Supplementary Information

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

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

Abbreviations AUROC, area under the receiver operating characteristic curve; HGNC, HUGO Gene Nomenclature Committee; IFN-β, interferon-β; KL, Kullback-Leibler; PBMC, peripheral blood mononuclear cell; SD, standard deviation; VAE, variational autoencoder. Terms not defined here are defined in the Glossary of the main text.

S1. The interferon-α latent (Q1, Q1a)

Definition

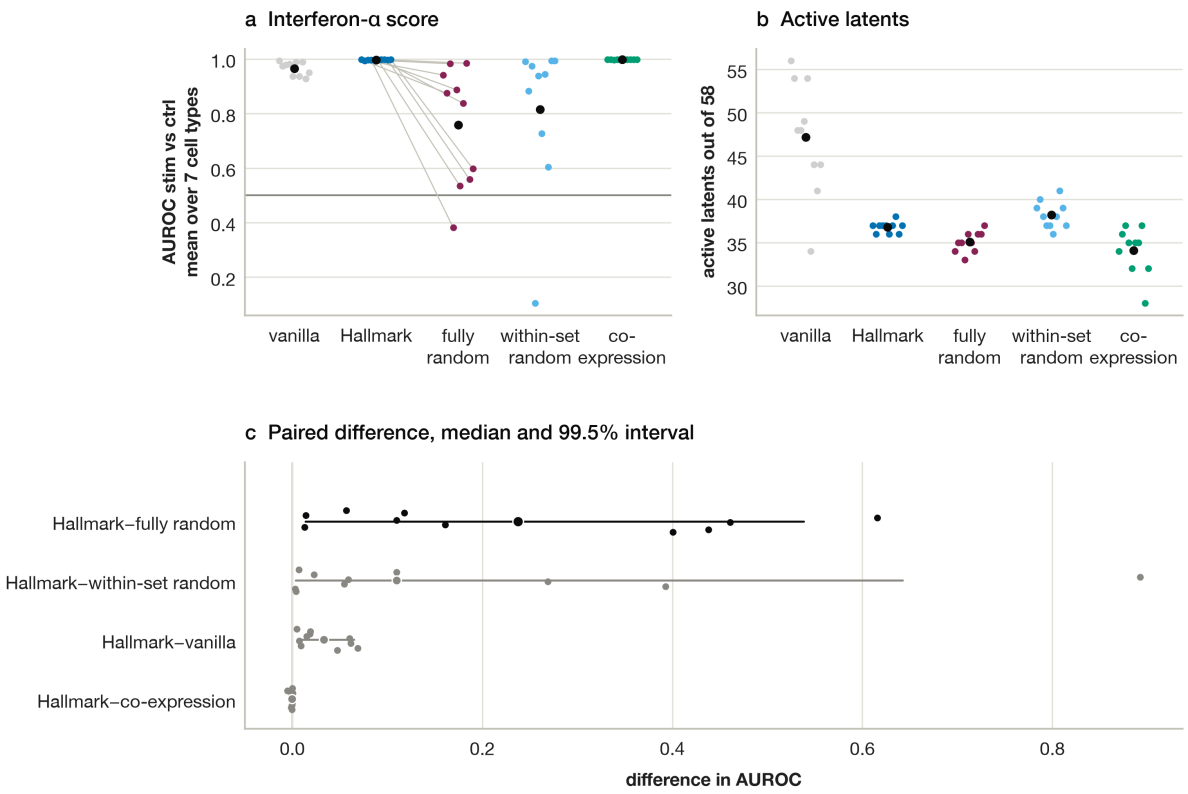
A latent labelled with the Hallmark interferon-α response set should be higher in stimulated cells. IFN-β is a type I interferon. These genes are induced in blood cells[1–3]. The score is built from the AUROC[4] of the latent's posterior mean, the probability that the latent value of a stimulated test cell exceeds that of a control test cell of the same cell type. The AUROC is computed within each cell type; a latent therefore cannot score by separating cell types. To count each type equally, the score is the mean of these values over the number of types, 7 (13 in the Theis et al. data). A score of 0.5 is chance. For the Hallmark and both random masks the latent at the interferon-α position is scored, oriented without labels by the sign of its mean decoder weight. It scores 0.5 when inactive (variance of its posterior means over validation cells at most 0.01). The vanilla and co-expression variants have no names. Their active latent with the largest |AUROC − 0.5| on validation cells is scored on test cells (the highest-scoring latent). The choice of the scored latent uses validation labels for these two variants.

The score shows where the response lies; it is not specific to interferon biology. Any latent whose genes respond to the stimulation can separate the conditions; the Hallmark interferon-α and -γ sets share 71 of 95 genes. The score approaches its maximum of 1.

Reference scores without a VAE: the mean log expression of the interferon-α genes (no labels) and the best of 58 principal components of the training cells (chosen on validation labels), scored like the latents. The best principal component is also refitted after removing the 216 genes of the Hallmark interferon-α and -γ sets, or the 100–2,000 genes with the largest stimulation effect in training cells.

