-5	128
1000-2-wide	0.0	1e-5	1e-5	128
1000-4-narrow	0.0	1e-4	0.0	128
1000-4-wide	0.0	1e-5	1e-7	128
1000-6-narrow	0.0	1e-6	0.0	512
1000-6-wide	0.0	1e-5	1e-7	512

Supplementary Table 3: **Simulation autoencoder hyperparameters**

Simulation	Hyperparameter	Choices
small	latent dimension	4
	learning rate	10^{-4}
	Number of layers	1
	Number of hidden neurons	n.a.
	Batch size	1
	Weight decay	0
	Random seed	0
large	latent dimension	$[20, 100, 150, 1000]$
	learning rate	see Table S2
	Number of layers	$[2, 4, 6]$
	Number of hidden neurons	["narrow", "wide"] (details in Appendix A A.2.2)
	Early stopping	20 epochs
	Batch size	see Table S2
	Weight decay	see Table S2
	Random seed	$[0, 42, 9307]$

Supplementary Table 4: **Simulation SAE hyperparameters**

Simulation	**Hyperparameter**	**Choices**
small	scaling factor	$[2, 5, 10, 20, 50, 100, 200, 1000]$
	learning rate	$[10^{-2}, 10^{-3}, 10^{-4}, 10^{-5}]$
	L1 weight	$[10^{-1}, 10^{-2}, 10^{-3}, 10^{-4}, 10^{-5}]$
	k (TopK percent active neurons)	$[5, 10, 20, 50, 75, 100]$
large	scaling factor	$[20, 100, 200, 500]$
	learning rate	$[10^{-4}, 10^{-5}, 10^{-6}]$
	L1 weight	$[10^{-1}, 10^{-2}, 10^{-3}, 10^{-4}]$
	Early stopping	20 epochs
	Batch size	128

Supplementary Table 5: **Simulation variable recovery in an autoencoder with latent dimension 150.** Per hidden generative variable, the maximum Pearson correlation of all features against all variable dimensions are reported. For the SAE, an average of highest correlations over 4 SAEs with differen hidden scaling factors are reported $\pm$ SEM.

Method	X''	X'	X	A	B
PCA	**0.72**	0.36	0.35	-0.12	0.56
ICA	0.54	**0.50**	**0.50**	0.02	-0.01
SVD	0.66	0.41	0.32	-0.09	0.62
SAE	**0.74** $\pm$ 0.02	0.43 $\pm$ 0.01	0.19 $\pm$ 0.01	-0.54 $\pm$ 0.01	**0.65** $\pm$ 0.01

Supplementary Table 6: **Robustness of number of live neurons and feature types for different random seeds**

Feature type	**Mean $\pm$ SEM**
Local (live)	914.00 ± 2.45
Global (live)	4427.33 ± 7.08
Dead	4658.67 ± 9.53

Supplementary Table 7: **Top 5 most abundant GO terms in the automated analysis**

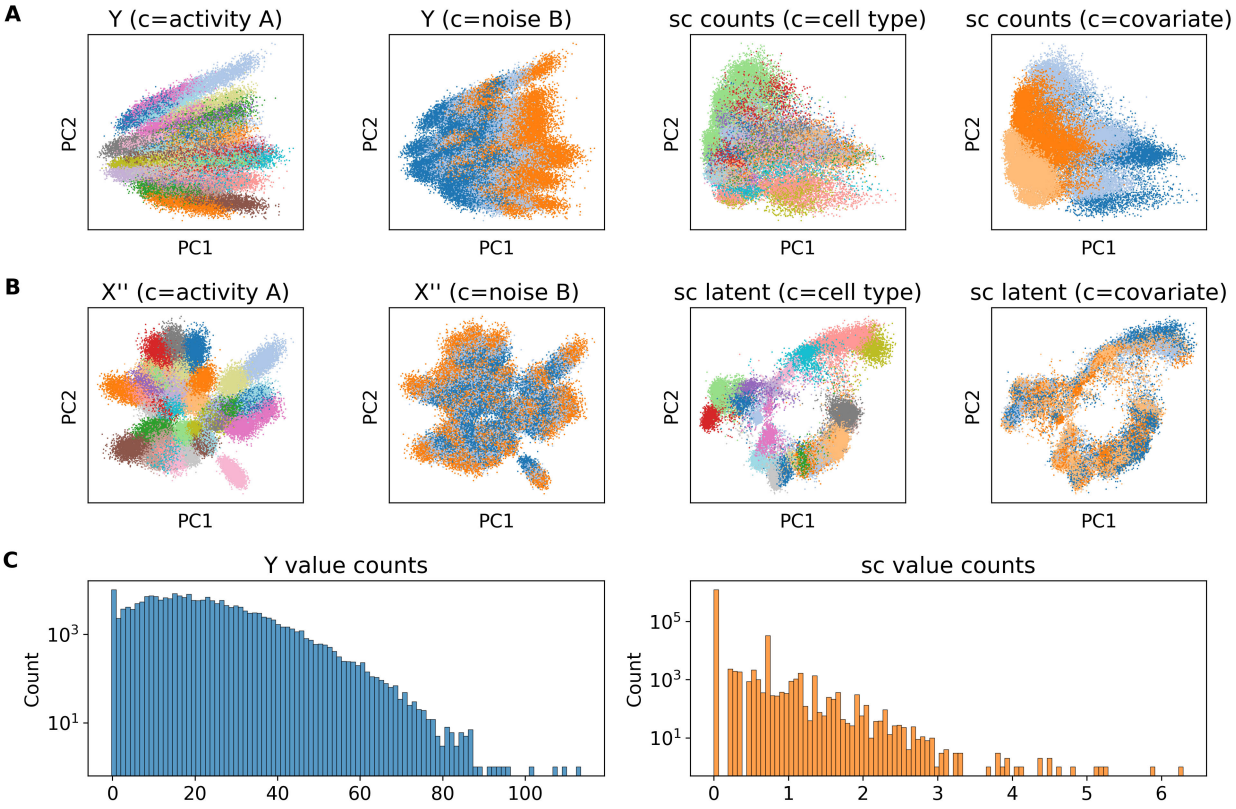
GO name (multiDGD)
immune response
cell surface receptor signaling pathway
structural constituent of ribosome
adaptive immune response
inflammatory response

GO name (Geneformer)
cytoplasmic translation
translation
structural constituent of ribosome
chromatin binding
mRNA splicing, via spliceosome

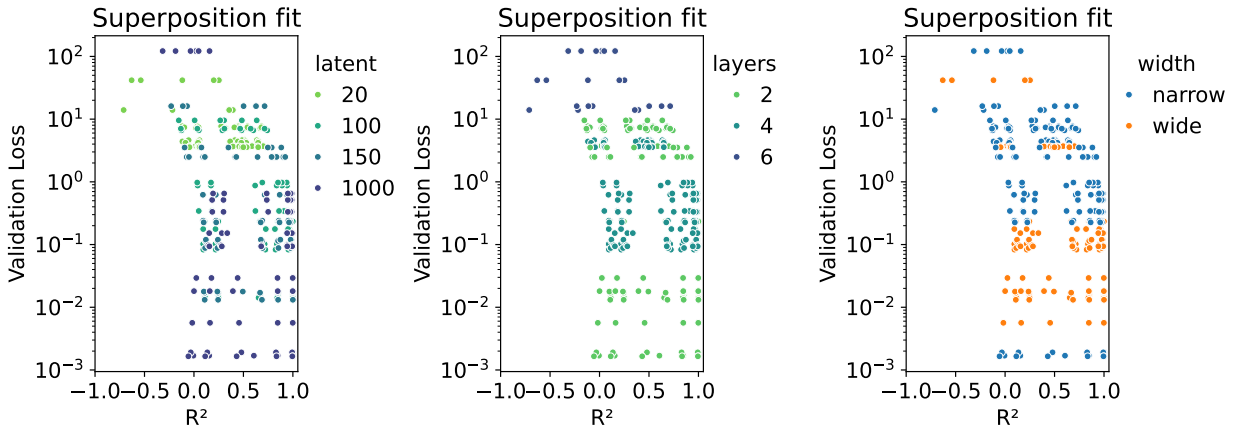
Supplementary Table 8: **Feature 2306 Perturbation GO terms.** Go terms associated with the gene lists derived from DEG analysis for each cell type perturbation experiment. Only highly specific GO terms are shown with maximum 400 gene references. GO terms appearing for more than one experiment are highlighted in bold font. Abbreviations: CT - cell type, HSC - hematopoietic stem cell, PE - proerythroblast, NK - natural killer cell, CD8T - CD8+ T cell.

GO term	**Present in CT perturbation**			
	HSC	PE	NK	CD8T
Intracellular **calcium** ion homeostasis	✓			
Carbon dioxide transport		✓	✓	
Oxygen transport		✓	✓	
Hydrogen peroxide catabolic process		✓	✓	
Positive regulation of myoblast differentiation		✓		
Erythrocyte development		✓		
Vascular process in circulatory system		✓		
Nitric oxide transport			✓	
Stimulatory C-type lecitin receptor signaling pathway			✓	
Positive regulation of natural killer cell mediated cytotoxicity			✓	
Myeloid leukocyte activation			✓	
Chemokine-mediated signaling pathway				✓
Calcium-mediated signaling				✓

