Research Square

Preprints are preliminary reports that have not undergone peer review.
They should not be considered conclusive, used to inform clinical practice,
or referenced by the media as validated information.

Knowledge or sparsity? Gene set masks in a single-cell VAE

George Miliarakis

contact@gmiliarakis.com

Independent Researcher https://orcid.org/0009-0009-3325-8386

Research Article

Keywords: single-cell RNA sequencing, variational autoencoder, gene set masks, knowledge-guided models, interpretable machine learning, sparsity, interferon-β response, Hallmark gene sets, Reactome pathways, linear decoder

Posted Date: November 2nd, 2026

DOI: https://doi.org/10.21203/rs.3.rs-11542814/v1

License: This work is licensed under a Creative Commons Attribution 4.0 International License.
Read Full License

Additional Declarations: The authors declare no competing interests.

Knowledge or sparsity?

Gene set masks in a single-cell VAE

George Miliarakis
Independent researcher
Amsterdam, Netherlands
contact@gmiliarakis.com

Abstract

Background Some knowledge-guided single-cell models assign gene sets to latent variables; each value then reports the activity of a known pathway. Earlier work shows that randomized versions of supervised pathway-informed networks, with the same share of zero weights, can match or outperform the curated ones. Whether the knowledge in the gene sets, or only their sparsity, explains how well unsupervised single-cell models capture a cell's response is not known.
Methods We trained a variational autoencoder (VAE) with a linear decoder on peripheral blood mononuclear cells from eight lupus patients, stimulated with interferon-β or left as controls. The decoder was either unrestricted (vanilla) or masked; the masks were the 50 Hallmark gene sets or co-expressed gene groups found in the data. Fully random, within-set random and expression-matched random masks with the same sparsity served as references. Variants were scored on held-out reconstruction and on how well the decoder reproduces the shift in expression on stimulation, with Spearman's rho to capture all genes without discrimination and Pearson's r to account for the strongest responders. We repeated the comparison with Reactome pathways and with a second dataset of interferon-β-stimulated cells (four benchmarks from two datasets).
Results Every mask worsened the fit relative to the vanilla model in all tested seeds. In all four benchmarks, gene set masks reproduced the shift in expression slightly better than the random masks with Spearman correlation and better with Pearson correlation. With the Hallmark sets on the lupus data, they recovered 10% (Spearman, 95% CI [8, 11]) and 56% (Pearson, 95% CI [49, 62]) of the gap between the fully random mask and the vanilla model. They also reached a lower test loss than the random masks in seven of eight comparisons. Co-expressed groups reproduced the shift better than the gene sets in all four benchmarks and fitted better than the Hallmark sets on the lupus data.
Conclusion In these data, random masks of the same sparsity decode the response almost on par with the gene set mask. Uncurated groups of co-expressed genes found in the data exceed the gain of the gene set mask over the random masks. The curated sets offset part of the fit lost by restricting the decoder. Only the curated sets assign each latent variable a biological name chosen before training.

1 Introduction

The state of a cell, such as its response to a stimulus or its phase in the cell cycle, is reflected in which genes it expresses and how strongly. RNA sequencing measures gene expression by counting the RNA transcripts of each gene. A tissue or blood sample consists of several cell types that may be in distinct states. Bulk sequencing pools the RNA of all the cells in a sample and can thus hide important information about the state of individual cells[1]. Single-cell RNA sequencing resolves gene expression per cell and has become a standard way to study tissues, development and disease[1]. Its

result is a high-dimensional table of transcript counts, with a row for every cell and a column for every gene, that is too large to analyse or visualize directly. Analysis usually represents every cell by a small number of latent variables, hidden factors that capture how its expression profile differs from that of other cells[2]. These variables serve to cluster cells, to infer trajectories, to compare conditions and to visualize the data[1,3]. Abbreviations and terms that may be unfamiliar are listed in the Abbreviations and the Glossary at the end of the paper.

Variational autoencoders (VAEs)[4] infer such latent variables from the transcript counts of every gene in a cell. A VAE comprises an encoder that maps the transcript counts of a cell to its latent variables, while a decoder maps the latent variables back to expected counts. The latent variables of an unconstrained model carry no names. Interpreting one means identifying afterwards which genes depend on it and testing whether they are enriched for known functions or pathways. A linear decoder simplifies this: the expected expression of a gene is a weighted sum of the latent variables, so the weight of each latent variable on each gene is directly visible[5].

Knowledge-guided models try to assign the latent variables a meaning before training; the output is then interpretable in biological terms. They draw on prior knowledge, here in the form of gene sets: lists of genes known to act together in a biological process or pathway[6]. This knowledge can enter a model in three ways[7]: as input features, such as gene set enrichment scores; in the architecture, as sparse layers that mirror gene sets; or as a graph of gene interactions. This study tests the architecture route.

In this route each latent variable is assigned one gene set. A mask fixes to zero the decoder weights from the latent variable to all genes outside its set; the latent variable then reaches only the genes of its set. Its value in a cell is then meant to represent the activity of that gene set. The decoder is called sparse, since most of its weights are zero. Such models are designed to be interpretable, since each latent variable carries the name of its gene set. They are expected to require fewer cells, because the prior knowledge narrows the space of models to search[7].

Several published models are built this way. VEGA[8] and expiMap[9] are single-cell VAEs whose decoder is masked by gene sets, and pmVAE[10] assigns one sub-network to each pathway. Others employ gene sets in related ways. f-scLVM[11], PLIER[12], Spectra[13] and MOFA-FLEX[14] are factor models that explain the expression of a cell by a few factors, here guided by gene sets. KPNN[15] is a supervised neural network whose connections follow a biological network, and Moullet et al.[16] present Tripso, a self-supervised transformer model that learns a separate embedding for each of several predefined gene programs.

In a study of 29 pathway-informed neural networks, randomized versions with the same architecture and sparsity matched or outperformed the curated ones on supervised prediction tasks[17]. VEGA[8] and expiMap[9] report masked decoders. Both are applied to peripheral blood mononuclear cells (PBMCs) stimulated with interferon-β. Interferons are signalling proteins of antiviral defence that switch on hundreds of interferon-stimulated genes[18,19]. Neither VEGA nor expiMap reports a random gene set control (a mask in which gene memberships are randomized). MOFA-FLEX[14] contains the closest controls. One is a model with uninformed sparse factors. The other swaps 30% of the genes of each set with random non-members and tests whether the original programs are recovered. It is not known whether the knowledge in the gene sets improves the fit and the decoded response of unsupervised single-cell models beyond the sparsity that the mask imposes. A search for unsupervised single-cell models that compare curated gene sets with size-matched random gene sets found none (supplementary, Literature search).

Several gaps remain in this literature. An improvement over a model without a mask could stem from the sparsity or from the knowledge in the sets. Gene sets are not independent of expression data

either. In bulk tumour data, a curated panel of 16 signatures—sets of genes whose joint expression marks a cell state or process—and, in an upper-bound sensitivity analysis in breast cancer, 45 of the 50 sets of the Hallmark collection[20] concentrate more of their variance in the leading principal components of expression (the directions along which expression varies most[21]) than size-matched random sets show[22]. The Hallmark collection, from the Molecular Signatures Database, consists of 50 gene sets that each represent a well-defined biological state or process, and its sets were refined for coherent expression. A Hallmark mask could thus outperform a random one purely because its genes are co-expressed, and it is not known whether pre-specified sets add anything beyond that.

Finally, it is not known whether an improvement over random masks generalizes over other kinds of datasets or gene sets. The Hallmark sets were refined from expression data, while Reactome[23] is a curated database of biological pathways. The result may therefore depend on the kind of knowledge. The dataset employed for this purpose, interferon-β-stimulated PBMCs of Kang et al.[24], is a single experiment in which the stimulation is confounded with the sequencing run: each condition was sequenced in one run, so stimulation cannot be told apart from technical differences between runs.

1.1 Research questions

To address these gaps we formulated one core question and three follow-ups that each vary one element of its design. Q1 asks whether the knowledge or the sparsity explains the fit. The follow-ups narrow down the conclusions that can be drawn from it.

- **Q1** Does the knowledge in the gene sets, or only the sparsity they impose, explain how well the model captures the response?
- **Q1a** Can pre-specified sets add anything beyond groups of co-expressed genes found in the data?
- **Q1b** Is the result reproduced with a second collection of gene sets?
- **Q1c** Is the result replicated in a second dataset?

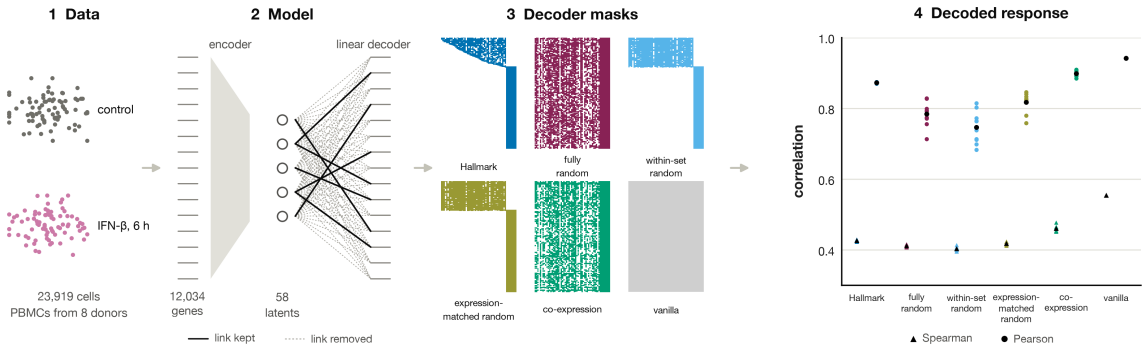


Figure 1. Overview of the study. (1) Data: 23,919 PBMCs from 8 donors, control and 6 h after interferon-β stimulation (Kang et al.). (2) Model: a VAE whose linear decoder maps 58 latents to 12,034 genes; a mask retains (solid) or removes (dashed) each link. (3) Decoder masks: the six masks as trained with seed 0, one dot per retained link, with the genes sorted by Hallmark membership. (4) Decoded response on the Kang et al. data with the Hallmark masks: the Spearman (triangles) and the Pearson correlation (circles) of the decoded with the observed shift in expression on stimulation.

2 Methods

2.1 Study design

Six variants of one model share the architecture and training setup and differ in the decoder mask (Figure 1, panel 3, Figure 4 and Table 1). The vanilla model and the random masks are baselines and controls[17,25,26]. Each variant is fit on four benchmarks (Table 2).

Table 1. The variants. Each differs from the others only in the decoder mask, and the last column says what the comparison with it tests.