Results

The interferon-α latent of the Hallmark mask scored 0.997 ± 0.001 (Supplementary Figure 1) and was its highest-scoring latent in 9 of 10 seeds (the interferon-γ latent in the tenth). It scored higher than the latent at the same position of the random masks in all 10 seeds (adjusted $p = 0.018$ for both). Those latents scored 0.758 ± 0.218 and 0.815 ± 0.281.



Supplementary Figure 1. Interferon-α score of the Hallmark latent (Kang et al. data). a, interferon-α score of every seed (lines join the Hallmark and random scores of the same seed). b, latents with an active posterior. c, paired difference of Hallmark minus each other variant, with the Hodges-Lehmann estimate[5] and the CI at the Bonferroni level; the primary comparison is in black.

In the random masks another latent reflects the response (best-latent scores of 0.989 ± 0.005 and 0.994 ± 0.002). The best latent is the interferon-α latent in 0 and 3 of 10 seeds. The score therefore does not single out a specific latent. A random set of about 100 genes scored 0.93 or higher in 6 of 10 within-set random seeds. Against co-expression the Hallmark latent is equivalent within ±0.005 (adjusted p = 0.010), and the difference of −0.0006 is itself detectable (adjusted p = 0.037).

Every variant scores near the maximum: without any VAE, the best principal component scores 0.997 and the mean of the interferon-α genes 0.996. After removing the 216 genes of the interferon-α and -γ sets the best component still scores 0.945. After removing the 500, 1,000 or 2,000 strongest stimulation genes it scores 0.744, 0.649 or 0.573. The separation depends on a few hundred genes. Each condition was sequenced in its own 10x Genomics run, so the separation is confounded with run.

S2. Cell-type classification, few-shot and difficult cell type pairs (Q1)

Supplementary Table 1. Descriptive cell-type classifiers. Balanced accuracy with 10, 50 and 200 validation cells per type, and AUROC between CD4 and CD8 T cells and between CD14+ and FCGR3A+ monocytes; mean ± SD over 10 seeds. Types with fewer than k validation cells contribute all of them; the classifier is described under Evaluation in the main text. Mean ± SD over 10 seeds of the balanced accuracy with 10, 50 and 200 validation cells per type, and of the AUROC between CD4 and CD8 T cells and between CD14+ and FCGR3A+ monocytes.

Cell-type classifier, descriptive	10 per type	50 per type	200 per type	CD4 vs CD8 T cells	CD14+ vs FCGR3A+ monocytes
Vanilla	0.815 ± 0.028	0.865 ± 0.008	0.884 ± 0.009	0.992 ± 0.001	0.962 ± 0.003
Hallmark	0.806 ± 0.023	0.839 ± 0.010	0.858 ± 0.003	0.984 ± 0.001	0.950 ± 0.004
Fully random	0.811 ± 0.026	0.840 ± 0.011	0.860 ± 0.008	0.986 ± 0.002	0.939 ± 0.005
Within-set random	0.805 ± 0.025	0.841 ± 0.012	0.857 ± 0.010	0.983 ± 0.003	0.949 ± 0.006
Co-expression	0.827 ± 0.017	0.854 ± 0.009	0.867 ± 0.009	0.991 ± 0.002	0.949 ± 0.009

S3. Alternative measures, the noise bound and the expression-matched mask (Q1)

Alternative measures and the noise bound

The decoded response of the main text (Spearman and Pearson correlation) is recomputed from the trained models with three further measures. They are computed across the genes detected in at least 1% of training cells and averaged over cell types. The correlation weighted by the absolute observed shift is one. The other two split the genes by the size of the observed shift in each cell type: the correlation on the 10% of genes with the largest shift and on the other 90%. The noise bound (Methods) appears in the last row. The bounds of the weighted correlation and of the two gene subsets use genes and weights chosen from the observed shift of the same test cells and describe the selection; on the other 90% of genes the vanilla VAE exceeds its bound (0.48 against 0.38). "None" marks a measure whose halves are uncorrelated. Means over 10 seeds (Supplementary Table 2).

Supplementary Table 2. Alternative decoded-response measures with all 12,034 genes. Kang et al. data with Hallmark sets; means over 10 seeds, with the noise bound in the last row.