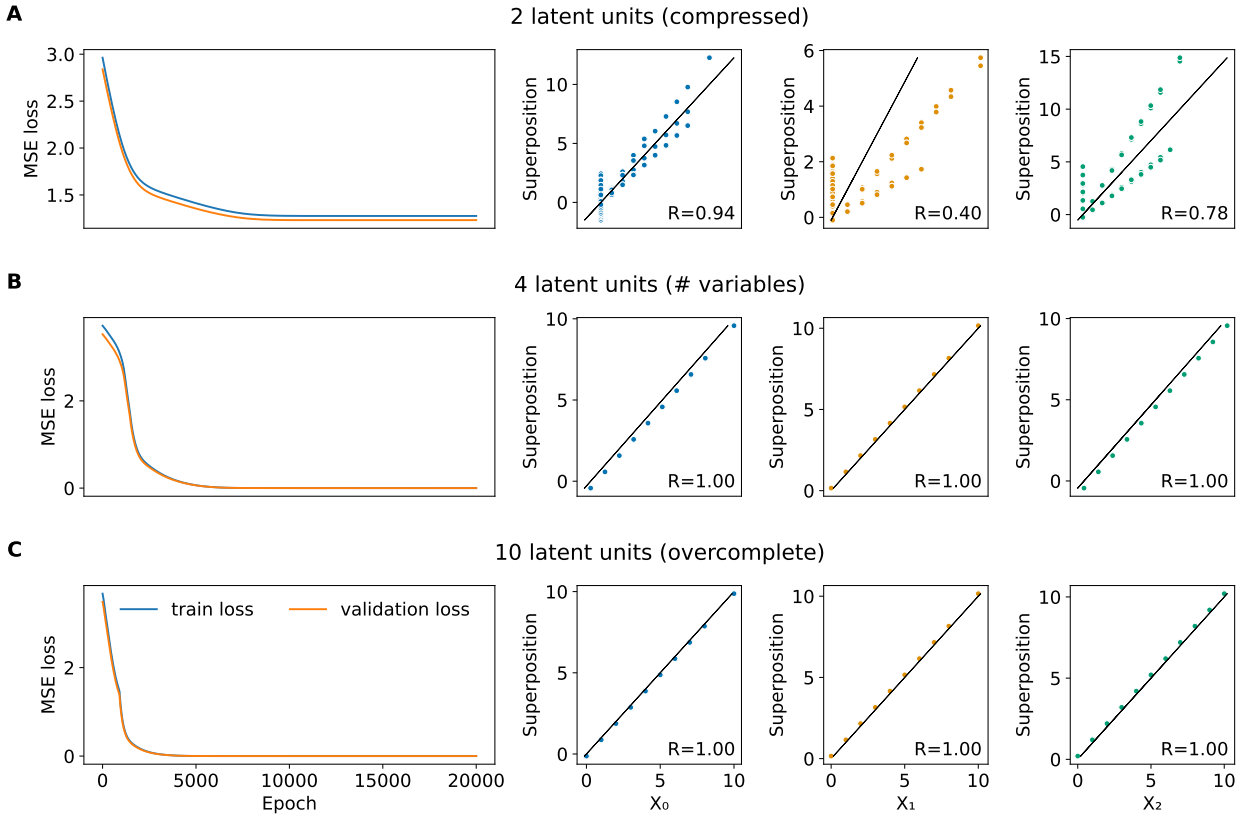
Figures



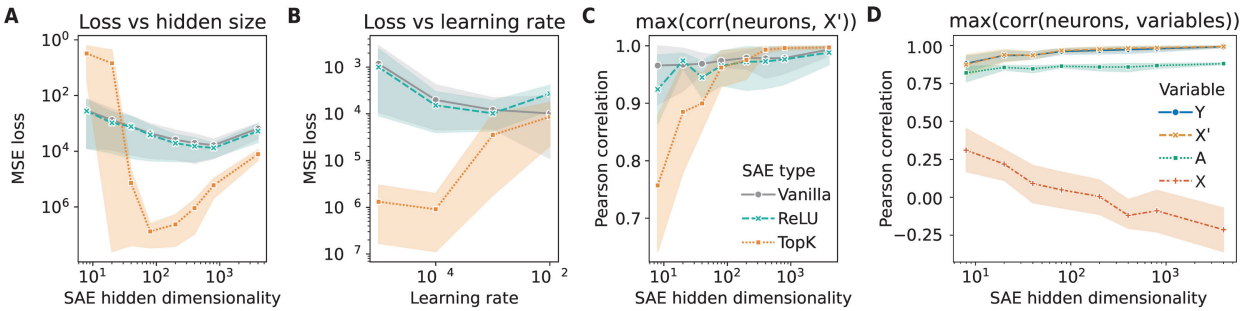
Supplementary Figure 1: **Simulated and single-cell data.** **A** PCAs of simulated observables Y and log-transformed single-cell (sc) counts colored by A/celltype and B/technical covariate, respectively. **B** PCAs of simulated latents X'' and inferred (not generative) latents from the sc model. PCAs are again colored by A/celltype and B/technical covariate, respectively. **C** Histograms of simulated Y values and real sc counts. Simulated data does not directly match the specific single-cell dataset presented here. However, clusters of A and B appear similar to our real-world comparison (cell type and technical covariate). The values in C are generally higher for the simulation and less sparse, but still match zero-inflated Negative Binomial distributions which are typically used to describe these count data.



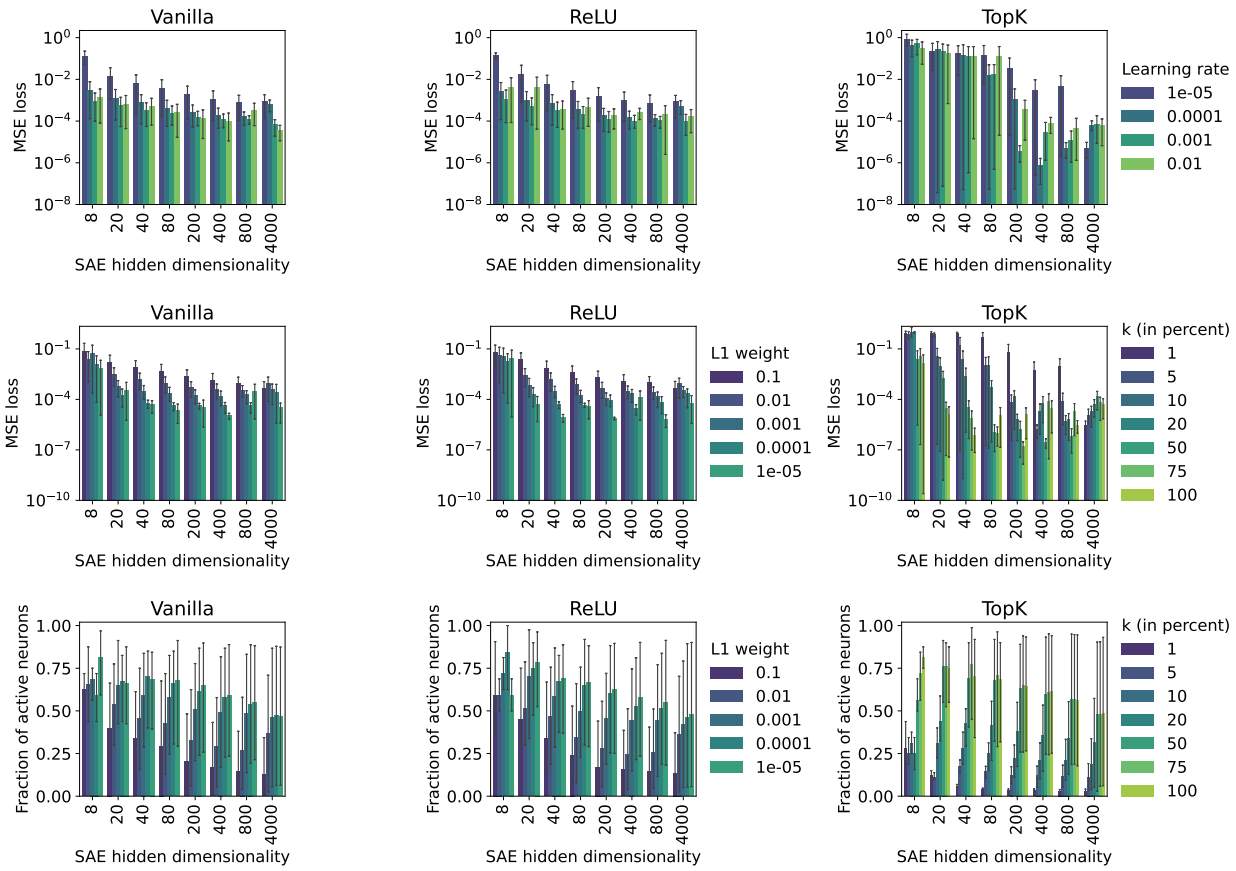
Supplementary Figure 2: **Validation loss against superposition fits for large simulation autoencoders.** AE performance (validation loss) vs. superposition fit. Coefficients of determination R^2 were computed based on linear regression performed on the AE latent representations w.r.t. each of the variables on the left. Colors present the latent dimension, number of hidden layers, and architecture width (details in Appendix A A.2.2), respectively.



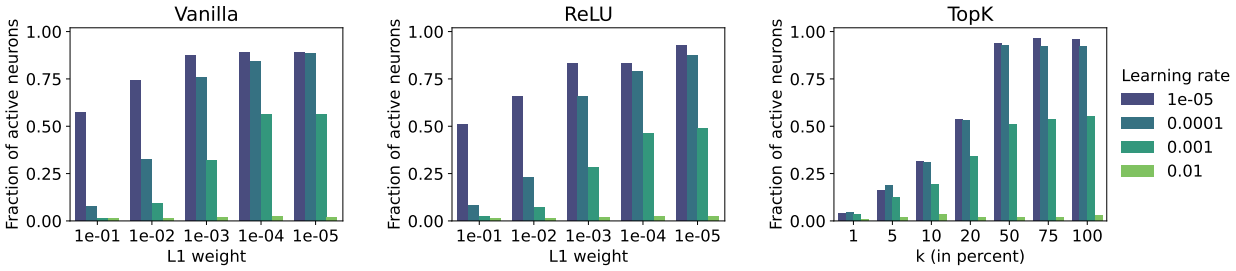
Supplementary Figure 3: **Superpositions in compressed, "ideal", and overcomplete autoencoders trained on simulated data. A)** The top row depicts learning curves of train and validation MSE loss over epochs (left, legend in C) and superpositions of the three variables X (right) of a single-layer autoencoder with a compressed bottleneck (2 dimensions). The superpositions are plotted as the product of the latent representations and coefficients from linear regression against the true values of X. Linear regression was performed between the latent representations and true X values. Points along the black line indicate a perfect fit of the superpositions (quantified by the R value rounded to two decimals in the bottom right corner (maximum 1). **B)** Same as A for the "ideal" case, in which the number of latent units is equal to the number of generative random variables. **C)** Same as A and B for the overcomplete case with 10 hidden units.



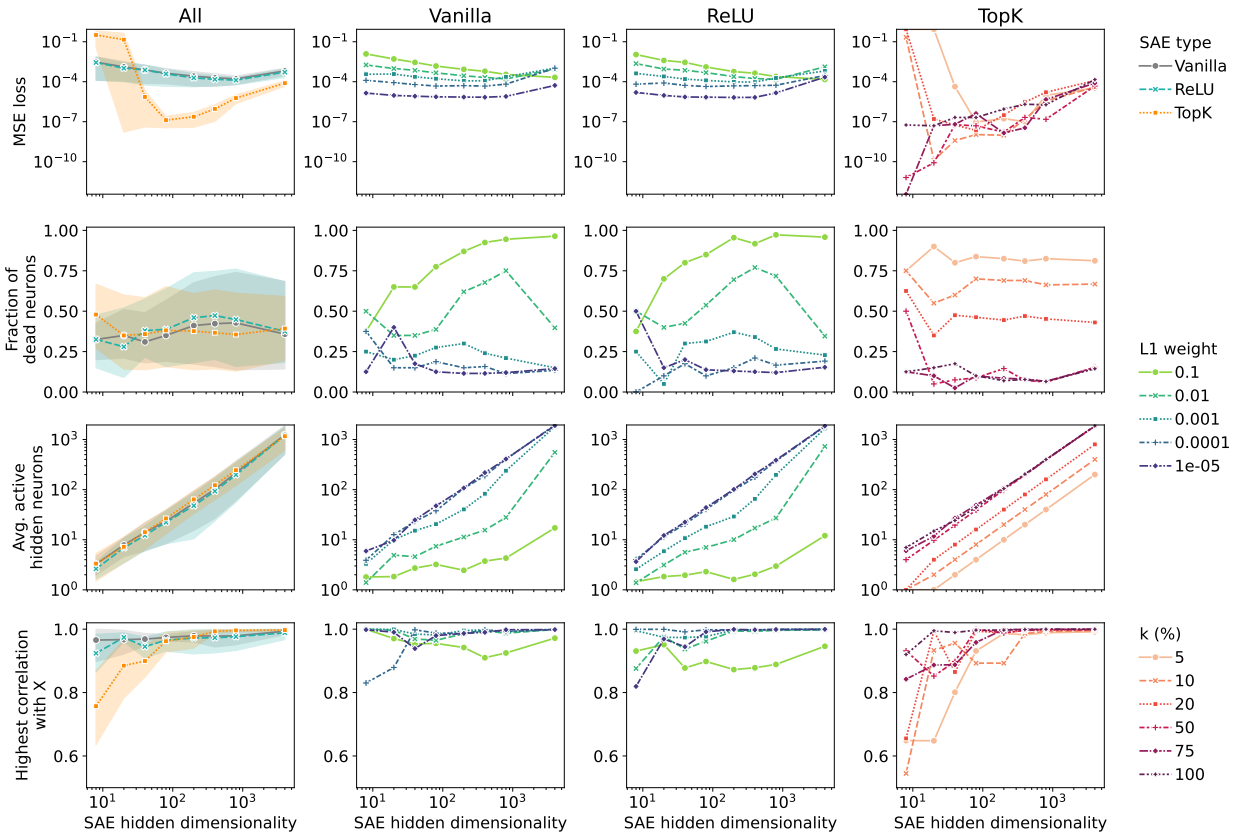
Supplementary Figure 4: **Performances of different SAE architectures on the small simulation data.** Performances of the three SAE types are presented as line plots with points depicting the average values over hyperparameter runs per SAE type (N listed with each plot) and lines and areas as projections of mean and 95 % confidence, respectively. Vanilla, ReLU, and TopK SAEs are identified in legend C. **A** MSE loss against hidden dimensionality (learning rate 10^{-4}, $N = 5$). **B** MSE loss against learning rates ($N = 40$). **C** Maximum Pearson correlation between SAE neurons and hidden variable X' of the simulated data against hidden dimensionality (learning rate 10^{-4}, $N = 5$). **D** Recovery of simulation variables. Maximum Pearson correlation between SAE neurons and hidden variables of the simulated data against hidden dimensionality. Variables are explained in the legend to the right (learning rate 10^{-4}, $N = 16$ samples per point including all SAE types).