Variant	Decoder mask	Tests
Vanilla (no mask, linear-decoder VAE, LDVAE[5])	none	reference without knowledge
Gene set mask	the 50 sets of the collection (Hallmark or Reactome)	knowledge + sparsity
Fully random	the gene sets with the labels of all genes randomly permuted	sparsity only; free latents can reach set genes
Within-set random	the gene sets with the labels of set genes randomly permuted among set genes	sparsity only; free latents reach the same genes as in the gene set mask
Co-expression	groups of co-expressed genes found in this data, same sizes	data-derived structure
Expression-matched random (Q1)	the gene sets with the labels of set genes randomly permuted among set genes of similar expression level and stimulation effect	sparsity only; every set retains its genes of each response strength

Each question varies one element of the main experiment, the Kang et al. data with the Hallmark sets (Figure 2). The comparison is the seed-paired difference between the gene set mask and another mask on four measures (Table 4), corrected for multiple testing (Statistics). Q1 compares the gene set mask with the fully random and within-set random masks. Q1a compares it with the co-expression mask. The expression-matched mask is a stricter control for Q1 (supplementary S3). Q1b compares the Reactome with the Hallmark benchmark within each dataset. Q1c compares the Kang et al. data with the Theis et al. data. Supplementary S7 documents the steps that apply knowledge or assume

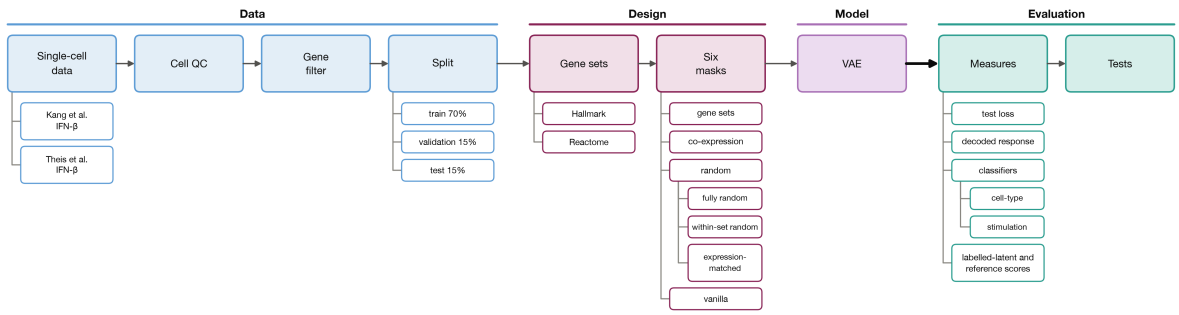


Figure 2. Pipeline. The analysis in four stages. Data: the single-cell data of Kang et al. and Theis et al. (interferon-β), cell quality control, the gene filter and the split into training, validation and test cells. Design: the gene sets (Hallmark or Reactome) and the six masks. Model: the VAE. Evaluation: the measures (test loss, decoded response, cell-type and stimulation classifiers, and the labelled-latent and reference scores) and the tests.

a property of the cells.

Table 2. The four benchmarks. The dataset, donors, cells, genes and stimulus (IFN-β, interferon-β) of each dataset, then the collection and the expected-response set of each benchmark.

Data	Donors	Cells	Genes	Stimulus	Collection	Expected-response set
Kang et al.[24]	8 (lupus)	23,919	12,034	IFN-β, 6 h	Hallmark (main)	interferon-α response
					Reactome	interferon α/β signaling
Theis et al.[27]	12 (healthy)	28,800	11,340	IFN-β, 24 h	Hallmark	interferon-α response
					Reactome	interferon α/β signaling

2.2 Data

The main dataset is from Kang et al.[24] (Gene Expression Omnibus accession GSE96583, batch 2). It contains PBMCs from 8 patients with systemic lupus erythematosus (SLE), called donors below, cultured for 6 h without (control) or with interferon-β (stimulated). Each condition was pooled and sequenced on the 10x Genomics platform. Every donor appears in both conditions.

2.2.1 Human Cytokine Dictionary

The second dataset is the Parse Biosciences dataset "10 Million Human PBMCs in a Single Experiment"[28], described as the Human Cytokine Dictionary by Theis et al.[27]. It contains PBMCs from 12 healthy donors on twelve 96-well plates, cultured for 24 h with one of 90 cytokines or with phosphate-buffered saline (PBS) as the control. Interferon-β is compared with the control; every donor appears in both conditions.

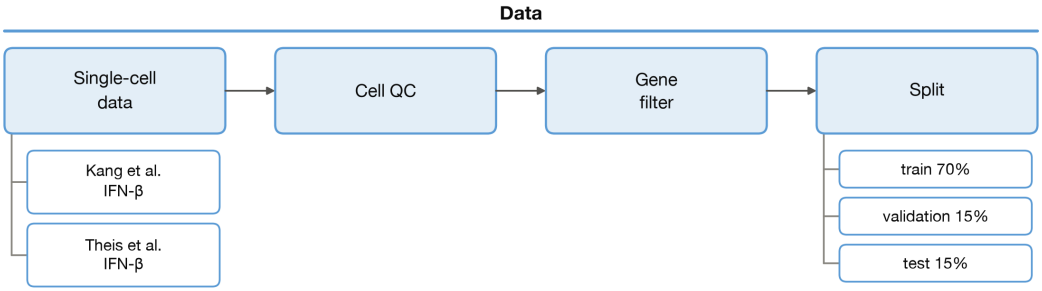


Figure 3. Data stage of the pipeline. The single-cell data of each stimulation pass cell quality control and a gene filter and are split into training, validation and test cells.

2.2.2 Preprocessing

Kang et al. data The deposited matrices contain 29,065 cells and 35,635 genes in two runs. Cells were retained if demuxlet called them singlets (24,679; the rest are 3,169 doublets and 1,217 ambiguous droplets) and if they carry a cell-type label (24,673). Cells were then filtered for quality

(QC; Figure 3) per condition and cell type, removing those beyond 5 median absolute deviations in log total counts or log detected genes, or in the share of counts in the top 20 genes[1,3]. The filter removed 754 of 24,673 cells (3.1%) and retained 23,919. Megakaryocytes lost 20.8% of their cells and are excluded from all scores. No ambient-RNA correction or mitochondrial filter was applied.

Theis et al. data Cells were read from the published archive until each donor × condition group reached 1,800. The raw counts were used, without the Harmony integration of the authors[28]. The same quality filter was applied. For interferon-β, 45,332 cells were read, 369 (0.8%) were removed and 44,963 retained. At most 1,200 cells per donor × condition group were retained, for 28,800 cells (12 donors × 2 conditions × 1,200).

Genes Genes detected in at least 20 cells were retained without selecting highly variable genes. This leaves 12,034 of the 35,635 genes in the Kang et al. data. In the Theis et al. data, 23,469 genes are detected in at least 20 of the 28,800 cells, and 11,340 of them are among the 12,031 distinct gene symbols of the Kang et al. data. Gene symbols were matched through Ensembl stable identifiers to current symbols of the HUGO Gene Nomenclature Committee (HGNC)[29], which renamed 1,256 symbols since 2017; the 427 genes that HGNC no longer lists retain their 2017 symbol.

Split Cells were split within every donor × condition × cell type group into 70% training, 15% validation and 15% test (at least one cell each in validation and test). The Kang et al. data form 128 groups (8 donors × 2 conditions × 8 cell types; smallest 4 cells) with 16,749 training, 3,585 validation and 3,585 test cells. The Theis et al. data form 388 groups (smallest 1 cell) with 20,103 training, 4,364 validation and 4,333 test cells. Training cells fit the model, validation cells stop the training and fit the two classifiers, and test cells yield all scores. The co-expressed groups and the expression bins of the random masks are built from the training cells only. A cell type is scored when the test split contains at least 20 cells of both conditions, which applies to 13 of 18 types in the Theis et al. data.

Model input The counts of each cell are normalized to the median total count of the training cells and log-transformed. A one-hot code of the donor is appended as encoder input. The encoder receives 12,042 values per cell for the Kang et al. data (12,034 genes and 8 donors) and 11,352 for the Theis et al. data (11,340 genes and 12 donors).

2.3 Gene sets and masks

Hallmark sets The Hallmark mask draws on the Hallmark collection[20] from the Molecular Signatures Database (MSigDB) v2024.1. It contains 50 sets of at least 12 genes (median 117.5) and no near-duplicate sets: the largest overlap has a Jaccard index of 0.34. It contains an unambiguous target for this data, the interferon-α response, with 95 of its 97 genes present. Among the 50 sets, 5,568 of 7,322 memberships (76%) remain after the gene filter and symbol matching. The Hallmark sets were designed as inputs for gene set enrichment analysis[20] and were built from founder sets by automated overlap analysis and manual curation.

Random masks Three random masks, serving as negative controls, replace the Hallmark sets with equally-sized (Figure 4). Each is built 10 times. The run with seed k applies version k. Genes are assigned to 25 bins of equal size (481–482 genes) by their mean count in training cells. The fully random mask permutes gene labels within each bin. Each gene, irrespective of if it is in a set or not, assumes the set memberships of another gene from its bin. A full shuffle replaces 98% of set members with genes of the same expression level. Every set keeps its size and its overlaps with other sets. The within-set random mask permutes only the genes that are in at least one set, among themselves and within the same 25 bins. Genes that belong to no set (unassigned genes) stay unassigned, and the

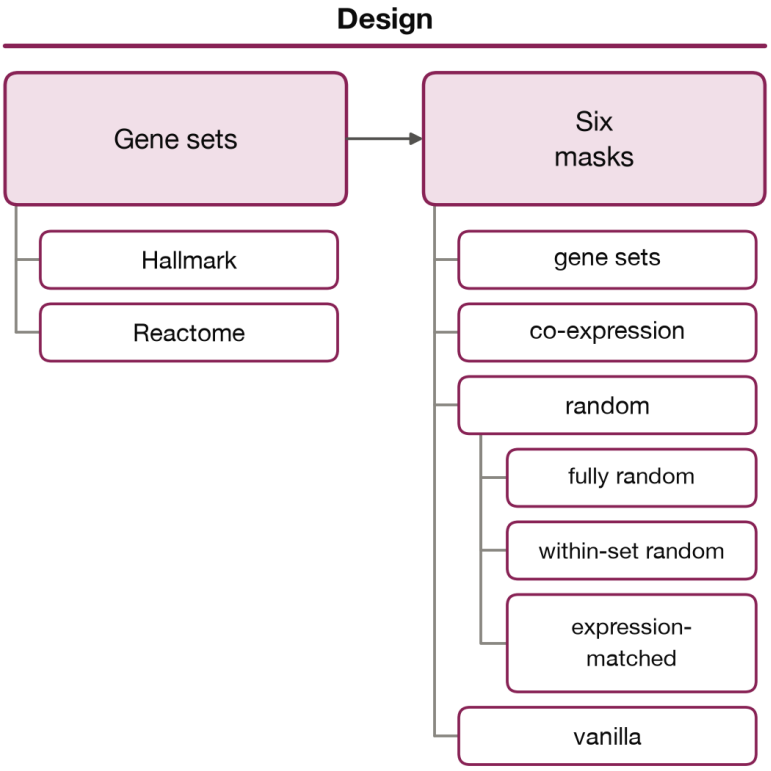


Figure 4. Design stage of the pipeline. The six masks are built from the two gene set collections.

permutation replaces 95% of set members.