Variant	Pearson	Spearman	Pearson weighted	Strongest 10% of genes	Other 90% of genes
Vanilla	0.942	0.555	0.987	0.965	0.482
Hallmark	0.873	0.426	0.959	0.899	0.337
Co-expression	0.898	0.461	0.973	0.924	0.355
Within-set random	0.746	0.404	0.816	0.780	0.307
Fully random	0.784	0.412	0.852	0.813	0.319
Noise bound	0.96	0.71	0.99	0.98	0.38

On every measure the Hallmark mask exceeds the fully random and within-set random masks in the main experiment, and in all four benchmarks (Supplementary Table 10) on the Pearson, Spearman, weighted and strongest-10% correlations. On the 90% of genes with the smaller shifts the noise bound is 0.38 in the Kang et al. data and the shifts of the halves are uncorrelated in the Theis et al. data. The Pearson difference between the gene set mask and the within-set random mask is

positive in at least 8 of 10 seeds in every benchmark, and the difference to the fully random mask is positive in 10, 2, 5 and 9 of 10 seeds (Supplementary Table 10).

Expression-matched random mask

The Hallmark improvement over random masks could arise from the sets containing the strongest response genes. The expression-matched mask tests this. It permutes the labels among the set genes only, inside bins of expression level and stimulation effect; every set therefore retains its number of genes of each response strength. It was fit with 10 seeds on the four benchmarks. The five scores form their own group of five tests per benchmark, corrected with Holm and separate from the 15 tests of the main comparison.

Supplementary Table 3. Hallmark minus the expression-matched mask. Hodges-Lehmann estimate of the paired differences over 10 seeds; an asterisk marks a Holm-adjusted p below 0.05 within the five tests of each benchmark, and a dagger marks such a p with a CI that touches or includes zero. Test loss is lower for the better mask.

Data	Collection	Test loss per cell	Decoded response, Spearman	Decoded response, Pearson	Cell-type balanced accuracy	Stimulation classifier AUROC
Kang et al. IFN-β	Hallmark	−7.1*	+0.008*	+0.048*	+0.002	+0.0003
Kang et al. IFN-β	Reactome	−13.1*	+0.017*	+0.130*	−0.006†	+0.0009*
Theis et al. IFN-β	Hallmark	−8.8*	+0.012*	+0.047*	−0.001	−0.0001
Theis et al. IFN-β	Reactome	−7.0*	+0.010*	+0.036*	+0.012†	+0.0067†

The Hallmark mask attains a lower test loss and a higher decoded response than the expression-matched mask in all four benchmarks, in every seed (adjusted p 0.010; Supplementary Table 3). The decoded response of the expression-matched mask is within 0.012 of that of the fully random mask in every benchmark. Four of the 8 tests of the cell-type and stimulation classifiers are significant, three of them in favour of the Hallmark mask.

S4. Reactome and the Human Cytokine Dictionary (Q1b, Q1c)

Design

The four benchmarks, the data and the expected-response sets are described in Methods. Architecture, training, scoring and the five scores are unchanged. The 15 tests of each benchmark (five scores, the gene set mask minus the three other masks) form their own multiple-testing correction.

Supplementary Table 4. Five scores in every benchmark. Means over 10 seeds (± SD) of the five scores. For the Kang et al. data with Hallmark, Table 6 of the main text applies.

Data	Collection	Variant	Test loss per cell	Decoded response, Spearman	Decoded response, Pearson	Cell-type balanced accuracy	Stimulation classifier AUROC
Kang et al. IFN-β	Reactome	Vanilla	1929.1 ± 1.0	0.555 ± 0.001	0.942 ± 0.001	0.885 ± 0.006	0.9989 ± 0.0003
Kang et al. IFN-β	Reactome	Gene set	1963.0 ± 1.2	0.437 ± 0.002	0.873 ± 0.002	0.861 ± 0.009	0.9987 ± 0.0002

Data	Collection	Variant	Test loss per cell	Decoded response, Spearman	Decoded response, Pearson	Cell-type balanced accuracy	Stimulation classifier AUROC
Kang et al. IFN-β	Reactome	Fully random	1973.8 ± 1.7	0.409 ± 0.005	0.733 ± 0.037	0.856 ± 0.005	0.9979 ± 0.0004
Kang et al. IFN-β	Reactome	Within-set random	1975.0 ± 0.7	0.410 ± 0.005	0.700 ± 0.046	0.865 ± 0.007	0.9976 ± 0.0005
Kang et al. IFN-β	Reactome	Co-expression	1958.8 ± 2.5	0.464 ± 0.008	0.896 ± 0.009	0.863 ± 0.008	0.9986 ± 0.0003
Theis et al. IFN-β	Hallmark	Vanilla	5111.8 ± 2.8	0.498 ± 0.002	0.794 ± 0.002	0.865 ± 0.005	0.9513 ± 0.0038
Theis et al. IFN-β	Hallmark	Gene set	5200.8 ± 2.3	0.384 ± 0.001	0.691 ± 0.001	0.756 ± 0.007	0.9203 ± 0.0054
Theis et al. IFN-β	Hallmark	Fully random	5200.2 ± 2.7	0.373 ± 0.005	0.636 ± 0.008	0.750 ± 0.014	0.9197 ± 0.0029
Theis et al. IFN-β	Hallmark	Within-set random	5213.1 ± 2.6	0.365 ± 0.003	0.611 ± 0.009	0.753 ± 0.008	0.9212 ± 0.0052
Theis et al. IFN-β	Hallmark	Co-expression	5192.8 ± 7.9	0.421 ± 0.006	0.703 ± 0.007	0.781 ± 0.013	0.9273 ± 0.0043
Theis et al. IFN-β	Reactome	Vanilla	5111.8 ± 2.8	0.498 ± 0.002	0.794 ± 0.002	0.865 ± 0.005	0.9513 ± 0.0038
Theis et al. IFN-β	Reactome	Gene set	5188.8 ± 3.1	0.382 ± 0.002	0.671 ± 0.004	0.768 ± 0.007	0.9256 ± 0.0033
Theis et al. IFN-β	Reactome	Fully random	5197.4 ± 2.0	0.368 ± 0.007	0.634 ± 0.013	0.744 ± 0.013	0.9184 ± 0.0037
Theis et al. IFN-β	Reactome	Within-set random	5195.2 ± 1.4	0.368 ± 0.004	0.620 ± 0.005	0.754 ± 0.008	0.9206 ± 0.0045
Theis et al. IFN-β	Reactome	Co-expression	5189.0 ± 3.3	0.425 ± 0.004	0.704 ± 0.007	0.785 ± 0.014	0.9262 ± 0.0048