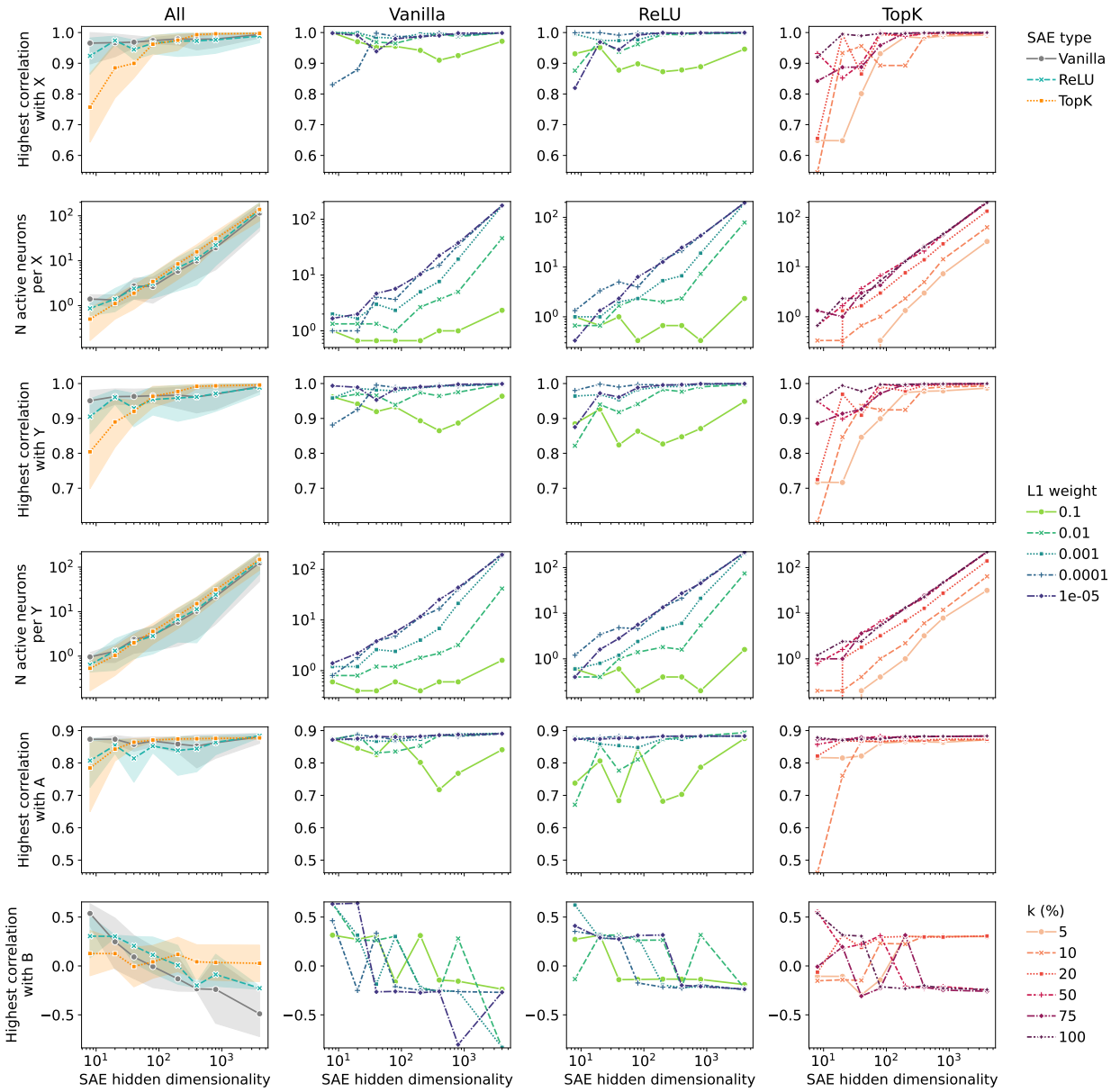
Supplementary Figure 5: **Hyperparameter bar plots of different types of SAEs trained on representations from the simulation experiment (latent dimension 4).** Columns depict performances for the SAE types Vanilla, ReLU, and TopK. Rows present different combinations of performance metrics. **A)** MSE loss against the hidden dimensionality colored by learning rate. $N = 5$ runs per bar. **B)** Same as A colored by the sparsity penalty (L_1 weight for Vanilla and ReLU, k in percent of hidden units for TopK). $N = 4$ runs per bar.**C)** Fraction of active neurons against the hidden dimensionality colored by the sparsity penalty. $N = 4$ runs per bar. Error bars indicate the 95th confidence interval.



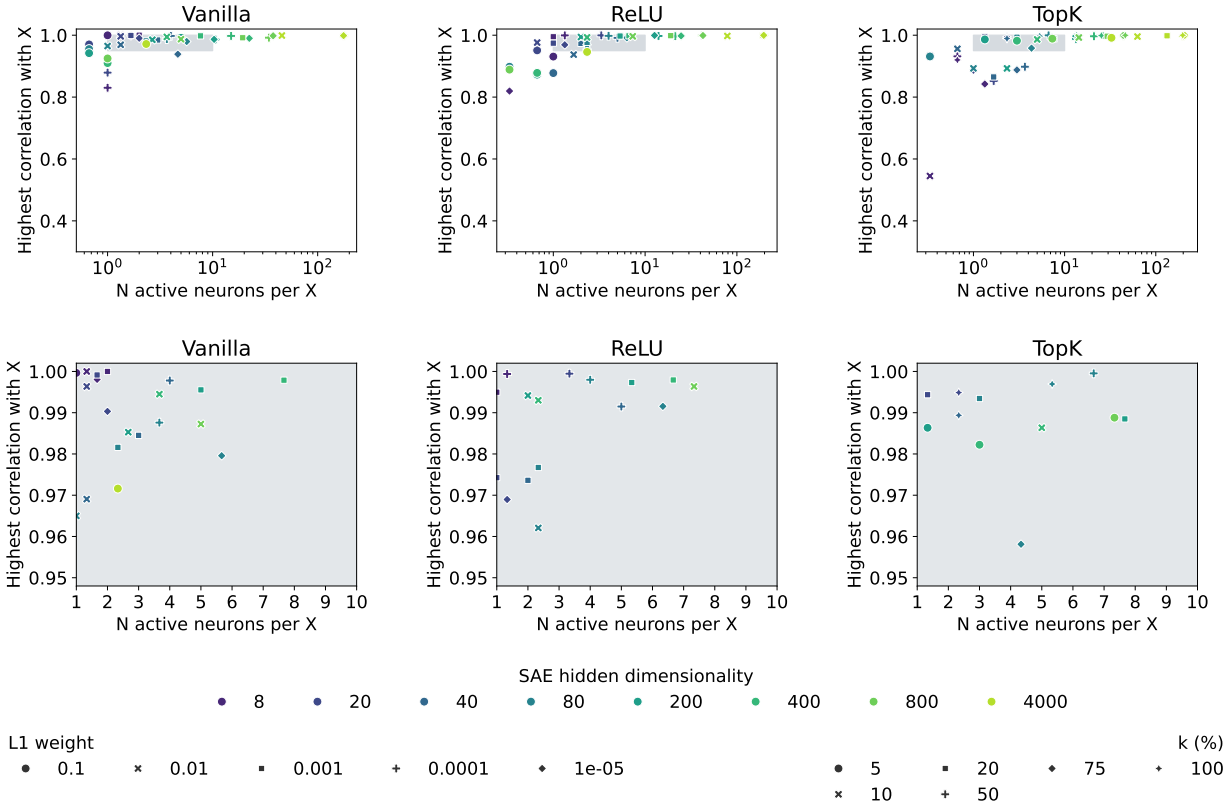
Supplementary Figure 6: **Influence of learning rate on the number of active neurons.** Bar plots of the three SAE types trained on the same representations as above for a hidden dimension of 400 ($100 \times$ latent). Columns depict performances for the SAE types Vanilla, ReLU, and TopK. The fraction of active neurons is plotted against the sparsity penalty colored by the learning rate ($N = 1$).