Reachable genes The free latents (Model) reach only the unassigned genes. The fully random mask shuffles all 12,034 genes. Set genes and unassigned genes therefore trade places. Depending on the seed, 43–59 of the 95 interferon-α genes then lie among the genes that no labelled latent can reach. The within-set random mask shuffles only the 3,187 Hallmark set genes among themselves. The unassigned genes stay as in the gene set mask. Every interferon-α gene remains reachable by a labelled latent. A difference from the Hallmark mask thus arises from the grouping of genes into sets.

Expression-matched mask The expression-matched random mask permutes the genes in sets among themselves, as the within-set random mask does, but within bins that cross expression level with stimulation effect. The 25 expression bins are merged five at a time into 5, and each is crossed with 5 bins of stimulation effect. The stimulation effect is the rank of $|t|$ of stimulated against control training cells. The 5 bins are cut at ranks 250, 1,000, 3,000 and 6,000. Every set preserves its overlaps and the distribution of its genes over response strength, including among the 250 strongest. The mask requires the stimulation labels of the training cells. It serves Q1 (supplementary S3).

Co-expression mask The co-expression mask replaces the Hallmark sets with groups of co-expressed genes found in this dataset and tests whether pre-specified sets add anything beyond them. It is built from the training cells only, without condition or cell-type labels. For each seed, 50 seed genes are drawn at random from the 6,753 genes detected in $\geq 1\%$ of training cells, and group j is seed gene j plus the n − 1 genes most positively correlated with it, where n is the size of Hallmark set j (12–197 genes). A group matches the size of the set it replaces and carries none of its biology, and the groups are built independently and may share genes. They are more coherent than the Hallmark

sets, with a median within-set correlation of 0.096 against 0.003, and in every seed the group most raised by IFN-β is enriched for Hallmark interferon genes (19–89% of its genes, against a base rate of 1.8%). The full recipe is in supplementary S5.

Reactome sets The comparison is repeated with Reactome[23] in place of Hallmark, drawing on the 1,736 sets of the MSigDB v2024.1 collection C2 CP:REACTOME. Reactome is a curated database of biological pathways. It tests the same question with knowledge of another kind. The 50 sets are chosen only from names, sizes, overlaps and the pathway hierarchy, without any use of expression or stimulation.

Reactome selection Five Reactome sets occupy the slots of the five Hallmark sets that are the expected responses of a stimulus: interferon alpha/beta, interferon gamma, TNF, interleukin-6 family and interleukin-2 family signaling. They are narrower than the Hallmark sets, with 62, 80, 52, 14 and 36 genes against 95, 192, 188, 74 and 167. The other 45 slots hold sets drawn at random from those that match the size of their Hallmark set (within 10%), overlap every chosen set by a Jaccard index below 0.34 (the largest overlap between two Hallmark sets), have at least 50% of their genes among the 12,034 QC genes, lie outside the Disease branch of the pathway hierarchy and contain none of INTERFERON, IFN, IRF, TNF, IL-6 or IL-2 in their name. Sets that cannot be matched to a pathway by name are excluded only when their name contains SARS, DEFECT, INFECTION, HIV, DISEASE, CANCER or MUTANT.

Selected draw Selection seeds 0, 1, 2 and so on were evaluated in order. The first draw with as many trainable decoder weights as the Hallmark mask (within 1%) was selected. It is seed 2, with 76,579 trainable weights against 76,344. Of its 6,501 memberships, 5,003 remain after the gene filter and symbol matching (3,087 genes in sets, 8,947 unassigned), against 5,568 memberships and 3,187 genes in sets for Hallmark. The random masks and the co-expressed groups are built from the Reactome sets by the recipes above, and the interferon α/β set is the target latent.

2.4 Model

The six variants share one VAE architecture[4] (Figure 5). The numbers below are for the Kang et al. data, with the Hallmark sets unless stated. The encoder (one hidden layer of 128 units, a rectified linear unit (ReLU) activation, dropout 0.1) reads the 12,034 genes after a shifted-log transform and a one-hot code for the 8 donors. It outputs a mean and a variance for each of the 58 latents (50 named, 8 free). The decoder is linear. The equations follow; the full architecture is in supplementary S6.

2.4.1 Model equations

The encoder input is the shifted logarithm[30] of the counts. Each count x_g is divided by the total count of the cell ℓ and multiplied by the median total m of the training cells:

$$\tilde{x}_g = \ln\left(1 + x_g \cdot m/\ell\right)$$

The encoder maps $\tilde{x}$ and the donor code $\mathbf{s}$ (a column of $\mathbf{S}$ in Figure 5) to the posterior $q(z \mid x, \mathbf{s})$, with a mean $\bar{\mu}$ and a log variance $\log \sigma^2$ for each latent. Training draws a point z from it[4]; the evaluation uses its mean.

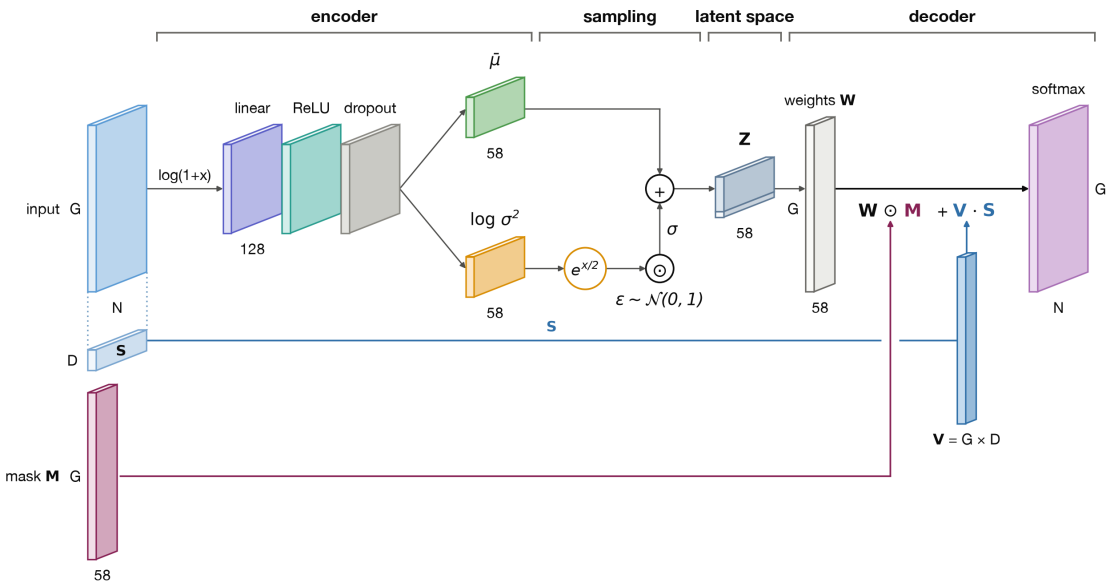


Figure 5. VAE architecture. Data pass through the layers as matrices with one column per cell. The height of a block suggests its number of units and its depth the number of cells; for the parameters **W**, **M** and **V** the depth is the number of columns. Sizes are not to scale. N is the number of cells, G the number of genes and D the number of donors. **S** ($D \times N$) is the one-hot donor code and is joined to the counts as input of the encoder. The encoder outputs, for each of the 58 latents (50 named, 8 free), a mean $\bar{\mu}$ and a log variance $\log \sigma^2$. The latent matrix **Z** ($58 \times N$) holds $z = \bar{\mu} + \sigma \odot \epsilon$ for each cell, with $\sigma = e^{(\log \sigma^2)/2}$ and ϵ from a standard normal. The decoder multiplies **Z** by the masked weights $\mathbf{W} \odot \mathbf{M}$, adds the donor offsets $\mathbf{V}\,\mathbf{S}$ (**V** is $G \times D$) and applies a softmax over genes. The mask removes decoder weights only. The encoder reads all genes and the loss covers all genes.

The decoder maps z to shares π_g of the counts of the cell[5,31]. A softmax over genes makes the shares sum to 1:

$$\pi = \mathrm{softmax}\big((\mathbf{W} \odot \mathbf{M})\, z + \mathbf{V}\,\mathbf{s}\big)$$

The observed count follows a negative binomial (NB) distribution with expectation $\ell\,\pi_g$ and a dispersion θ_g for each gene[31,32]:

$$x_g \sim \mathrm{NB}\big(\mu_g = \ell \cdot \pi_g,\ \theta_g\big)$$

The loss is the NB negative log-likelihood plus the Kullback-Leibler (KL) divergence of the posterior of each cell from a standard normal prior. The KL divergence measures how far one probability distribution lies from another[33].

$$\mathcal{L} = -\log p_{\mathrm{NB}}(x \mid \mu, \theta) + \beta \cdot \mathrm{KL}\big(q(z \mid x, \mathbf{s}) \parallel \mathcal{N}(0, I)\big)$$

where, for one cell:

- $\mathcal{L}$: the loss per cell, averaged over each batch. With $\beta = 1$ it is the negative evidence lower bound (ELBO)[4].
- $-\log p_{\mathrm{NB}}(x \mid \mu, \theta)$: the reconstruction term, the negative log-likelihood summed over genes.
- x_g: the raw count of gene g. $\ell = \sum_g x_g$ is the total count of the cell.

- $\mu_g = \ell\,\pi_g$: the expected count of gene g.
- π_g: the share of gene g from a softmax over genes; the shares are positive and sum to 1.
- $\mathbf{W} \in \mathbb{R}^{G\times 58}$: the latent $\to$ gene weights, with G the number of genes (12,034 in the Kang et al. data). $\mathbf{M} \in \{0,1\}^{G\times 58}$ is the mask, with 1 where latent k may reach gene g and 0 elsewhere. The product $\odot$ is elementwise.
- $\mathbf{V} \in \mathbb{R}^{G\times D}$: the donor offsets, which also set the baseline of each gene, with D the number of donors (8 in the Kang et al. data).
- θ_g: the dispersion of gene g with $\mathrm{Var}(x_g) = \mu_g + \mu_g^2/\theta_g$.
- β: the KL weight. It rises linearly from 0 to 1 over the first $\approx$ 38 epochs (KL warm-up[34]) and then remains at 1.
- $\mathrm{KL}\big(q \,\|\, \mathcal{N}(0,I)\big)$: the KL divergence of the posterior from the standard normal prior, summed over latents. For the diagonal normal posterior it equals $\frac{1}{2}\sum_k \left(\bar{\mu}_k^2 + \sigma_k^2 - \log\sigma_k^2 - 1\right)$[4].
- $z \in \mathbb{R}^{58}$: the latent values of the cell, one column of the latent matrix $\mathbf{Z}$ $(58 \times N)$ in Figure 5. Training draws them as $z = \bar{\mu} + \sigma \odot \epsilon$ from $q(z \mid x, \mathbf{s}) = \mathcal{N}\big(\bar{\mu},\ \mathrm{diag}\,\sigma^2\big)$, with $\sigma = \exp(\frac{1}{2}\log\sigma^2)$ and $\epsilon \sim \mathcal{N}(0,I)$.
- $\mathbf{s} \in \{0,1\}^D$: the donor one-hot code of the cell, one column of the donor matrix $\mathbf{S}$ $(D \times N$ for N cells) in Figure 5.
- $\tilde{x}_g$: the encoder input, the shifted logarithm of the count of gene g defined above.