Supplementary Table 5. Test loss per cell, gene set mask minus each other mask. Lower is better. Hodges-Lehmann estimate of the paired differences over 10 seeds; an asterisk marks a Holm-adjusted p below 0.05 within the 15 tests of each benchmark, and a dagger marks such a p with a CI that touches or includes zero.

Data	Collection	− fully random	− within-set random	− co-expression
Kang et al. IFN-β	Hallmark	−3.7*	−10.1*	+11.8*
Kang et al. IFN-β	Reactome	−10.8*	−11.9*	+4.6†
Theis et al. IFN-β	Hallmark	+0.8	−12.3*	+8.7
Theis et al. IFN-β	Reactome	−8.5*	−7.3*	−0.3

Supplementary Table 6. Decoded response (Spearman correlation), gene set mask minus each other mask. As in Supplementary Table 5.

Data	Collection	− fully random	− within-set random	− co-expression
Kang et al. IFN-β	Hallmark	+0.014*	+0.022*	−0.033*
Kang et al. IFN-β	Reactome	+0.027*	+0.028*	−0.027*
Theis et al. IFN-β	Hallmark	+0.012*	+0.020*	−0.037*
Theis et al. IFN-β	Reactome	+0.014*	+0.014*	−0.043*

Supplementary Table 7. Decoded response (Pearson correlation), gene set mask minus each other mask. As in Supplementary Table 5.

Data	Collection	− fully random	− within-set random	− co-expression
Kang et al. IFN-β	Hallmark	+0.087*	+0.129*	−0.026*
Kang et al. IFN-β	Reactome	+0.142*	+0.176*	−0.023*
Theis et al. IFN-β	Hallmark	+0.054*	+0.079*	−0.014†
Theis et al. IFN-β	Reactome	+0.037*	+0.052*	−0.033*

Supplementary Table 8. Cell-type balanced accuracy, gene set mask minus each other mask. As in Supplementary Table 5.