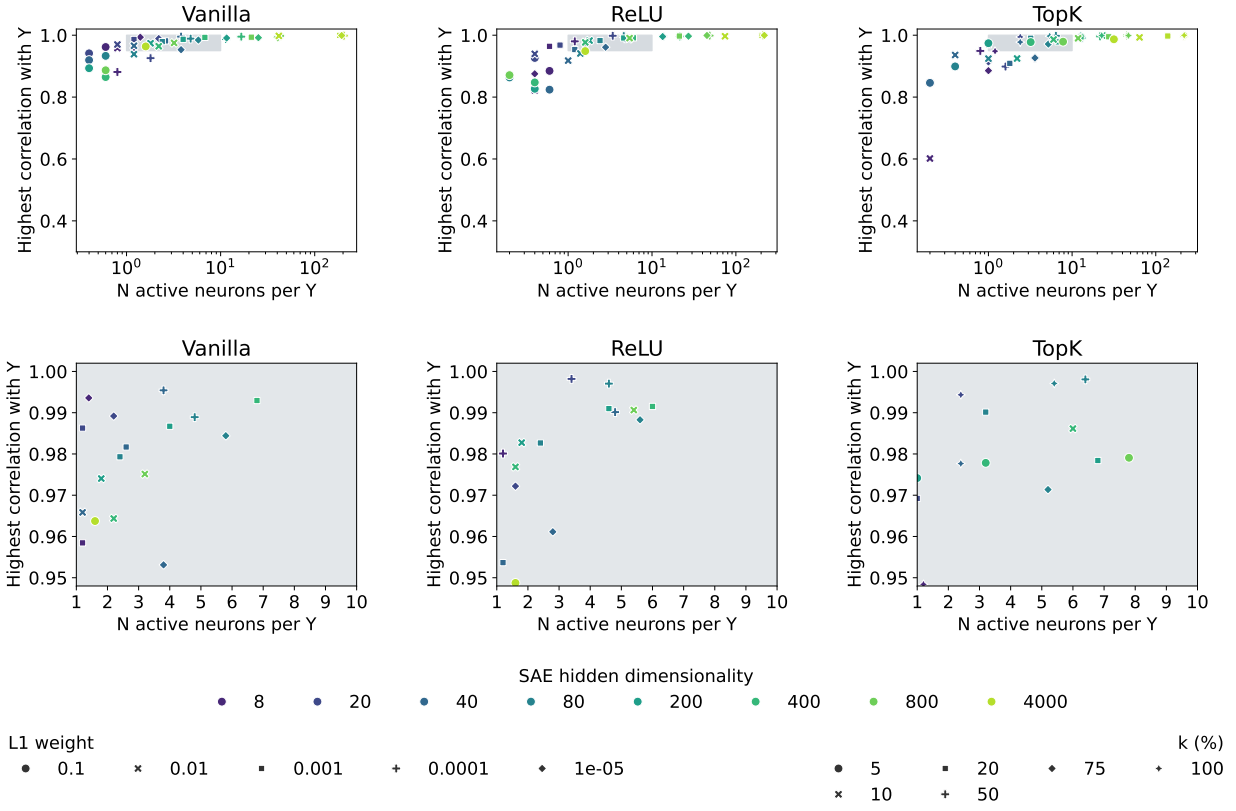
Supplementary Figure 7: **Performance comparison of SAEs for learning rate 10^{-4}.** The first column presents accumulated line plots of specific metrics for the three different model types over the hidden dimensionality ($N = 5$ and $N = 6$ samples per point for Vanilla/ReLU and TopK, respectively) with the area as the 95th confidence interval. The other three columns show the individual data points as line plots colored by the sparsity penalty. Legends to the right. The rows depict different metrics on the y axes: MSE loss, fraction of dead neurons, average number of firing neurons per sample, highest Pearson correlation of SAE neurons with variables X.



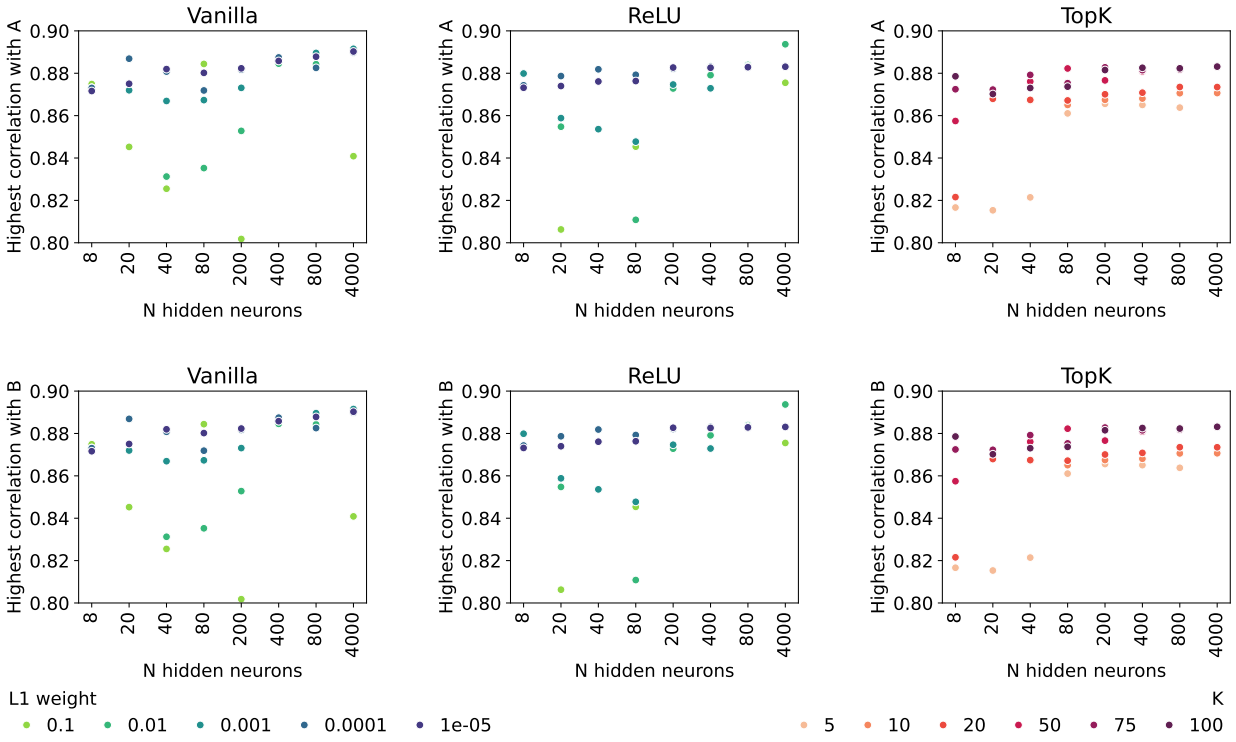
Supplementary Figure 8: **Comparison of variable recovery in different SAEs for learning rate 10^{-4}.** Same as Supplementary Figure 7 with different metrics on the y axes. Metrics refer to the highest Pearson correlation of SAE neurons with the simulation variables, as well as the number of corresponding SAE neurons with a correlation threshold of $> 95\,\%$.



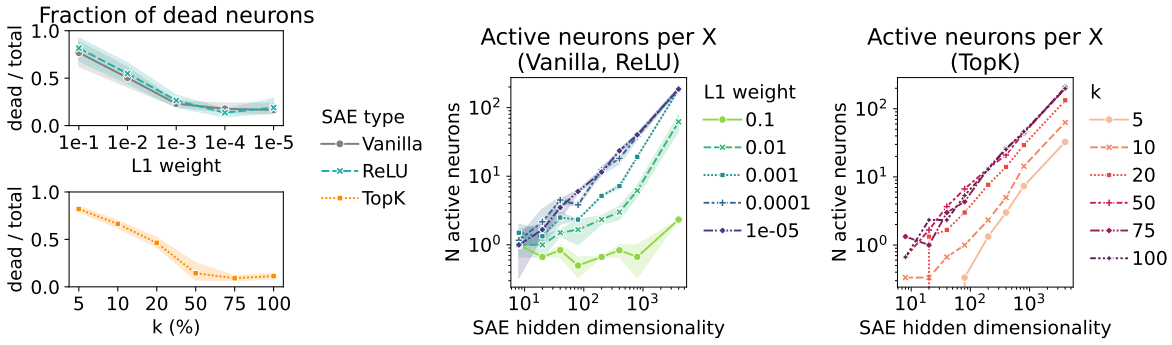
Supplementary Figure 9: **Sensitivity and specificity of SAE neurons for variable X.** Highest correlations of SAE neurons plotted against the number of active SAE neurons with a correlation threshold of $> 95\,\%$ for Vanilla, ReLU, and TopK SAEs (columns from left to right). Colors indicate the hidden dimensionality. Data point styles indicate the sparsity penalty, explained in the legend at the bottom. The top row shows all model setups. The bottom row depicts the area highlighted as a grey box in the top row.



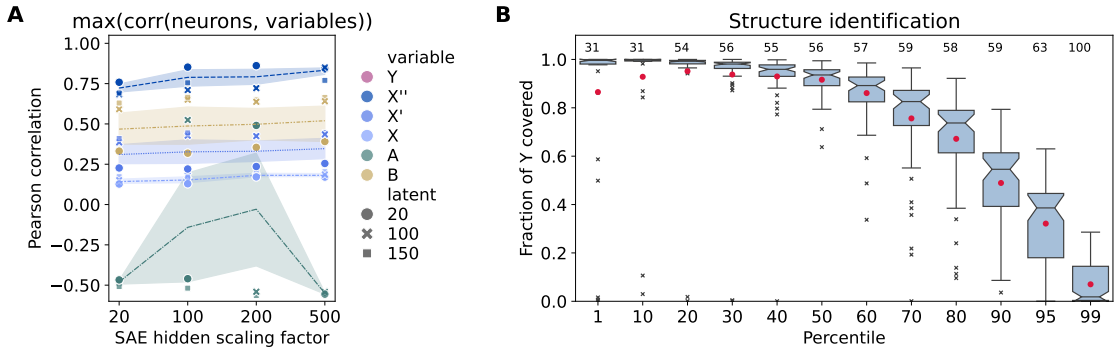
Supplementary Figure 10: **Sensitivity and specificity of SAE neurons for variable Y.** Highest correlations of SAE neurons plotted against the number of active SAE neurons with a correlation threshold of $> 95\,\%$ for Vanilla, ReLU, and TopK SAEs (columns from left to right). Colors indicate the hidden dimensionality. Data point styles indicate the sparsity penalty, explained in the legend at the bottom. The top row shows all model setups. The bottom row depicts the area highlighted as a grey box in the top row.



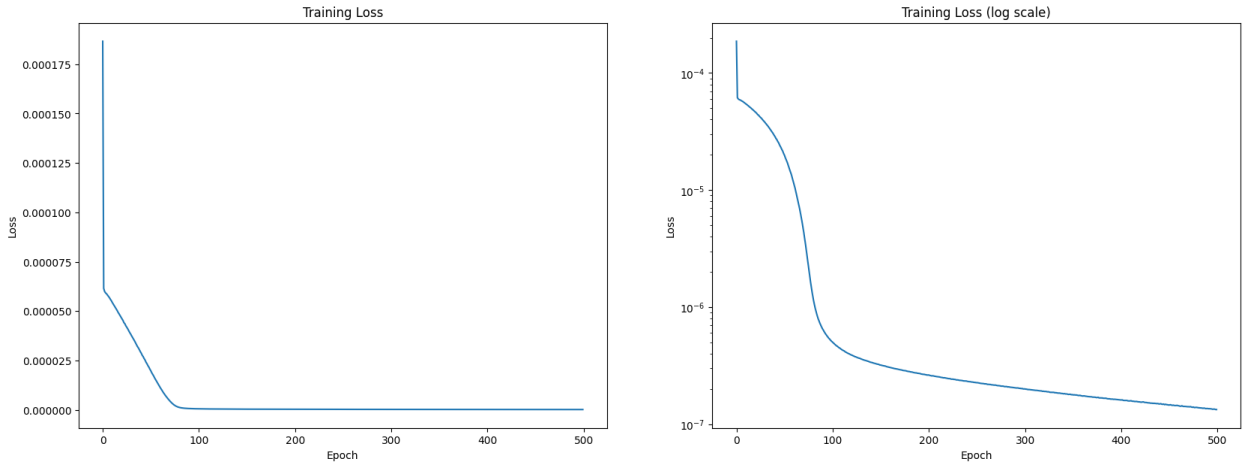
Supplementary Figure 11: **Sensitivity and specificity of SAE neurons for variables A (top) and B (bottom).** Highest correlations of SAE neurons plotted against the dimensionality of the SAE hidden space for Vanilla, ReLU, and TopK SAEs (columns from left to right). Colors indicate the sparsity penalty, explained in the legend at the bottom. The top row shows all model setups.