The mask sets the number of trainable decoder weights (Kang et al. data: 12,034 genes and 58 latents). The vanilla VAE trains all 697,972. The gene set mask and the random masks train 76,344 weights with Hallmark (5,568 set memberships plus 8 free latents × 8,847 unassigned genes) and 76,579 with Reactome (5,003 plus 8 × 8,947). The co-expression mask trains 75,976–83,136 weights by seed with Hallmark and 77,251–83,827 with Reactome. The Theis et al. data include 11,340 genes and 12 donors, so the vanilla model holds 657,720 decoder weights. The gene set masks train 70,956 (Hallmark) or 71,116 (Reactome) of them.

The decoder is linear in every variant. In the masked variants, latents 1–50 reach only the genes of their set. The mask reserves the unassigned genes for the 8 free latents. They reach 8,847 genes (74%) in the Hallmark mask and, in the co-expression mask, the genes in no group. Donor is a covariate in encoder and decoder. Condition is not an input; the latents find the stimulation themselves. The likelihood is negative binomial on raw counts, with the observed total count as library size and no zero inflation[32]. All variants share the training settings (Table 3).

Table 3. Training settings. All variants are trained with the same settings.

Setting	Value
Optimizer	Adam[35]
Learning rate	1e-3
Batch size	128
KL warm-up	β rises linearly from 0 to 1 over 10,000 steps at batch 64, counted in cells processed: ≈ 38 epochs for the Kang et al. data and ≈ 32 for the Theis et al. data
Stopping	early stopping on validation loss with a patience of 3 epochs[36], counted after the KL warm-up; any decrease counts as an improvement; at most 2, 000 epochs (no run of the main experiment reached it)
Encoder width	128, chosen on the vanilla VAE
Seeds	10 per variant; each seed sets the initial weights, the batches and (for the random and co-expression masks) the mask. The train/validation/test split is fixed (seed 0)

Setting	Value
Validation loss	computed with the same latent draws at every epoch. Early stopping then follows the model and not the sampling noise
Software	versions in Table 5; the last digits of the losses differ between runs on the central processing unit (CPU) and on the graphics processing unit (GPU)

2.5 Evaluation

Four main measures are computed for every run on the test cells (Figure 6; Table 4): test loss, decoded response, a cell-type classifier and a stimulation classifier. Spearman and Pearson correlation score the decoded response. Each run yields five scores. An effect on the decoded response is stated as a direction when it holds for both correlations; its size is reported for each. None of the measures depends on which latent represents the interferon response. The two classifiers are linear probes of the latent means. Validation cells also select the best latent and the best principal component of supplementary S1. The AUROC, the area under the receiver operating characteristic curve, is the probability that a random stimulated cell scores above a random control cell (0.5 is chance, 1 is perfect separation). Megakaryocytes are mostly platelets. They are included in training but excluded from all scores. The interferon-α latent score and the reference scores without a VAE are described in the supplementary results. The score of the labelled latent requires an expected response, the Hallmark interferon-α response set (supplementary S1). The labelled latent is oriented without labels, by the sign of its mean weight on its own genes.

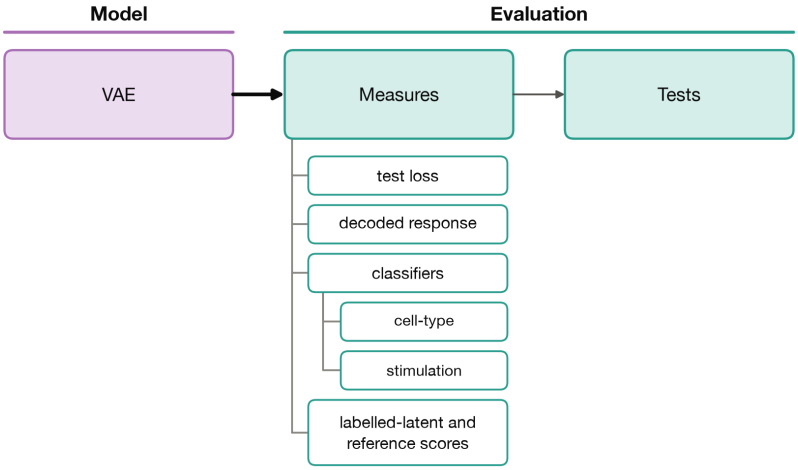


Figure 6. Model and evaluation stages of the pipeline. The trained VAE is scored by the four measures, and the tests compare the variants.

Table 4. The four main measures. What each measure asks and how it is computed.

Measure	Asks	Computation
Test loss	How well does the model reconstruct unseen cells?	Negative log-likelihood per test cell plus the KL term at weight 1 (natural logarithm, in nats), the training objective; lower is better
Decoded response	Does the decoder reproduce the stimulation effect?	The decoder returns the expected share of every gene for a test cell from its latent means and donor. For each cell type, the mean decoded log expression (the encoder's shifted-log scale) of stimulated minus control test cells is correlated across the genes detected in at least 1% of training cells (6,753 in the Kang et al. data) with the observed shift. The correlations are averaged over the scored cell types (7 for the Kang et al. data, 13 for the Theis et al. data). The Spearman correlation[37] scores all genes and counts each gene once. The Pearson correlation weights each gene by the size of its shift and therefore scores the large responses. The labels enter only through the observed shift. A further Spearman value is computed from the 95 interferon-α genes
Cell-type classifier	Do the latents represent cell type, a label unseen in training?	Multinomial logistic regression on standardized latent means, fitted on validation cells; balanced accuracy on test cells. Descriptive variants: 10, 50 and 200 validation cells per type, and the AUROC between CD4 and CD8 T cells and between CD14+ and FCGR3A+ monocytes
Stimulation classifier	Does the representation reflect the stimulation response?	Logistic regression on all 58 latents fitted on validation cells; AUROC[38] per cell type on test cells

2.5.1 Noise bound

Definition The shift in expression on stimulation is estimated from a finite number of test cells and is thus noisy. A decoder that reproduced the true shift would still not correlate perfectly with the observed one. Reliability ceilings are used in neural encoding[39,40]. Split halves of the data are an empirical reference in single-cell perturbation prediction[41–43]. Let T be the true shift across genes and $Y = T + \varepsilon$ the observed shift, with noise ε uncorrelated with T. The reliability

$$\rho = \frac{\mathrm{Var}(T)}{\mathrm{Var}(Y)}$$

is the share of the variation between genes that is real. The correlation of a prediction X with Y is

$$r(X, Y) = r(X, T)\sqrt{\rho} \leq \sqrt{\rho}.$$

No prediction exceeds $\sqrt{\rho}$, the noise bound[37,44]. An exact decoder, with $X = T$, reaches it.

Estimation The reliability is estimated from the shift measured in two random halves of the test

cells (Algorithm 1)[40,43]. The Spearman-Brown formula corrects for the halves being noisier than the full set[45,46]:

$$\rho = \frac{2\, r_{\text{half}}}{1 + r_{\text{half}}}.$$

The halves are correlated with the same correlation as the score. The Spearman score and the Pearson score each receive their own bound. The bound is the square root of the mean reliability over the 50 splits.

Algorithm 1. Noise bound of the decoded response.

Require: test cells of each scored cell type c, stimulated and control; genes G detected in at least 1% of the training cells; number of splits $K = 50$
Ensure: noise bound b
for $k = 1, \ldots, K$ **do**
  for each scored cell type c **do**
    split the stimulated and the control test cells of c at random into halves A and B
    $\Delta_A \leftarrow$ mean shifted-log expression of the stimulated cells in A minus that of the control cells in A; likewise Δ_B
    $r_c \leftarrow$ correlation of Δ_A and Δ_B across the genes G
  end for
  $r_k \leftarrow$ mean of r_c over the cell types
  $\rho_k \leftarrow 2r_k/(1 + r_k)$
end for
$\rho \leftarrow$ mean of ρ_k over the splits
return $b = \sqrt{\rho}$

Assumptions The derivation is exact for the Pearson correlation under additive noise that is uncorrelated with the true shift and with the prediction. For the Spearman correlation, Algorithm 1 yields an empirical estimate. The decoder reconstructs the test cells whose shift is the target. Noise shared with those cells can lift a score above the bound. The score is a mean of correlations per cell type, whereas Algorithm 1 averages the half-sample correlations over the cell types before the correction. A per-type correction lowers the bound by 0.02–0.03. The donors of the test cells also appear in the training cells.

2.5.2 Statistics

Tests Each variant is run with 10 seeds, paired by seed across variants. The replicate is the training run. With the split fixed, the tests concern variation in training and in the random masks. For a measure and two variants, $d_i = a_i - b_i$ is the difference of seed i. The p-value is the exact two-sided Wilcoxon signed-rank test[47] of $d_1, \ldots, d_{10}$ against 0. Its smallest value with 10 pairs is 0.002.

Estimate and CI The effect is the Hodges-Lehmann estimate[48], the median of the 55 Walsh averages,

$$\hat{\Delta} = \text{median}\left\{\tfrac{d_i + d_j}{2} : 1 \le i \le j \le 10\right\}.$$

Its exact CI follows from the distribution of the Wilcoxon signed-rank statistic. Descriptive summaries are the mean and standard deviation over the seeds.

Multiple testing The 15 tests of a benchmark (five scores, gene set mask minus each other mask) are corrected together. The 5 tests of the expression-matched control (supplementary S3) and the 10 tests on the interferon-α score (supplementary S1) form separate groups. For m tests in a group, with ordered p-values $p_{(1)} \leq \cdots \leq p_{(m)}$, Holm's procedure[49] gives the adjusted p-values

$$\tilde{p}_{(k)} = \max_{j \leq k} \ \min\left(1,\ (m - j + 1)\, p_{(j)}\right)$$

and controls the family-wise error rate at 0.05. CIs use the Bonferroni level[50], $1 - 0.05/m$, for a simultaneous coverage of 95%. "Significant" means an adjusted p below 0.05 within its group. Holm's procedure cannot reject at the smallest exact p of 0.002 when more than 25 tests are corrected together, and the 90 tests therefore form separate groups. With 15 tests the smallest adjusted p is 0.029.

Shares The share of the gap between the fully random mask and the vanilla VAE that the gene set mask recovers is the ratio of two Hodges-Lehmann estimates. Its 95% CI is bounded by the 2.5th and 97.5th percentiles of 2,000 resamples of the seeds, conditional on the fixed split.

Equivalence The Hallmark and co-expression masks are equivalent on the interferon-α score when both one-sided Wilcoxon signed-rank tests of $d + \varepsilon > 0$ and $d - \varepsilon < 0$ reject at 0.05. The margin $\varepsilon = 0.005$ AUROC is half a percentage point in the probability that a stimulated cell outranks a control cell.

Symmetry check The Wilcoxon signed-rank CI assumes symmetric differences. A sign-test CI of the same level needs no such assumption. It is consistent with the Wilcoxon signed-rank CI on excluding zero in all 60 benchmark tests (4 benchmarks × 15) and in 19 of the 20 tests of the expression-matched control. The exception is the stimulation classifier in the Kang et al. data with Reactome sets. Among the 10 comparisons of supplementary S1, one differs (the highest-scoring latent of the within-set random mask minus that of the vanilla VAE).