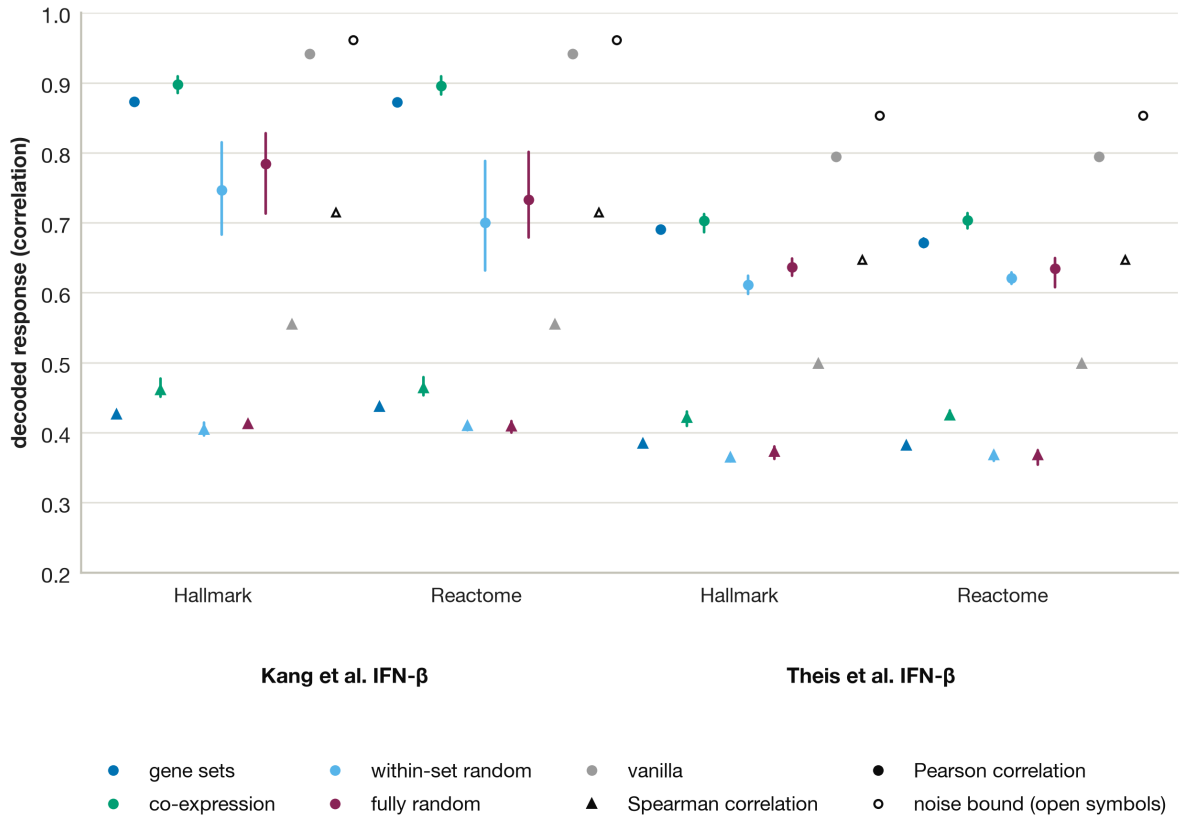
Data	Collection	− fully random	− within-set random	− co-expression
Kang et al. IFN-β	Hallmark	+0.001	+0.006	−0.009†
Kang et al. IFN-β	Reactome	+0.005	−0.003	−0.001
Theis et al. IFN-β	Hallmark	+0.006	+0.002	−0.026*
Theis et al. IFN-β	Reactome	+0.023*	+0.014†	−0.016

Supplementary Table 9. Stimulation classifier AUROC, gene set mask minus each other mask. As in Supplementary Table 5.

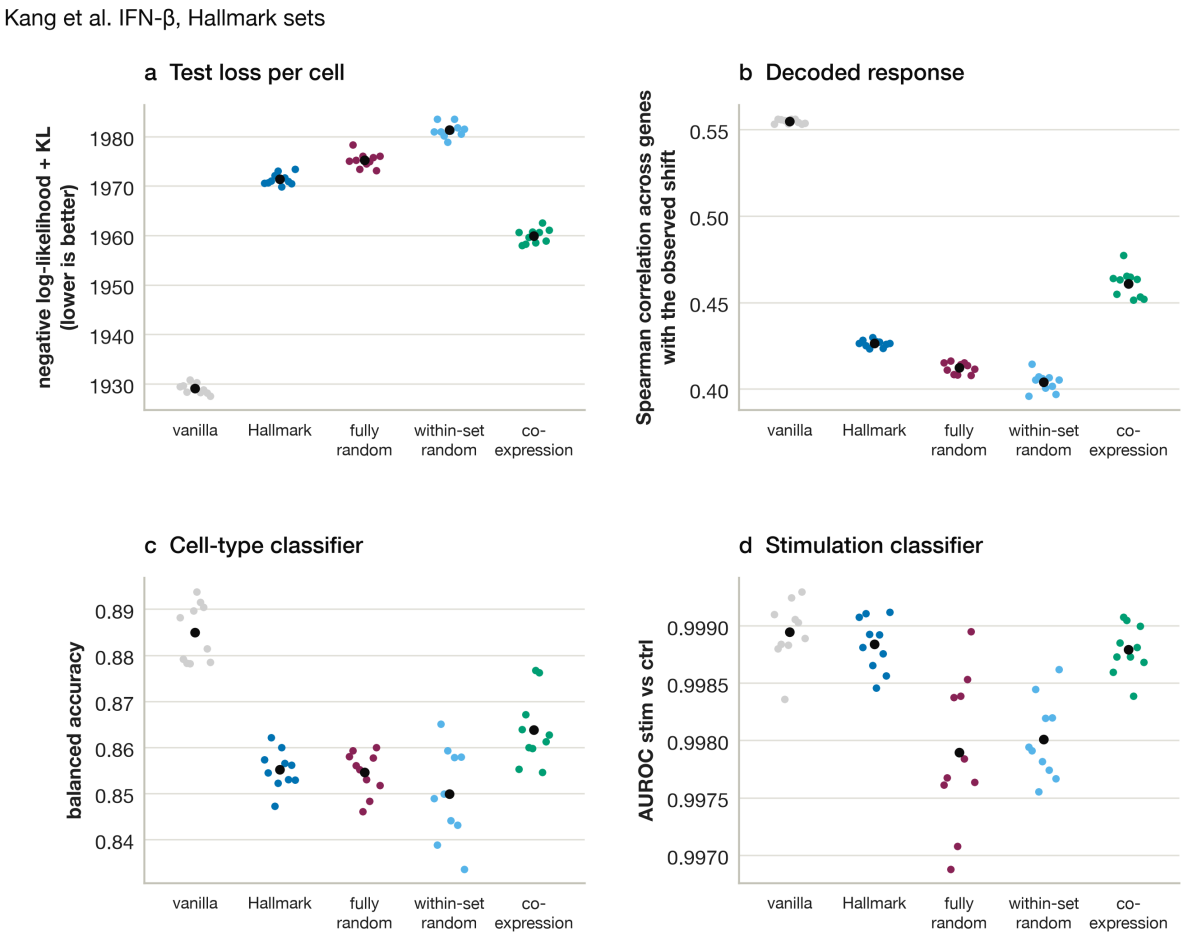
Data	Collection	− fully random	− within-set random	− co-expression
Kang et al. IFN-β	Hallmark	+0.0009†	+0.0008*	+0.0001
Kang et al. IFN-β	Reactome	+0.0009*	+0.0011*	+0.0001
Theis et al. IFN-β	Hallmark	+0.0005	−0.0018	−0.0077
Theis et al. IFN-β	Reactome	+0.0073†	+0.0043	−0.0007