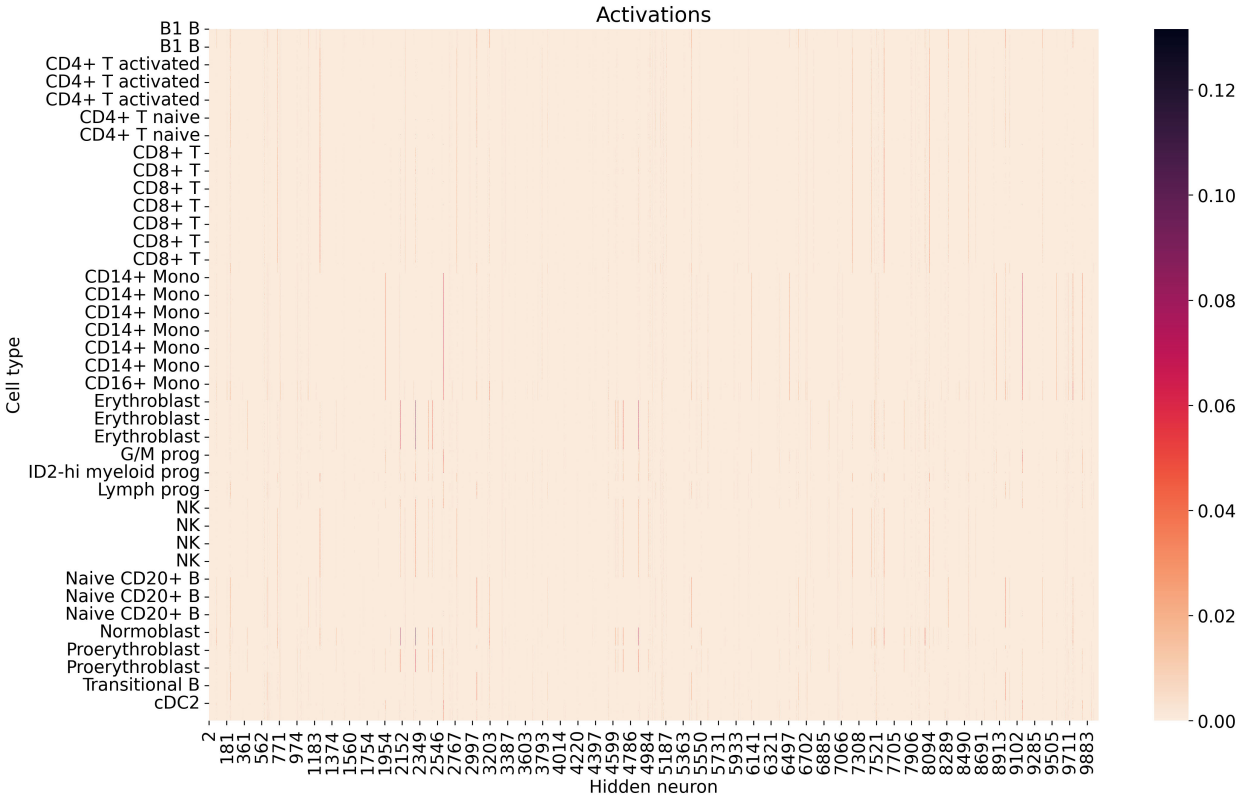
Supplementary Figure 12: **Redundancy of SAE features.** The two line plots show the number of active neurons per variable X colored by sparsity parameter for Vanilla/ReLU (sparsity parameter: L1 weight) and TopK (sparsity parameter: k) SAEs, respectively. The number of features are plotted against the total number of hidden neurons in the SAE. Line plots are set up as in Figure 4 with $N = 2$ and $N = 1$ samples per point, respectively.



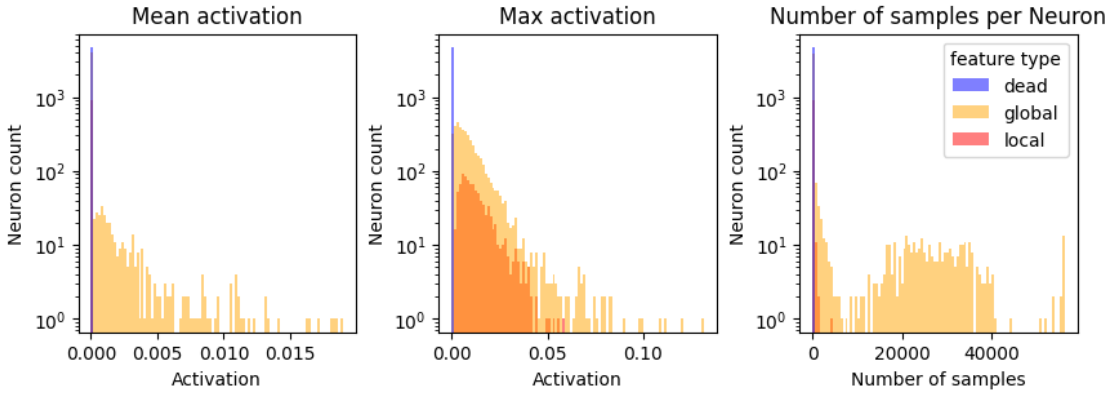
Supplementary Figure 13: **Recovery of large simulation variables and structure in SAE features. Left:** Maximum Pearson correlation between SAE neurons and hidden variables of the simulated data against hidden scaling factor ($N = 3$). Points are colored by variable and the style depicts the AE latent dimensionality (legend on the right). **Right:** Boxplot of the fraction of "genes" Y regulated by individual X'' variables connected to best matching SAE features. The x axis presents percentiles of the cosine similarities between SAE features and Y. The boxplot center line depicts the median, notches the $95\,\%$ confidence interval, and error bars 1.5 times the interquartile range. Red dots present the means and numbers above indicate the number of samples per boxplot (= the number of X variables out of 100 that were matched with an SAE feature).



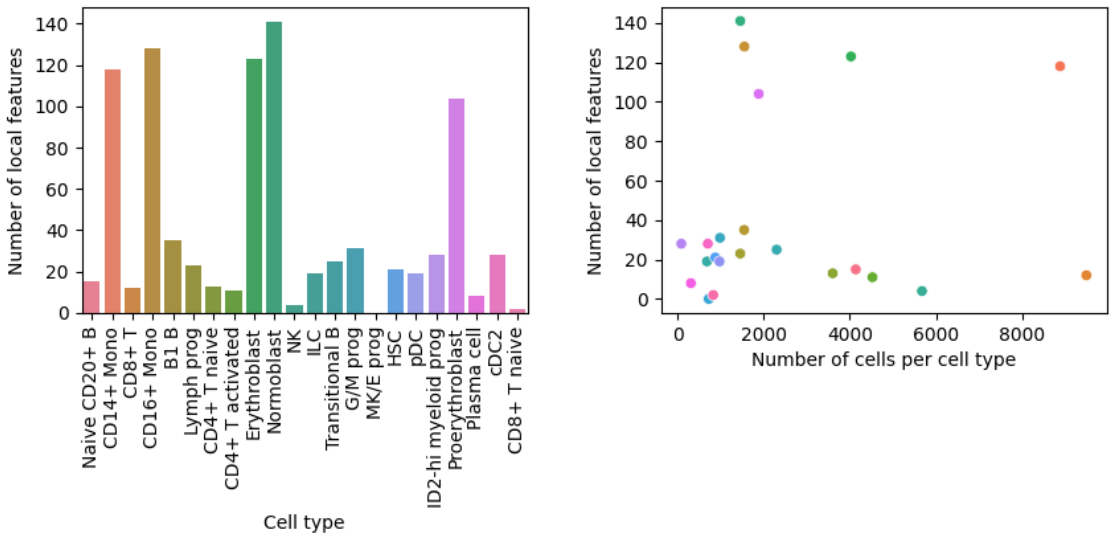
Supplementary Figure 14: **SAE training loss curve for the human bone marrow model.** The reconstruction loss (MSE) is plotted against the epochs. The right plot depicts the log-scaled loss.



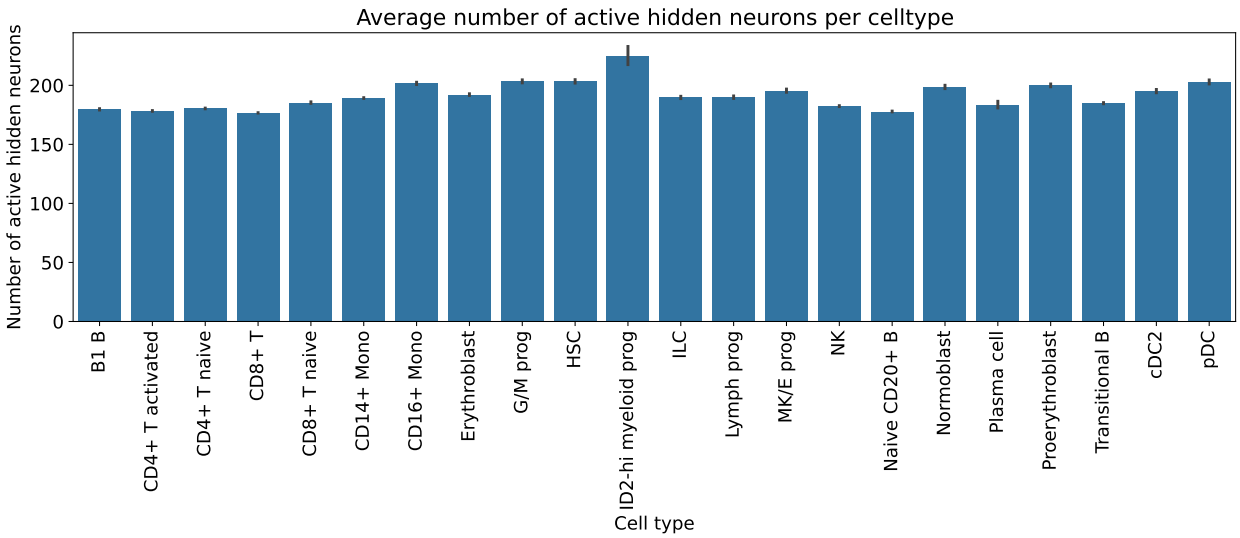
Supplementary Figure 15: **Heatmap of SAE activations from human bone marrow.** All samples are sorted by cell type on the y axis. All activations of active neurons are plotted on the x axis. The legend on the right describes the color range of the activations.