2.6 Software and reproducibility

All analyses ran in Python 3.13.15 (packages and versions in Table 5) on an Apple M1 Pro with 16 GB of memory under macOS 15.7. The VAE was trained on the Apple GPU through PyTorch[51].

Table 5. Software and versions. Software used, with version and role.

Package	Version	Used for
PyTorch[51]	2.14.1	training the VAE on the Apple GPU
scanpy[3]	1.12.4	quality control and gene filtering
AnnData[52]	0.13.4	count matrices and their annotations
NumPy[53]	2.5.3	array computation
pandas[54]	3.0.6	tables of scores and tests
SciPy[55]	1.18.1	Wilcoxon signed-rank tests and other statistics
scikit-learn[56]	1.9.1	the classifiers and the principal components
Matplotlib[57]	3.11.2	figures
h5py	3.16.0	reading the Theis et al. archive
scikit-misc	0.5.2	the highly variable gene comparison

The scripts run in a fixed order and produce every number in the paper. A seed fixes the initial weights and the batches of a run. In the random and co-expression variants it also fixes the structure of the mask. Code and tables are described under Data and code availability.

3 Results

The main experiment answers Q1 and Q1a; the four benchmarks answer Q1b and Q1c.

3.1 Q1: sparsity costs fit, and the gene sets offset part of that cost

The decoded response is the correlation between the decoded and the observed shift in expression. Its noise bound—the highest correlation it can reach, given the noise in the observed shift (Methods)—is 0.71 for the Spearman and 0.96 for the Pearson correlation in the Kang et al. data. Every masked variant carries 31–52 more units of loss per cell than the vanilla VAE. It also scores lower on the cell-type classifier and the decoded response, in every one of the 10 seeds (Table 6).

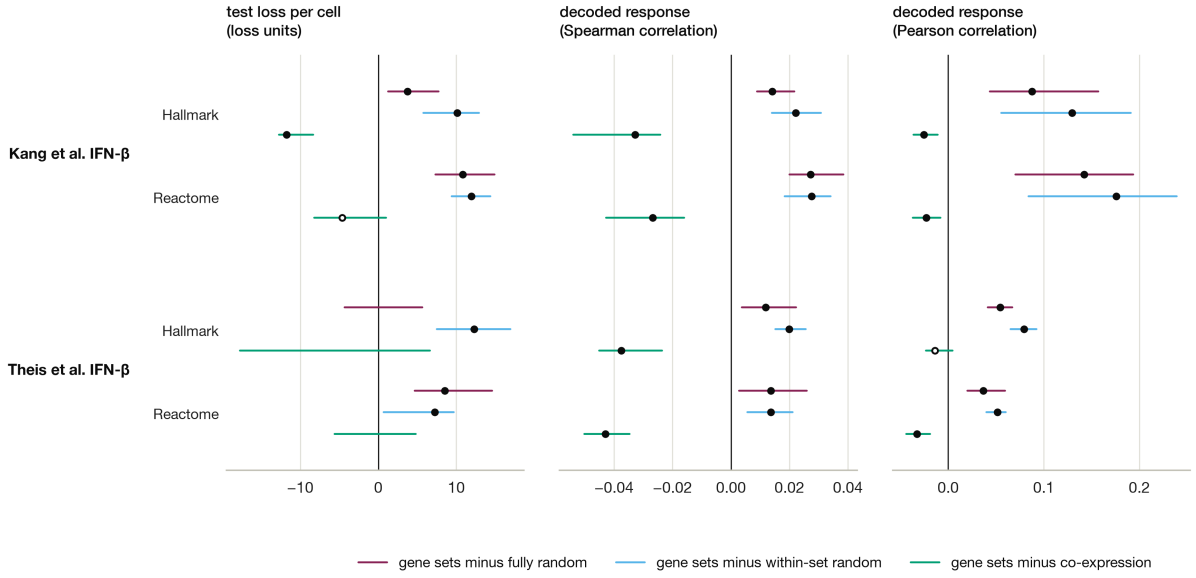


Figure 7. Gene sets against the other masks in test loss and decoded response. Each line is the CI of the paired difference over 10 seeds (gene sets minus the other mask, at the Bonferroni level). A line right of zero means the gene sets are better; for test loss the sign is reversed and a lower loss lies to the right. Each line is drawn in the colour of the mask it is compared with. A black dot at the Hodges-Lehmann estimate marks an adjusted p below 0.05. The dot is open when the CI touches zero. The two right panels show the decoded response with the Spearman and the Pearson correlation. Classification of cell type and stimulation is in Supplementary Figure 7.

The Hallmark mask reaches a lower test loss than the fully random mask by 3.7 units (Table 7; first row of Figure 7). The difference to the within-set random mask is 10.1 units. These are about 8% and 19%, respectively, of the loss that the fully random and the within-set random mask add over the vanilla VAE. It also reproduces the shift in expression on stimulation better. The improvement is slight over the fully random mask (Spearman +0.014, Pearson +0.087). It is small over the within-set random mask (Spearman +0.022, Pearson +0.129; Table 7). Across all genes (Spearman correlation) the Hallmark mask recovers 10% (95% CI [8, 11]) of the gap between the fully random mask and the vanilla VAE. In the large responses (Pearson correlation) it recovers 56% (95% CI [49, 62]). The shares for the within-set random mask are in Figure 8 (every seed in supplementary Figure 3; the shares are ratios of Hodges-Lehmann estimates with percentile CIs over the seeds). The decoded response of each variant as a share of the bound is in Table 6. On the 95 interferon-α genes, the Hallmark decoder leads the random masks by 0.25–0.29 in Spearman correlation (Table 6). The set that defines the target latent supplies these genes, so this score favours the Hallmark mask by

construction.

Table 6. Main benchmark: every variant on every measure. Kang et al. data with Hallmark sets; mean ± standard deviation (SD) over 10 seeds. The expression-matched random mask is described below. The last two rows give the decoded response as a share of the noise bound (0.715 for the Spearman and 0.961 for the Pearson correlation).

Measure	Metric	Vanilla	Hallmark	Fully random	Within-set random	Co-expression
Test loss per cell		1929.1 ± 1.0	1971.4 ± 1.2	1975.3 ± 1.5	1981.3 ± 1.4	1959.9 ± 1.5
Decoded response, all genes	Spearman	0.555 ± 0.001	0.426 ± 0.002	0.412 ± 0.003	0.404 ± 0.005	0.461 ± 0.008
	Pearson	0.942 ± 0.001	0.873 ± 0.001	0.784 ± 0.034	0.746 ± 0.044	0.898 ± 0.008
Decoded response, interferon-α genes	Spearman	0.972 ± 0.001	0.940 ± 0.003	0.688 ± 0.050	0.652 ± 0.048	0.923 ± 0.037
Cell-type classifier	Balanced accuracy	0.885 ± 0.006	0.855 ± 0.004	0.855 ± 0.005	0.850 ± 0.010	0.864 ± 0.008
Stimulation classifier	AUROC	0.9989 ± 0.0003	0.9988 ± 0.0002	0.9979 ± 0.0007	0.9980 ± 0.0003	0.9988 ± 0.0002
Share of noise bound, all genes	Spearman	78%	60%	58%	57%	65%
	Pearson	98%	91%	82%	78%	93%

Table 7. Hallmark minus each other mask over 10 seeds. The first two columns answer Q1 (random masks, same sparsity), the third Q1a (co-expression mask). Each cell reports the Hodges-Lehmann estimate of the difference, the Holm-adjusted p[49] (15 tests) and the CI at the Bonferroni level.

Measure	Hallmark – fully random (Q1)	Hallmark – within-set random (Q1)	Hallmark – co-expression (Q1a)
Test loss per cell (lower is better)	−3.7 p = 0.029 [−7.7, −1.3]	−10.1 p = 0.029 [−12.9, −5.8]	+11.8 p = 0.029 [+8.4, +12.7]
Decoded response (Spearman correlation)	+0.014 p = 0.029 [+0.009, +0.022]	+0.022 p = 0.029 [+0.014, +0.031]	−0.033 p = 0.029 [−0.054, −0.024]
Decoded response (Pearson correlation)	+0.087 p = 0.029 [+0.044, +0.157]	+0.129 p = 0.029 [+0.055, +0.191]	−0.026 p = 0.029 [−0.036, −0.011]
Cell-type balanced accuracy	+0.001 p = 1.000 [−0.008, +0.008]	+0.006 p = 0.697 [−0.007, +0.019]	−0.009 p = 0.029† [−0.020, +0.002]
Stimulation classifier (AUROC)	+0.0009 p = 0.039† [−0.0004, +0.0022]	+0.0008 p = 0.029 [+0.0003, +0.0014]	+0.0001 p = 1.000 [−0.0004, +0.0005]

† Adjusted p below 0.05 with a CI at the Bonferroni level that touches or includes zero (Holm rejects at a looser threshold). The value 0.029 is the smallest possible with 10 seeds and 15 tests (the exact p of 2/1024, or 0.00195, times 15) and is reached when all 10 seeds differ in the same direction.

Against the expression-matched random mask the Hallmark mask achieves a lower test loss and a slightly higher decoded response in all four benchmarks, in every one of the 10 seeds (adjusted p 0.010; supplementary Table 3). The decoded response of the expression-matched mask lies within

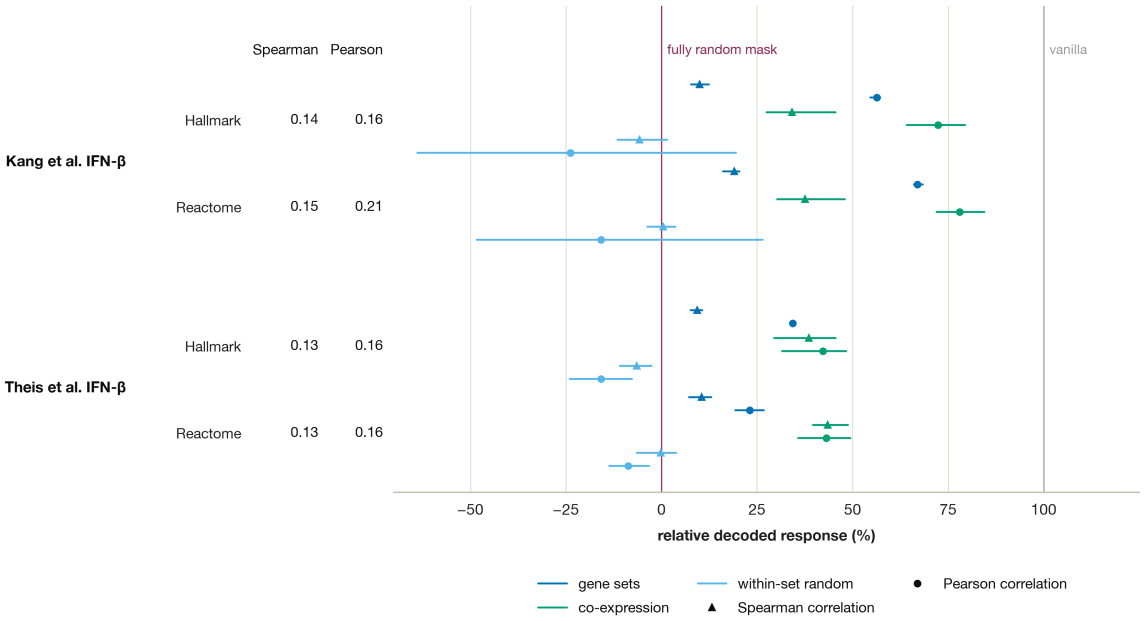


Figure 8. Decoded response relative to the fully random mask. Each variant is placed between a fully random mask (0) and the vanilla VAE (100) in each of the four benchmarks, with the Spearman correlation (triangles) and the Pearson correlation (circles); the point is the mean of 10 seeds and the line the range. The value 100 marks the vanilla VAE. In the Kang et al. data it reaches 98% of the Pearson and 78% of the Spearman noise bound. The fully random mask is the zero reference. The columns next to the benchmark names show the decoded response of the vanilla VAE minus that of the fully random mask for each correlation, the denominator of the scale. The absolute values are in supplementary S4.