Supplementary Table 10. Gene set mask minus the random masks under alternative decoded-response measures. Mean difference over 10 seeds, with the number of seeds with a positive difference in brackets. Pearson and Spearman are the two correlations of the main text; weighted is the correlation weighted by the absolute observed shift; strongest 10% and other 90% are the correlations on the genes with the largest observed shift in each cell type and on the remaining genes.

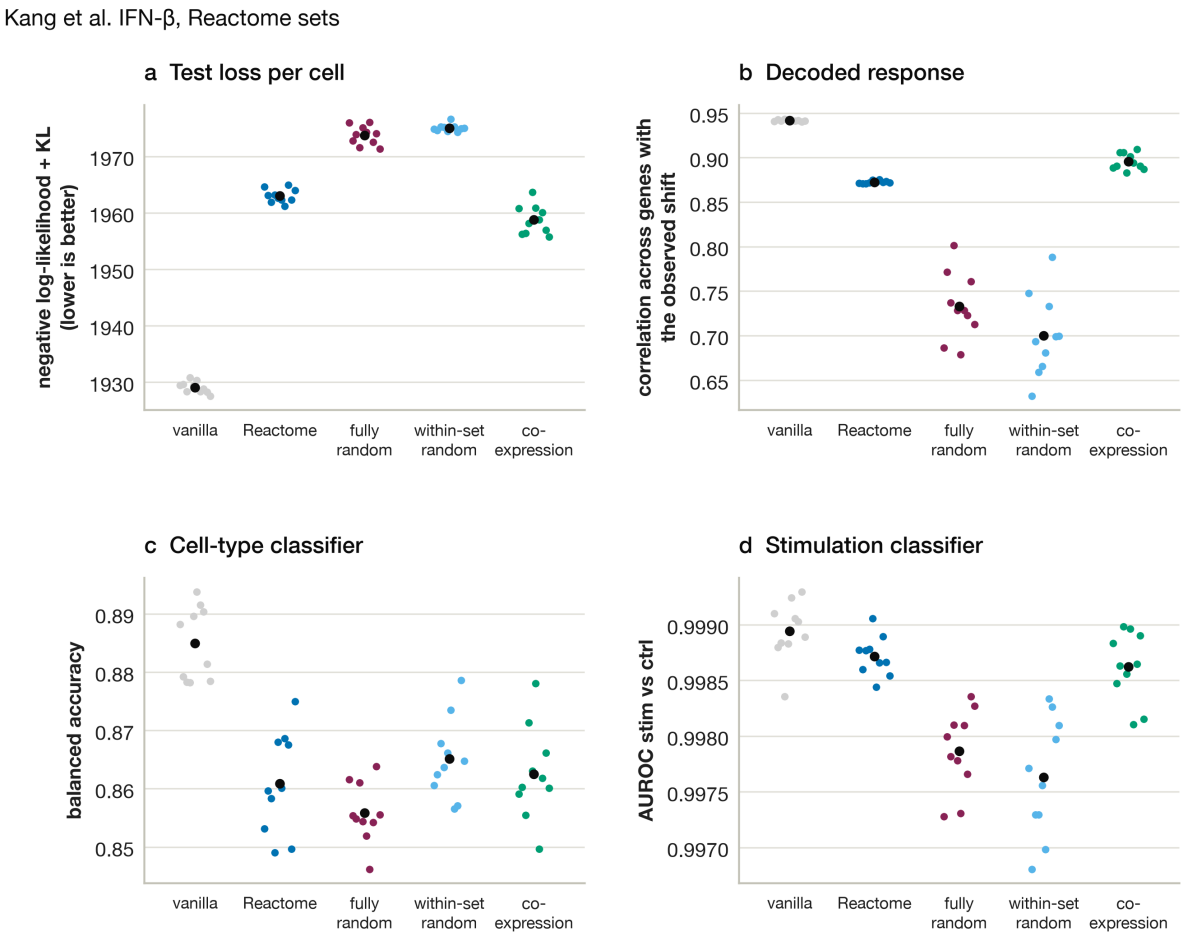
Data	Collection	Random mask	Pearson	Spearman	Weighted	Strongest 10%	Other 90%
Kang et al. IFN-β	Hallmark	fully random	+0.089 (10)	+0.014 (10)	+0.107 (10)	+0.086 (10)	+0.018 (10)
Kang et al. IFN-β	Hallmark	within-set	+0.127 (10)	+0.022 (10)	+0.144 (10)	+0.119 (10)	+0.030 (10)
Kang et al. IFN-β	Reactome	fully random	+0.140 (10)	+0.028 (10)	+0.188 (10)	+0.144 (10)	−0.009 (2)
Kang et al. IFN-β	Reactome	within-set	+0.173 (10)	+0.027 (10)	+0.206 (10)	+0.172 (10)	+0.009 (8)
Theis et al. IFN-β	Hallmark	fully random	+0.054 (10)	+0.012 (10)	+0.075 (10)	+0.057 (10)	+0.001 (5)
Theis et al. IFN-β	Hallmark	within-set	+0.079 (10)	+0.020 (10)	+0.094 (10)	+0.074 (10)	+0.012 (10)
Theis et al. IFN-β	Reactome	fully random	+0.037 (10)	+0.014 (10)	+0.048 (10)	+0.038 (10)	+0.006 (9)
Theis et al. IFN-β	Reactome	within-set	+0.051 (10)	+0.014 (10)	+0.064 (10)	+0.050 (10)	+0.006 (10)