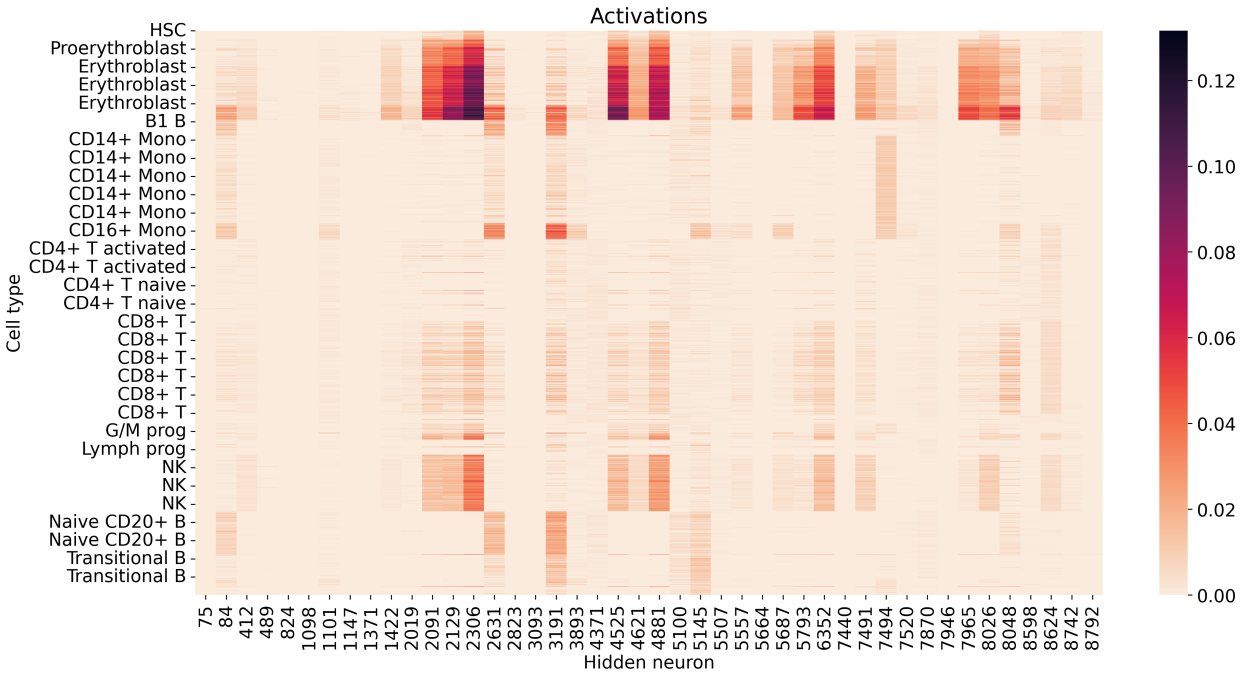
Supplementary Figure 16: **Feature activations of the SAE trained on human bone marrow single-cell data.** Log-scale neuron counts are plotted against mean activation, maximum activation, and the number of samples per neuron. Histograms are colored by the type of neuron.



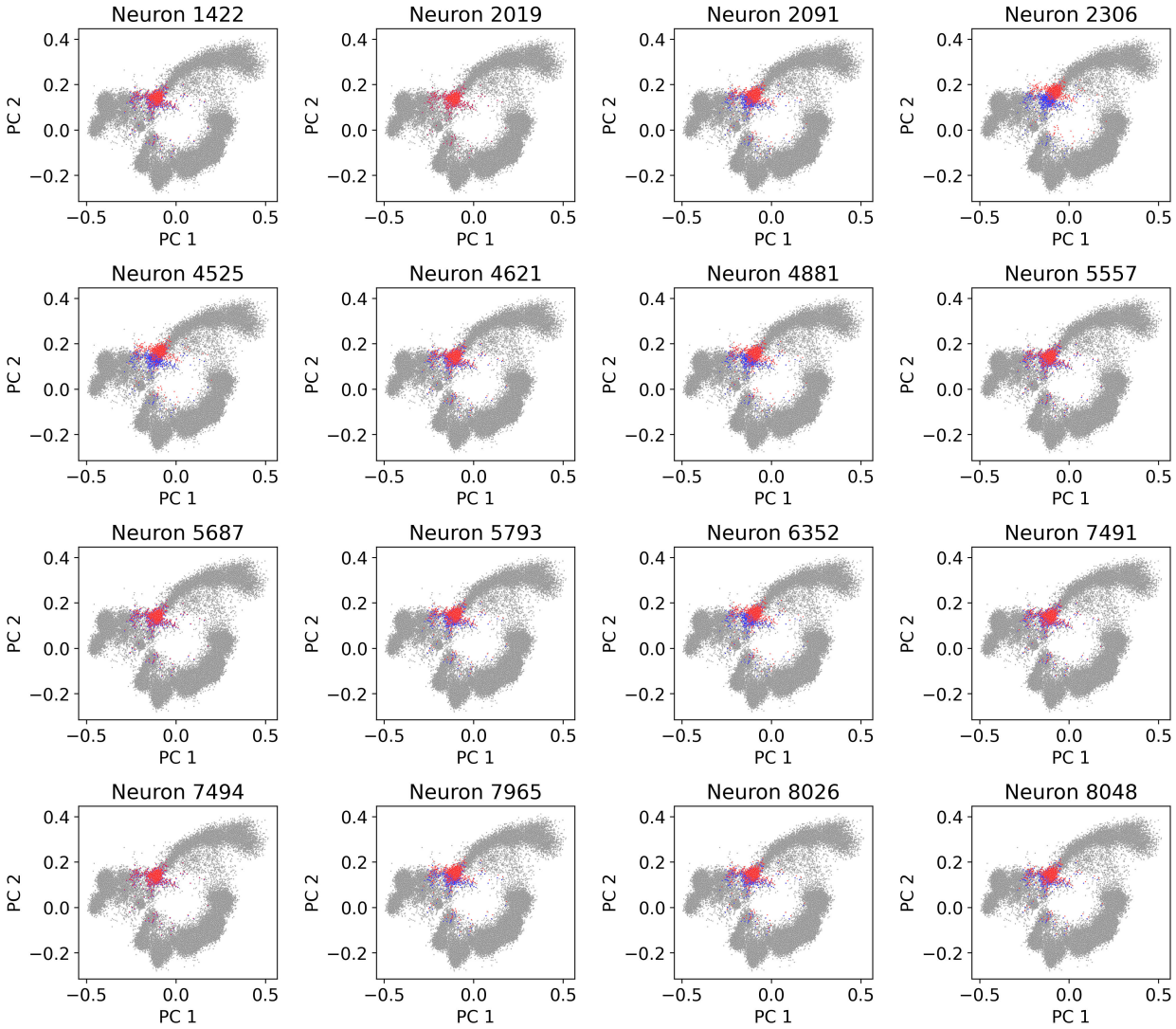
Supplementary Figure 17: **Distribution of local features among cell types.** The left shows a bar plot of the number of local features associated with each cell type. The right shows the number of local features plotted against the number of cells per cell type. Colors are the same as on the left.



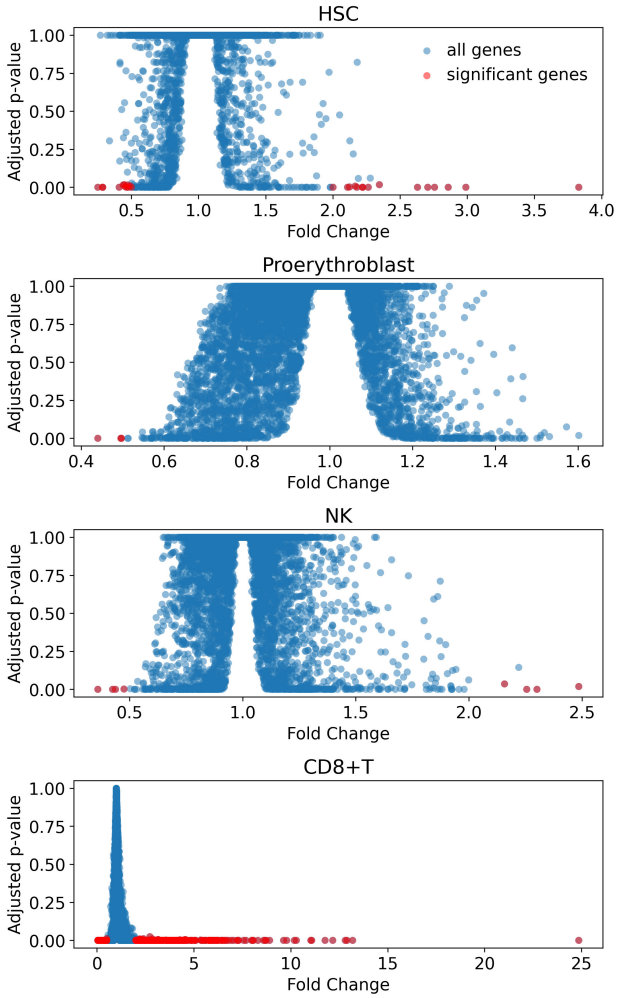
Supplementary Figure 18: **Average firing neurons per cell type.** Bar plots of the number of firing neurons per sample, plotted by cell type. Error bars indicate the 95th confidence interval.



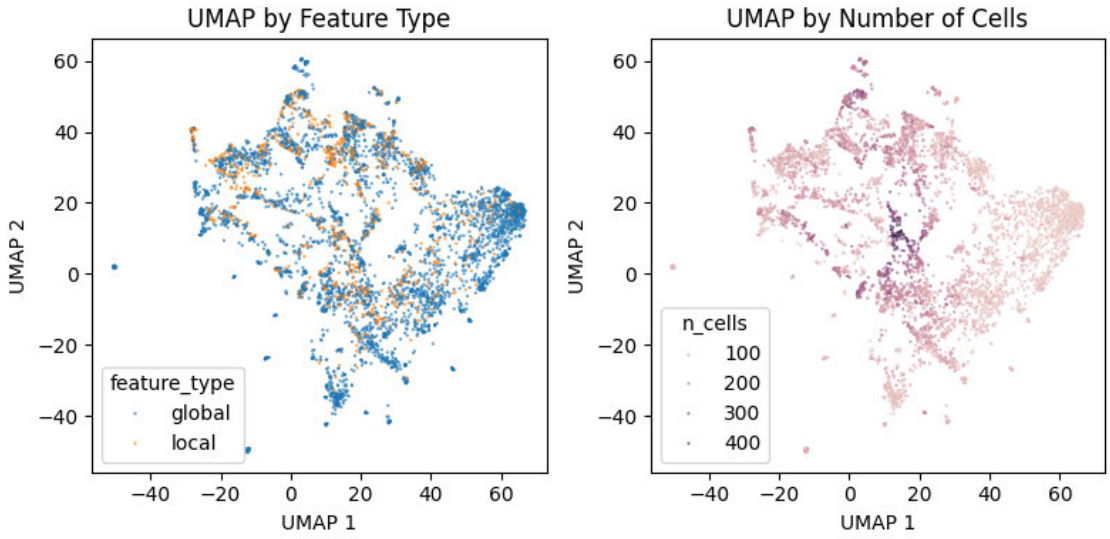
Supplementary Figure 19: **Activations of potential features for red blood cell development.** All samples are sorted by cell type on the y axis. Activations of neurons that fulfilled the requirements for red blood cell development are plotted on the x axis. The legend on the right describes the color range of the activations.