0.011 (Spearman) and 0.034 (Pearson) of that of the fully random mask. Four of the 8 tests of the cell-type and stimulation classifiers are significant, three of them in favour of the Hallmark mask.

The latent labelled with the interferon-α set carries the response, whereas the random masks' latent at that position carries it unreliably (supplementary S1). Cell-type classification fails to separate the Hallmark mask from the random masks (adjusted p 1.0 and 0.70 for the fully random and the within-set random mask, respectively); classifiers trained on few cells and the two difficult cell type pairs show the same picture (supplementary S2).

All variants score 0.998–0.999 on the stimulation classifier, close to its maximum. Without any VAE, the best principal component scores 0.997 and the mean of the interferon-α genes 0.996 (supplementary S1).

3.2 Q1a: co-expressed groups exceed the Hallmark sets on fit and decoded response

The co-expression mask fits better than the Hallmark mask by 11.8 units. It reproduces the shift in expression better by a small margin (Spearman +0.033, Pearson +0.026). On the 95 interferon-α genes, the two masks are close (Spearman 0.940 for Hallmark, 0.923 for co-expression). The interferon-α latent of the Hallmark mask is equivalent to the highest-scoring latent of the co-expression mask within ±0.005 AUROC. It scores 0.0006 lower (adjusted p 0.037; supplementary S1). Only the Hallmark sets assign each latent a biological name chosen before training, such as "interferon-α response"; the meaning of a group must be determined after training.

3.3 Q1b: the response result is reproduced in both collections, the test-loss result is not

The comparison of Q1 was repeated with Reactome sets on the Kang et al. data and with both collections on interferon-β in the Theis et al. data. Each benchmark is run with 10 seeds and the six variants, and its 15 tests are corrected together. Across the four benchmarks the gene set mask reproduces the shift in expression on stimulation better than the fully random and within-set random masks (adjusted p 0.029) and worse than the co-expression mask (adjusted p 0.029–0.031), with both correlations (Figure 8; paired differences with their CIs in Figure 7). It also fits better than the within-set random mask in every benchmark. The correlation weighted by the shift and the correlation on the strongest 10% of genes show the same direction (supplementary Table 10). For test loss, the Reactome mask outperforms the fully random mask in both datasets, by 8.5–10.8 units (supplementary S4), and the Hallmark mask only in the Kang et al. data. The co-expression mask fits better than the Hallmark mask in the Kang et al. data (11.8 units). It also fits better than the Reactome mask there (4.6 units; adjusted p 0.029[†]). In the Theis et al. data the co-expression mask is not distinguished from either collection (supplementary Table 5).

3.4 Q1c: the response improvement is positive in the second dataset

On the decoded response, the gene set mask exceeds the fully random and within-set random masks in both datasets for interferon-β, with both correlations. The noise bound of the Theis et al. data is 0.65 for the Spearman and 0.85 for the Pearson correlation. The share of the bound that each variant reaches is in supplementary Table 11. In the Kang et al. data the improvement over the fully random mask is slight with Hallmark (Spearman +0.014) and small with Reactome (+0.027). In the Theis et al. data it is slight with both collections (+0.012 and +0.014, respectively; Pearson in supplementary Table 7). On test loss, the Hallmark mask is not distinguished from the fully random mask in the Theis et al. data (+0.8; CI [–5.7, +4.3]). In the Kang et al. data it is lower by 3.7 units. The Reactome mask retains the improvement in both datasets (10.8 and 8.5 units in the Kang and the Theis et al. data, respectively). Both collections outperform the within-set random mask in both datasets. Stimulation classification is not close to its maximum in the Theis et al. data (0.92–0.95). Reactome on interferon-β is the only benchmark there in which the stimulation classifier scores higher for the gene set mask than for the fully random mask (+0.0073; adjusted p 0.029†). It is also the only collection for which the cell-type classifier differs between the gene set mask and the random masks, clearly against the fully random mask (+0.0226) and weakly against the within-set random mask (+0.0137; adjusted p 0.029†).

4 Discussion

A latent assigned to a gene set is read as the activity of that pathway. This presumes that the curated knowledge accounts for the improvement over a model without a mask. A simple comparison of the gene set mask with the vanilla model cannot say whether the improvement stems from the sparsity of the mask, the knowledge in the sets, the structure inherent to any set of co-expressed genes or the response strength of the genes in the sets. The controls of this study separate these four sources.

In the main experiment, fixing 89% of the decoder weights at zero raised the test loss by 31–52 units per cell in every masked variant. It also lowered the cell-type balanced accuracy and the decoded response (Table 6). Against this cost, the gene sets outperform random masks of the same sparsity

[†]Adjusted p below 0.05 with a CI at the Bonferroni level that touches or includes zero.

on the decoded response in all four benchmarks, with both correlations.

VEGA and expiMap recover interferon programs in the Kang et al. data without a control for sparsity. The interferon-α/β latent of VEGA separates stimulated from untreated cells, and expiMap ranks interferon pathways among its top differential programs[8,9]. The latent labelled with the interferon-α set carries the response here as well (supplementary S1). VEGA reports a slightly worse fit for the latents guided by gene sets than for the full architecture[8], in line with the cost of sparsity found here. expiMap outperforms an unmasked linear decoder only with few training cells, and its authors read this as more efficient learning from prior knowledge[9]. The unmasked decoder is dense. That comparison therefore cannot distinguish knowledge from sparsity. The random masks here reflect that distinction; the gene set mask keeps a slight advantage over them in the Spearman correlation of the decoded response. pmVAE separates stimulated cells from controls with a classifier on its interferon module[10]. Every variant here separates the conditions (AUROC 0.998–0.999). However, such a classifier cannot rank the masks.

MOFA-FLEX is the one study that compares knowledge with sparsity in an unsupervised model. On murine CITE-seq data a model with uninformed sparse factors and without gene sets, gave worse batch correction and biological conservation. Its authors conclude that "the domain knowledge itself and not merely the assumption of sparsity leads to this improvement"[14]. The two results agree in direction. Knowledge adds to sparsity. The addition here is slight in Spearman correlation and larger in Pearson correlation. Nevertheless, the measures differ: MOFA-FLEX reports batch correction and biological conservation with a soft prior, while this study reports held-out loss and decoded response with a hard mask. PLIER reports that reconstruction error is smallest without the prior[12], in line with the cost of sparsity here. Spectra and Tripso report factors and programs that depart from the curated sets[13,16], as the co-expressed groups here differ from the Hallmark sets and fit better.

KPNN, a supervised network, matches unconstrained networks with many more edges on AUC, while shuffled edges generally reduce its performance[15]. The decoded response behaves similarly here: the random masks score below the Hallmark mask. Caranzano et al. explain the success of biologically informed networks largely by their structural sparsity[17]. They caution that this does not make pathways irrelevant. The gene sets here keep an advantage in the decoded response over random masks of the same sparsity, while the advantage in test loss depends on the kind of random control. The tasks differ, as those networks are supervised. Thapa et al. state that the benefit of prior knowledge depends on its relevance to the data[7]. The interferon-α set is relevant to the stimulus here, and the advantage is largest for the strongest responses.

Zhu et al. report in bulk tumour RNA-seq that curated signatures concentrate more of their variance in the leading principal components than size-matched random sets[22]. In these single-cell data the Hallmark members barely co-vary (median within-set correlation 0.003, against 0.096 for groups built from the data), and the groups still exceed the Hallmark sets.

The fully random and within-set random masks answer different questions. The difference between the Hallmark mask and the fully random mask partly measures where the interferon-α genes lie. The fully random mask places most of them among the genes that only the free latents reach. Conversely, the within-set random mask leaves every interferon-α gene reachable by a labelled latent and isolates the value of knowing which genes belong together. The Hallmark improvement on test loss is much smaller against the fully random mask than against the within-set random mask (Table 7). The choice of random control thus affects the result. The expression-matched random mask retains in every set the number of genes of each response strength and still underperforms the Hallmark mask on test loss and decoded response in all four benchmarks (supplementary S3).

In these data, co-expressed groups reproduced the decoded response better than the Hallmark sets

in all four benchmarks and fitted better in the Kang et al. data (not distinguished in the Theis et al. data). The co-expression mask carries only expression coherence and exceeds the Hallmark sets, while the within-set random mask removes both curated knowledge and expression coherence. Hence, the comparison does not separate the two.

The decoded-response improvement over the random masks was positive in all four benchmarks, with a second collection and a second dataset. The Reactome mask outperformed the fully random mask on test loss in both datasets and the Hallmark mask only in the main experiment.

The differences between the masks are small against the effect of masking itself. The fully random mask decodes the response almost on par with the Hallmark mask, at 97% of its Spearman and 90% of its Pearson score (Table 6). The co-expression mask exceeds the Hallmark mask. Most of what a masked decoder achieves in these data is set by its sparsity. The curated sets add little over all genes. They add more on the 95 interferon-α genes (Results), a score that favours the Hallmark mask by construction. The sets place the interferon genes under one labelled latent, which the random masks do not. Curated knowledge names each latent before training. In these data, the labelled latent carries it. Groups of co-expressed genes found in the data fit better and reproduce the response better without any curated set, but their meaning has to be determined after training. A choice of masked decoder for interpretability can use curated sets at a cost in fit. In the main experiment the vanilla VAE has a lower test loss and a higher decoded response than every masked variant, so a choice made for fit needs no mask.

For work that applies gene set masks, a random mask that retains the sparsity, size and overlaps of the sets is the minimum control for a claim about knowledge. The genes that the random mask leaves reachable by labelled latents must be stated. A labelled latent should be read as pathway activity only after its name is checked against the expected response.

4.1 Limitations

Data The main analysis rests on one stimulus, interferon-β, and one expected response, the Hallmark interferon-α set. The Kang et al. data come from 8 patients with SLE. Their type I interferon activity is often raised at baseline[58]. The Theis et al. data add interferon-β in 12 healthy donors. The two datasets differ in donors, stimulus time, platform and scored cell types. The interferon-α and -γ sets share 71 of 95 genes. The expected response is therefore not specific to type I interferon.