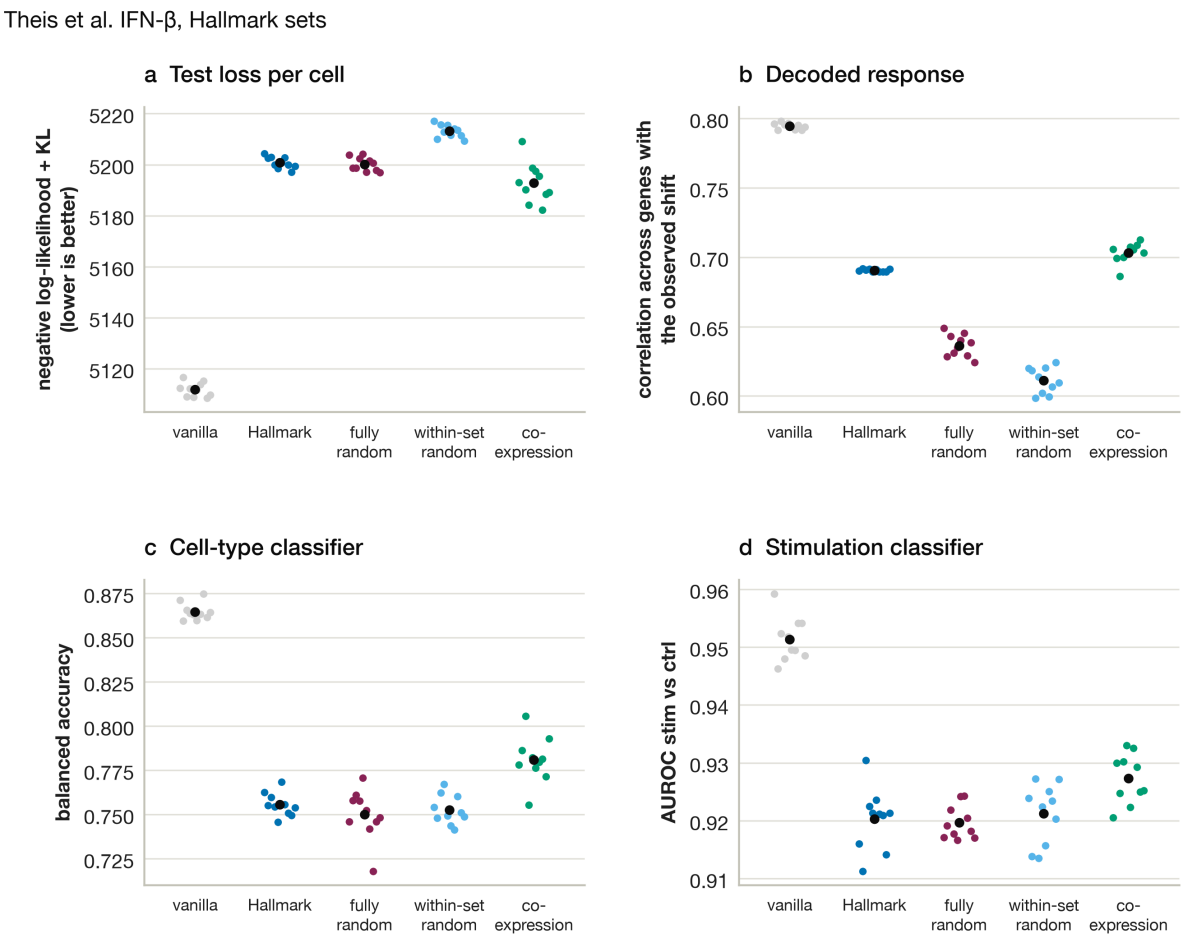
Supplementary Figure 2. Decoded response of every variant in the four benchmarks. Each point is the mean of 10 seeds and the line the range over seeds. Triangles show the Spearman and circles the Pearson correlation; the open symbols mark the noise bound (Methods).