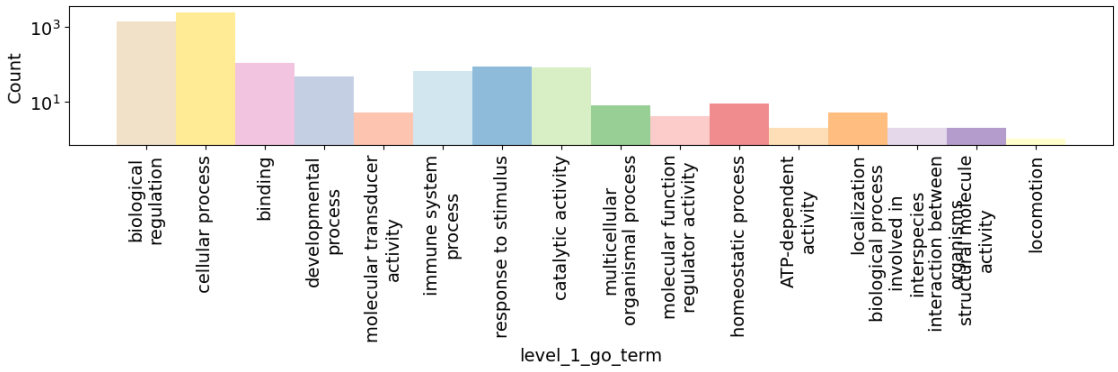
Supplementary Figure 20: **Effect of perturbations on potential features for red blood cell development.** PCA plots of the extracted single-cell representations (grey dots). Titles indicate the neuron that was perturbed. Blue and red dots present normal and perturbed samples, respectively.



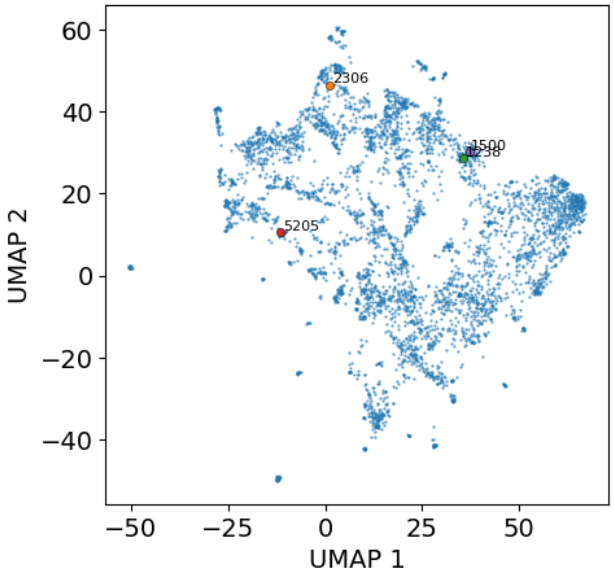
Supplementary Figure 21: **Differential gene expression analysis of perturbation experiments.** The plot shows the adjusted p-values against the fold change for all genes modeled by multiDGD [31]. Each row shows the results of one of the four experiments indicated by the plot titles. Red data points depict genes with an adjusted p-value below 0.05 and a fold change below 0.5 or above 2 (see legend in the top plot).



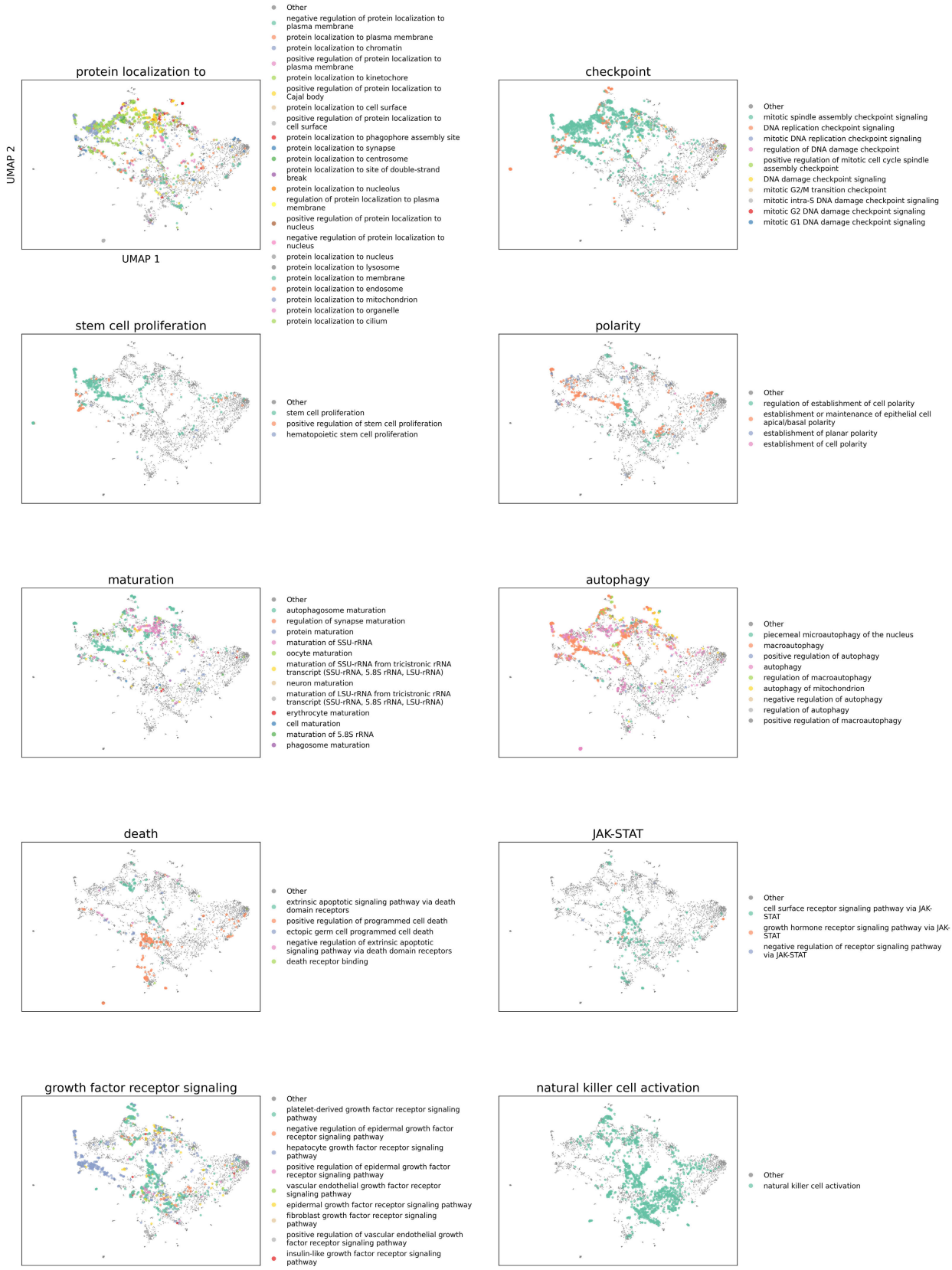
Supplementary Figure 22: **multiDGD SAE feature space UMAP.** The UMAP was computed with a minimum distance of 1, 10 neighbours, random seed 0, and a spread of 10. It is colored by feature type (left) and number of cells in which the feature is active (right).



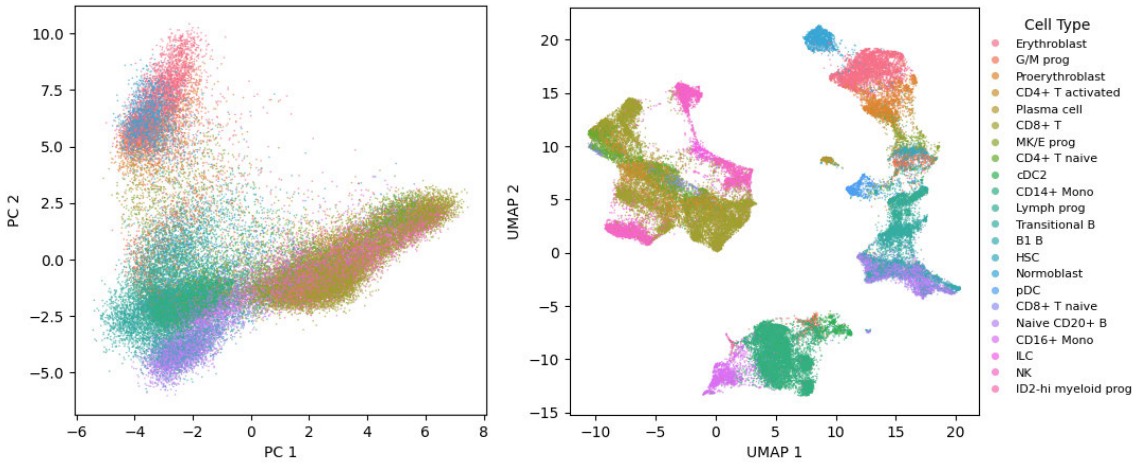
Supplementary Figure 23: **Frequency of individual GO terms.** The plot shows count histograms of all unique GO terms identified in the automated analysis colored by associated feature type (left) and GO term category (right).



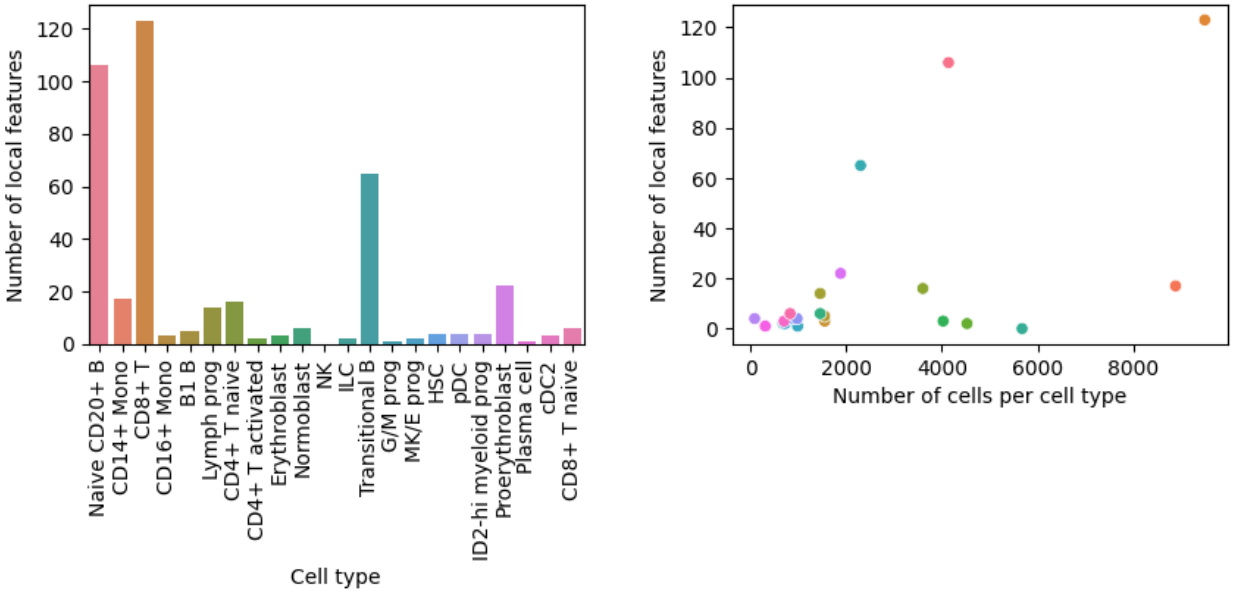
Supplementary Figure 24: **Single-cell SAE feature space UMAP indicating features from manual analysis.** Features 2306, 1238, 5205, and 1500 are highlighted by large colored dots and the feature id in black. All other features are depicted in blue.