Signal The stimulation is so strong that a principal component, the mean of the interferon-α genes and every variant separate the conditions (supplementary S1). Stimulation classification therefore cannot rank the variants. In the Kang et al. data each condition was pooled and sequenced in one 10x run. Every stimulation score there potentially includes run effects. Removing the 500 strongest stimulation genes lowers the best principal component from 0.997 to 0.744 (2,000 genes: 0.573; supplementary S1). The separation depends mainly on a few hundred genes. The Theis et al. data come from a different platform, Parse split-pool barcoding with Ultima sequencing[27]. Each donor occupies one 96-well plate with the 90 cytokines and 6 controls[28]. Neither the preprint nor the dataset page states how sublibraries and sequencing runs relate to the conditions. Run effects cannot be excluded in either dataset. Replicate runs would settle it.

Comparison The better fit and decoded response of the Hallmark mask may arise mainly from sets that respond to the stimulation, such as interferon and TNF-α. A dataset without such sets may differ. An exploratory repeat after removing the 250 strongest stimulation genes did not distinguish the gene set mask from the fully random mask. It does not test whether the improvement depends on these genes. The Hallmark sets and the groups of co-expressed genes both depend on expression

data. The groups are built from the training cells and thus favour fit. The comparison is between sets named in advance and sets found in this data. Reactome, built by manual pathway curation, adds a collection that is independent of expression data. Notably, the Reactome comparison rests on one draw of 50 sets from about 950 eligible. The 10 seeds do not measure this draw. Thus, a difference between the two collections refers to this semi-random draw. The expected-response sets of Reactome are narrower than those of Hallmark. A smaller Reactome improvement can reflect these narrower sets.

Statistics The 10 seeds vary the initial weights, the batch order, the dropout and the latent sampling. For the random and co-expression masks a seed also selects the mask. The gene set mask is one fixed mask, compared with the spread of 10 random masks and the training noise. The split, the donors and the cells are fixed. The test cells share their donors with the training cells. No donor was held out. The tests and CIs therefore describe training and mask randomness on this split, and say nothing about new donors. The SD of the random masks (0.034 and 0.044 for the Pearson correlation of the decoded response, 0.003 and 0.005 for the Spearman correlation, for the fully random and the within-set random mask, respectively) confounds the draw of the mask with the training noise. The classifiers, the score of the labelled latent and the per-cell-type and per-donor scores are descriptive.

Model The design is restricted to a hard mask and a linear decoder. Soft masks, nonlinear decoders, gene set scores as input features and graph models were not tested. Letting the 8 free latents reach only unassigned genes followed trial runs. With access to all genes they absorbed the response and the labelled latents became inactive. The training settings of Table 3 were not tuned for each variant. The encoder width was chosen on validation loss on the vanilla VAE. On the fully random mask, widths 128 and 256 differed by 0.4 units, against 3.1 on the vanilla VAE. A learning rate of 1e-4 improved the vanilla VAE by 1.1 units at 4.6 times the epochs. The masked models did not reach early stopping within 400 epochs at that rate. Other settings could still narrow or widen the gap between the vanilla VAE and the masked variants. The vanilla VAE is an unmasked reference with about nine times more decoder weights (697,972 against 76,344). The random masks are the matched controls. The shares in Figure 8 are positions between the fully random mask and the vanilla VAE. They do not measure the fraction of an attainable score.

Measurement The decoded response is noisiest for the 90% of genes with the smaller shifts. On these genes the gene set mask exceeds the fully random mask in only some of the benchmarks (supplementary Table 10). The Spearman bound is an empirical estimate (Methods). The Theis et al. data are subsampled to at most 1,200 cells per donor × condition group after QC. Five cell types with too few test cells are not scored.

Processing The 2017 gene annotation (hg19) and the labels of the data authors (cell types and demuxlet calls) are used as provided. Doublets of two cells from the same donor are not detected (an estimated 2% of cells). Ambient RNA is not corrected.

Outside the scope Two further analyses are outside the scope of this paper: learning curves on subsamples of the training cells and a second stimulus, IL-2. In the learning curves the interferon-α score is saturated and the vanilla VAE led at every size. The response to IL-2 is dominated by interferon genes. It is therefore not an independent test of another stimulus. The scripts and tables of both analyses accompany the code.

4.2 Future work

Perturbations of the transcription factors IRF4 and PRDM1 in B cell differentiation offer ground truth for testing whether the labelled latents recover known regulators[59]. Cellular indexing of tran-

scriptomes and epitopes by sequencing (CITE-seq) would show whether RNA-based pathway activity agrees with surface protein measured in the same cells. Incomplete or species-transferred gene sets, for example human sets applied to mouse gastrulation data[60], would show how the result depends on the quality of the sets. A soft mask initialized from random sets would show whether training refines it to the same gene sets.

Holding out whole donors of the Theis et al. data would test whether the result transfers to donors unseen in training. Further draws of Reactome sets would measure the variation between draws (Limitations), and a new split for each seed would separate the split from the training noise.

Abbreviations

AUROC Area under the receiver operating characteristic curve

CITE-seq Cellular indexing of transcriptomes and epitopes by sequencing

CPU Central processing unit

ELBO Evidence lower bound

GPU Graphics processing unit

HGNC HUGO Gene Nomenclature Committee

IFN Interferon

IL-2 Interleukin-2

IL-6 Interleukin-6

KL Kullback-Leibler

LDVAE Linear-decoder variational autoencoder

MSigDB Molecular Signatures Database

NB Negative binomial

PBMC Peripheral blood mononuclear cell

PBS Phosphate-buffered saline

QC Quality control

ReLU Rectified linear unit

SD Standard deviation

SLE Systemic lupus erythematosus

TNF Tumour necrosis factor

VAE Variational autoencoder

Glossary

Ambient RNA RNA that floats free in the cell suspension and enters droplets next to the cells.

Balanced accuracy The share of cells classified correctly, averaged over the classes so that rare classes count as much as common ones.

Batch The cells that the model processes together in one update of its weights.

Bonferroni level The coverage $1 - \alpha/m$ of each of m CIs, chosen so that all m CIs cover their targets together at $1 - \alpha$.

Cell state The condition of a cell at a given time, such as its place in a developmental process, its phase in the cell cycle, its response to a stimulus or its stress level. Cells of any type can be in different states.

CITE-seq A method that measures RNA and surface proteins in the same cells.

Co-expressed genes Genes whose expression rises and falls together across cells.

Cytokine A signalling protein that cells release to influence other cells. Interferons and interleukins are cytokines.

Doublet A droplet that contains two cells. Demuxlet separates doublets from singlets by the genotypes of the donors.

Dropout A training technique that sets a random share of the hidden units to zero in every update; the share is 0.1 here.

Early stopping Training stops when the validation loss has not improved for a set number of epochs, the patience (3 epochs here).

Embedding A list of numbers that represents a cell, a gene or a gene program.

Epoch One pass of the model through all training cells.

Gene program A coordinated set of biologically related genes. Some papers apply the term to what this paper calls a gene set.

Highly variable genes Genes whose expression varies most across cells; selecting a few thousand of them is a common step before modelling.

Holm correction A procedure that adjusts p-values for several tests run together and keeps the chance of any false positive at or below the chosen level.

Jaccard index The size of the overlap of two sets divided by the size of their union: 0 for sets without shared genes and 1 for identical sets.

Kullback-Leibler divergence A measure of how far one probability distribution lies from another. In the VAE it measures how far the distribution of the latent variables of a cell lies from a standard normal distribution.

Learning rate The size of each update of the model weights.

Library size The total count of a cell.

Median absolute deviation The median of the absolute differences of values from their median, a measure of spread that a few extreme cells barely move.

One-hot code A vector with a 1 at the position of a category, here the donor, and 0 elsewhere.

Posterior The distribution of the latent variables of a cell given its counts, proposed by the encoder.

Prior The standard normal distribution towards which the KL term draws the posterior.

Self-supervised models Models trained without labels on a task derived from the data itself.

Shifted-log transform The logarithm of normalized counts after a constant has been added.

Singlet A droplet that contains one cell.

Supervised models Models trained with cell labels.

Transformer A neural network that relates every part of its input to every other part.

Unsupervised models Models trained without cell labels.

Wilcoxon signed-rank test A paired test of whether the median of the differences is zero, based on the ranks of their absolute values. The Walsh averages are the means of all pairs of differences, and their median is the Hodges-Lehmann estimate of the effect.

Use of generative AI

The author used Claude Opus 5.5 and Claude Sonnet 5.5 (Anthropic) between September and October 2026 to help write and debug analysis code and to improve the readability of the text. The author conceived the study, designed the experiments, ran all analyses and wrote the text. The author reviewed all AI-assisted code and text and assumes full responsibility for the accuracy, integrity and originality of this work.

Data and code availability

All data are public. The data of Kang et al.[24] (8 donors, interferon-β) are in the Gene Expression Omnibus under GSE96583 (batch 2). The data of Theis et al.[27] are the Parse Biosciences dataset "10 Million Human PBMCs in a Single Experiment"[28], released under CC BY-NC 4.0. The analysis code reads that archive in place without redistributing it. The gene sets are the Hallmark collection[20] and the Reactome collection (C2 CP:REACTOME) of the Molecular Signatures Database v2024.1. The Reactome pathway hierarchy[23] is the release of 20 October 2026. Gene symbols follow the HGNC[29].

The tables behind every figure, test and number in this paper (scores per run, paired differences and tests for each benchmark) are provided with the code. The code regenerates the trained models.

The code is at `https://github.com/gmiliarakis/knowledge-or-sparsity-sc-vae` under the MIT licence. The version archived on Zenodo is 10.5281/zenodo.23729999.

Competing interests

The author declares no competing interests.

Funding

This research received no specific funding.

References

[1] Heumos, L. et al. Best practices for single-cell analysis across modalities. *Nat Rev Genet* 24, 550–572 (2023). doi:10.1038/s41576-023-00586-w

[2] Luecken, M. D. & Theis, F. J. Current best practices in single-cell RNA-seq analysis: a tutorial. *Mol Syst Biol* 15, e8746 (2019). doi:10.15252/msb.20188746

[3] Wolf, F. A., Angerer, P. & Theis, F. J. SCANPY: large-scale single-cell gene expression data analysis. *Genome Biol* 19, 15 (2018). doi:10.1186/s13059-017-1382-0

[4] Kingma, D. P. & Welling, M. Auto-encoding variational Bayes. *arXiv* (2013), presented at ICLR 2014. doi:10.48550/arXiv.1312.6114

[5] Svensson, V., Gayoso, A., Yosef, N. & Pachter, L. Interpretable factor models of single-cell RNA-seq via variational autoencoders. *Bioinformatics* 36, 3418–3421 (2020). doi:10.1093/bioinformatics/btaa169

[6] Subramanian, A. et al. Gene set enrichment analysis: a knowledge-based approach for interpreting genome-wide expression profiles. *Proc Natl Acad Sci USA* 102, 15545–15550 (2005). doi:10.1073/pnas.0506580102

[7] Thapa, K., Kinali, M., Pei, S., Luna, A. & Babur, Ö. Strategies to include prior knowledge in omics analysis with deep neural networks. *Patterns* 6, 101203 (2025). doi:10.1016/j.patter.2025.101203

[8] Seninge, L., Anastopoulos, I., Ding, H. & Stuart, J. VEGA is an interpretable generative model for inferring biological network activity in single-cell transcriptomics. *Nat Commun* 12, 5684 (2021). doi:10.1038/s41467-021-26017-0

[9] Lotfollahi, M. et al. Biologically informed deep learning to query gene programs in single-cell atlases. *Nat Cell Biol* 25, 337–350 (2023). doi:10.1038/s41556-022-01072-x

[10] Gut, G., Stark, S. G., Rätsch, G. & Davidson, N. R. pmVAE: learning interpretable single-cell representations with pathway modules. *bioRxiv* (2021). doi:10.1101/2021.01.28.428664

[11] Buettner, F., Pratanwanich, N., McCarthy, D. J., Marioni, J. C. & Stegle, O. f-scLVM: scalable and versatile factor analysis for single-cell RNA-seq. *Genome Biol* 18, 212 (2017). doi:10.1186/s13059-017-1334-8

[12] Mao, W., Zaslavsky, E., Hartmann, B. M., Sealfon, S. C. & Chikina, M. Pathway-level information extractor (PLIER) for gene expression data. *Nat Methods* 16, 607–610 (2019). doi:10.1038/s41592-019-0456-1

[13] Kunes, R. Z., Walle, T., Land, M., Nawy, T. & Pe'er, D. Supervised discovery of interpretable gene programs from single-cell data. *Nat Biotechnol* 42, 1084–1095 (2024). doi:10.1038/s41587-023-01940-3

[14] Qoku, A. et al. MOFA-FLEX: a factor model framework for integrating omics data with prior knowledge. *bioRxiv* (2025). doi:10.1101/2025.11.03.686250

[15] Fortelny, N. & Bock, C. Knowledge-primed neural networks enable biologically interpretable deep learning on single-cell sequencing data. *Genome Biol* 21, 190 (2020). doi:10.1186/s13059-020-02100-5

[16] Moullet, M. et al. Self-supervised learning for a gene program-centric view of cell states. *bioRxiv* (2026). doi:10.64898/2026.03.24.713961

[17] Caranzano, I. et al. Sparsity is all you need: rethinking biologically informed neural networks. *Brief Bioinform* 27, bbag425 (2026). doi:10.1093/bib/bbag425

[18] Schneider, W. M., Chevillotte, M. D. & Rice, C. M. Interferon-stimulated genes: a complex web of host defenses. *Annu Rev Immunol* 32, 513–545 (2014). doi:10.1146/annurev-immunol-032713-120231

[19] Ivashkiv, L. B. & Donlin, L. T. Regulation of type I interferon responses. *Nat Rev Immunol* 14, 36–49 (2014). doi:10.1038/nri3581

[20] Liberzon, A. et al. The Molecular Signatures Database (MSigDB) hallmark gene set collection. *Cell Syst* 1, 417–425 (2015). doi:10.1016/j.cels.2015.12.004

[21] Jolliffe, I. T. & Cadima, J. Principal component analysis: a review and recent developments. *Philos Trans A Math Phys Eng Sci* 374, 20150202 (2016). doi:10.1098/rsta.2015.0202

[22] Zhu, Y., Zhang, C., Calhoun, V. D. & Bi, Y. A residual-ratio framework for auditing transcriptomic gene signatures against background expression structure. *bioRxiv* (2026). doi:10.64898/2026.04.11.717907

[23] Milacic, M. et al. The Reactome Pathway Knowledgebase 2024. *Nucleic Acids Res* 52, D672–D678 (2024).

doi:10.1093/nar/gkad1025

[24] Kang, H. M. et al. Multiplexed droplet single-cell RNA-sequencing using natural genetic variation. *Nat Biotechnol* 36, 89–94 (2018). doi:10.1038/nbt.4042

[25] Makrodimitris, S., Pronk, B., Abdelaal, T. & Reinders, M. An in-depth comparison of linear and non-linear joint embedding methods for bulk and single-cell multi-omics. *Brief Bioinform* 25, bbad416 (2023). doi:10.1093/bib/bbad416

[26] Ahlmann-Eltze, C., Huber, W. & Anders, S. Deep-learning-based gene perturbation effect prediction does not yet outperform simple linear baselines. *Nat Methods* 22, 1657–1661 (2025). doi:10.1038/s41592-025-02772-6

[27] Theis, F. et al. A single-cell cytokine dictionary of human peripheral blood. *Research Square* (2025), preprint without peer review. doi:10.21203/rs.3.rs-8337240/v1

[28] Parse Biosciences. 10 Million Human PBMCs in a Single Experiment (data set, CC BY-NC 4.0). https://www.parsebiosciences.com/datasets/10-million-human-pbmcs-in-a-single-experiment/ (accessed 21 October 2026).

[29] Seal, R. L. et al. Genenames.org: the HGNC resources in 2023. *Nucleic Acids Res* 51, D1003–D1009 (2023). doi:10.1093/nar/gkac888

[30] Ahlmann-Eltze, C. & Huber, W. Comparison of transformations for single-cell RNA-seq data. *Nat Methods* 20, 665–672 (2023). doi:10.1038/s41592-023-01814-1

[31] Lopez, R., Regier, J., Cole, M. B., Jordan, M. I. & Yosef, N. Deep generative modeling for single-cell transcriptomics. *Nat Methods* 15, 1053–1058 (2018). doi:10.1038/s41592-018-0229-2

[32] Svensson, V. Droplet scRNA-seq is not zero-inflated. *Nat Biotechnol* 38, 147–150 (2020). doi:10.1038/s41587-019-0379-5

[33] Kullback, S. & Leibler, R. A. On information and sufficiency. *Ann Math Stat* 22, 79–86 (1951). doi:10.1214/aoms/1177729694

[34] Sønderby, C. K., Raiko, T., Maaløe, L., Sønderby, S. K. & Winther, O. Ladder variational autoencoders. *arXiv* (2016), presented at NeurIPS 2016. doi:10.48550/arXiv.1602.02282

[35] Kingma, D. P. & Ba, J. Adam: a method for stochastic optimization. *arXiv* (2014), presented at ICLR 2015. doi:10.48550/arXiv.1412.6980

[36] Eltager, M. et al. Benchmarking variational autoencoders on cancer transcriptomics data. *PLoS One* 18, e0292126 (2023). doi:10.1371/journal.pone.0292126

[37] Spearman, C. The proof and measurement of association between two things. *Am J Psychol* 15, 72–101 (1904). doi:10.2307/1412159

[38] Hanley, J. A. & McNeil, B. J. The meaning and use of the area under a receiver operating characteristic (ROC) curve. *Radiology* 143, 29–36 (1982). doi:10.1148/radiology.143.1.7063747

[39] Hsu, A., Borst, A. & Theunissen, F. E. Quantifying variability in neural responses and its application for the validation of model predictions. *Network* 15, 91–109 (2004). PMID 15214701

[40] Lage-Castellanos, A., Valente, G., Formisano, E. & De Martino, F. Methods for computing the maximum performance of computational models of fMRI responses. *PLoS Comput Biol* 15, e1006397 (2019). doi:10.1371/journal.pcbi.1006397

[41] Wenkel, F. et al. TxPert: using multiple knowledge graphs for prediction of transcriptomic perturbation effects. *Nat Biotechnol* (2026). doi:10.1038/s41587-026-03113-4

[42] Vollenweider, M. & Bühlmann, P. Signal, bounds, and baselines: principles for rigorous evaluation of high-dimensional biological perturbation prediction. *bioRxiv* (2026). doi:10.64898/2026.04.20.719650

[43] Wang, X., Kuipers, J., Hugi, F., Platt, R. J. & Beerenwinkel, N. Reliable single-cell perturbations explain and improve model performance. *bioRxiv* (2026). doi:10.64898/2026.08.11.744177

[44] Lord, F. M. & Novick, M. R. *Statistical Theories of Mental Test Scores* (Addison-Wesley, Reading, MA, 1968).

[45] Spearman, C. Correlation calculated from faulty data. *Br J Psychol* 3, 271–295 (1910). doi:10.1111/j.2044-8295.1910.tb00206.x

[46] Brown, W. Some experimental results in the correlation of mental abilities. *Br J Psychol* 3, 296–322

(1910). doi:10.1111/j.2044-8295.1910.tb00207.x

[47] Wilcoxon, F. Individual comparisons by ranking methods. *Biometrics Bull* 1, 80–83 (1945). doi:10.2307/3001968

[48] Hodges, J. L. & Lehmann, E. L. Estimates of location based on rank tests. *Ann Math Stat* 34, 598–611 (1963). doi:10.1214/aoms/1177704172

[49] Holm, S. A simple sequentially rejective multiple test procedure. *Scand J Stat* 6, 65–70 (1979). jstor.org/stable/4615733

[50] Dunn, O. J. Multiple comparisons among means. *J Am Stat Assoc* 56, 52–64 (1961). doi:10.1080/01621459.1961.10482090

[51] Paszke, A. et al. PyTorch: an imperative style, high-performance deep learning library. *arXiv* (2019), presented at NeurIPS 2019. doi:10.48550/arXiv.1912.01703

[52] Virshup, I. et al. anndata: access and store annotated data matrices. *J Open Source Softw* 9, 4371 (2024). doi:10.21105/joss.04371

[53] Harris, C. R. et al. Array programming with NumPy. *Nature* 585, 357–362 (2020). doi:10.1038/s41586-020-2649-2

[54] McKinney, W. Data structures for statistical computing in Python. *Proc 9th Python in Science Conf* 56–61 (2010). doi:10.25080/Majora-92bf1922-00a

[55] Virtanen, P. et al. SciPy 1.0: fundamental algorithms for scientific computing in Python. *Nat Methods* 17, 261–272 (2020). doi:10.1038/s41592-019-0686-2

[56] Pedregosa, F. et al. Scikit-learn: machine learning in Python. *J Mach Learn Res* 12, 2825–2830 (2011). jmlr.org/papers/v12/pedregosa11a.html

[57] Hunter, J. D. Matplotlib: a 2D graphics environment. *Comput Sci Eng* 9, 90–95 (2007). doi:10.1109/MCSE.2007.55

[58] Sim, T. M., Ong, S. J., Mak, A. & Tay, S. H. Type I interferons in systemic lupus erythematosus: a journey from bench to bedside. *Int J Mol Sci* 23, 2505 (2022). doi:10.3390/ijms23052505

[59] Demela, P. et al. Competing gene regulatory networks drive naive and memory B cell differentiation. *Mol Syst Biol* 22, 1097–1117 (2026). doi:10.1038/s44320-026-00207-8

[60] Pijuan-Sala, B. et al. A single-cell molecular map of mouse gastrulation and early organogenesis. *Nature* 566, 490–495 (2019). doi:10.1038/s41586-019-0933-9

Supplementary Files

This is a list of supplementary files associated with this preprint. Click to download.

- supplementary.pdf