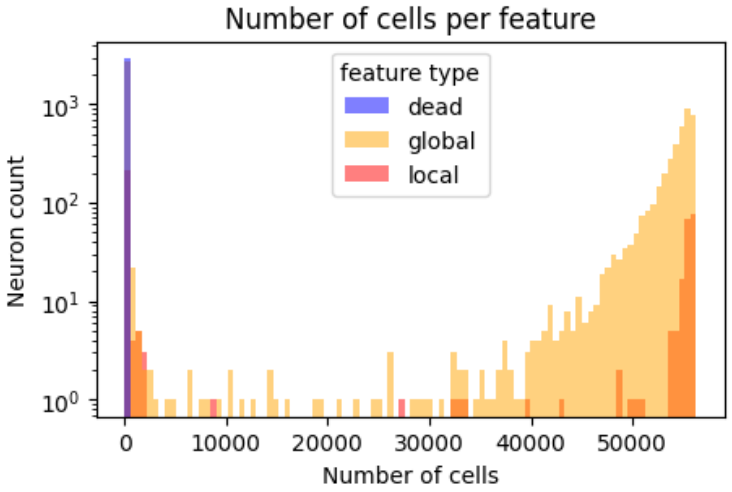
Supplementary Figure 25: **Probing multiDGD human bone marrow SAE feature space.** Words or concept snippets used for probing GO terms in the feature space are depicted in the title of each plot. Grey small background dots present all features ("other"). Colored, larger dots present all features in which the probing term was found. They are colored by the actual GO terms (legends to the right).



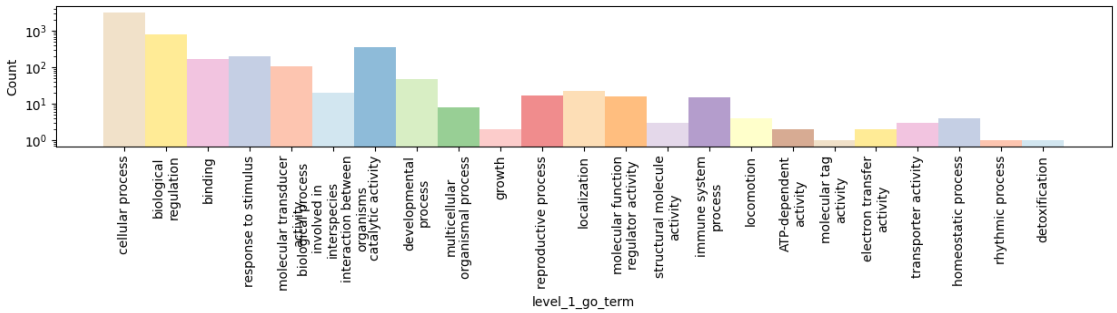
Supplementary Figure 26: **Geneformer embeddings of the human bone marrow data**. PCA on the left, colored by cell type (legend to the right). The right plot shows a UMAP with minimum distance 0.2, 20 neightbors, a spread of 0, and random seed 0.



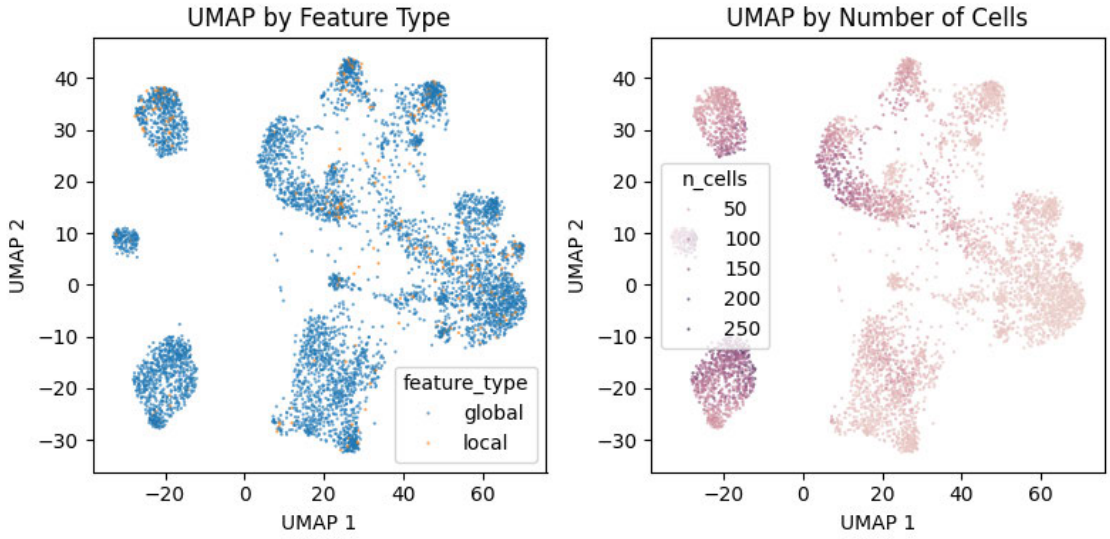
Supplementary Figure 27: **Distribution of local features among cell types in Geneformer embedding SAE.** The left shows a bar plot of the number of local features associated with each cell type. The right shows the number of local features plotted against the number of cells per cell type. Colors are the same as on the left.



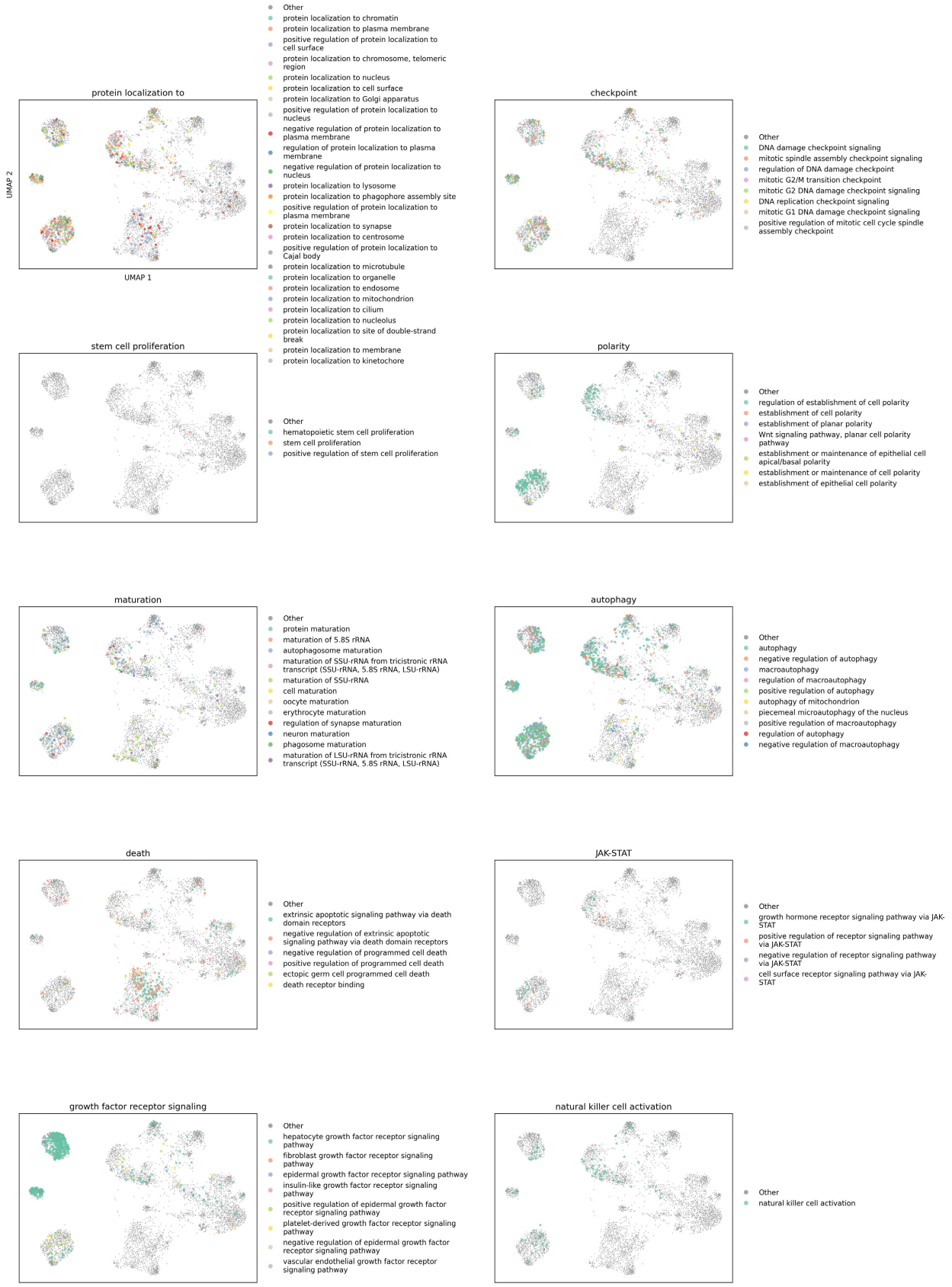
Supplementary Figure 28: **Histogram of features over cells of the Geneformer SAE trained on human bone marrow single-cell data.**



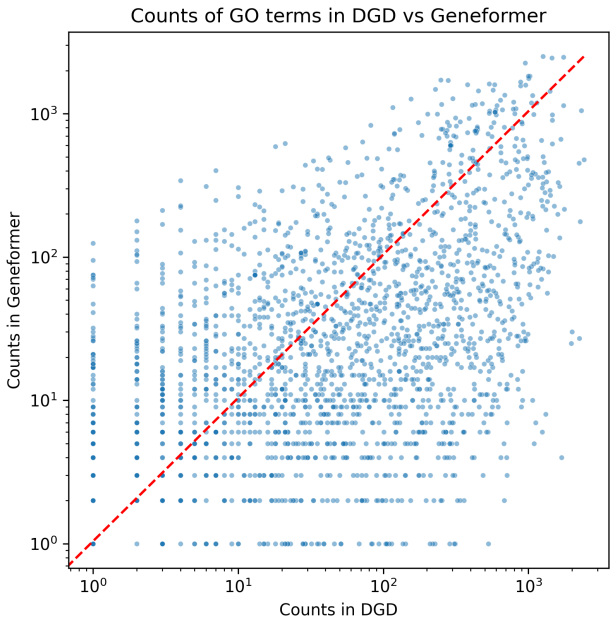
Supplementary Figure 29: **Frequency of individual GO terms.** The plot shows count histograms of all unique GO terms identified in the automated analysis colored by associated feature type (left) and GO term category (right).



Supplementary Figure 30: **Geneformer SAE feature space UMAP.** The UMAP is colored by feature type (left) and number of cells in which the feature is active (right).



Supplementary Figure 31: **Probing Geneformer human bone marrow SAE feature space.** Words or concept snippets used for probing GO terms in the feature space are depicted in the title of each plot. Grey small background dots present all features ("other"). Colored, larger dots present all features in which the probing term was found. They are colored by the actual GO terms (legends to the right).



Supplementary Figure 32: **Value counts in DGD and Geneformer SAE feature space per shared GO term.**