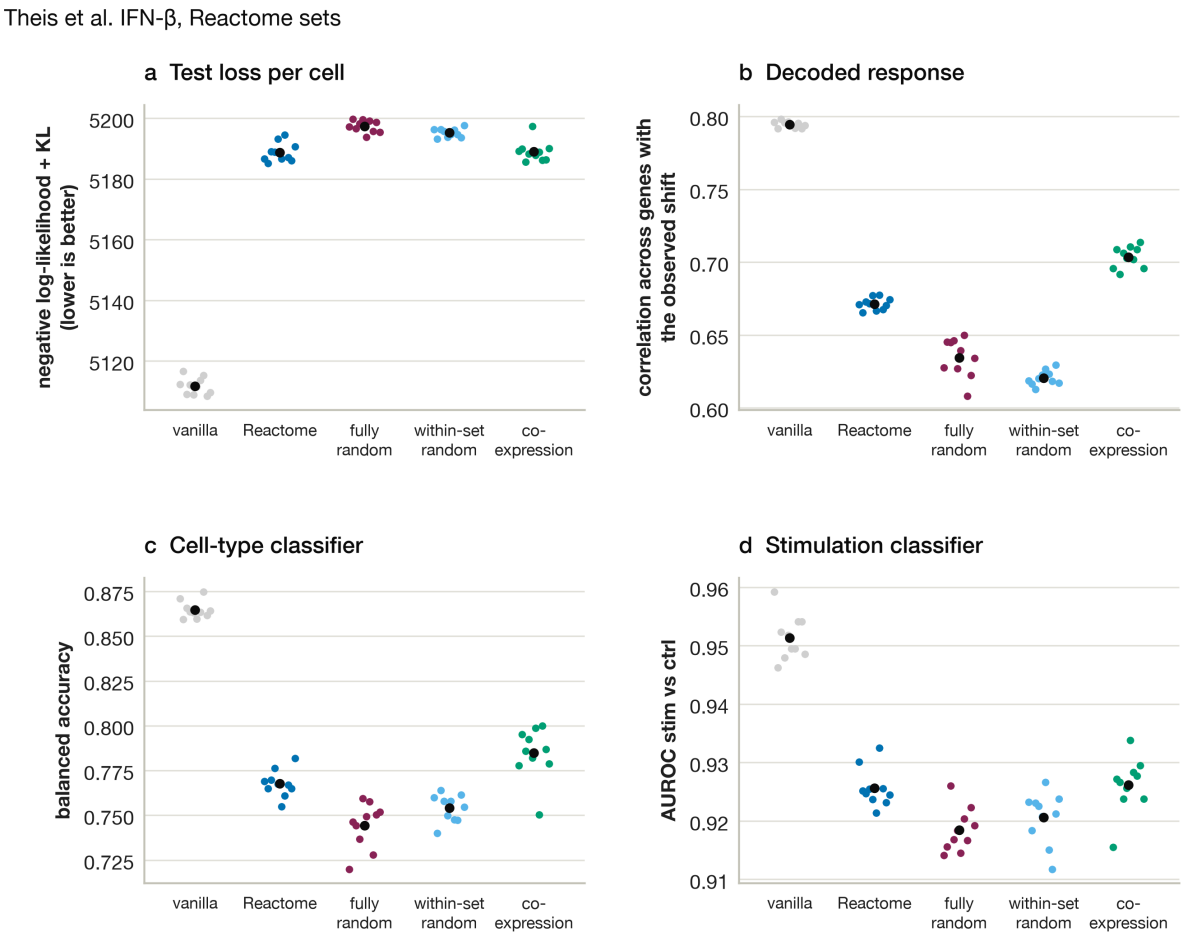
Supplementary Figure 3. Kang et al. interferon-β, Hallmark sets. a to d, test loss, decoded response (Spearman correlation), cell-type classifier and stimulation classifier (one dot per seed, black dot at the mean).



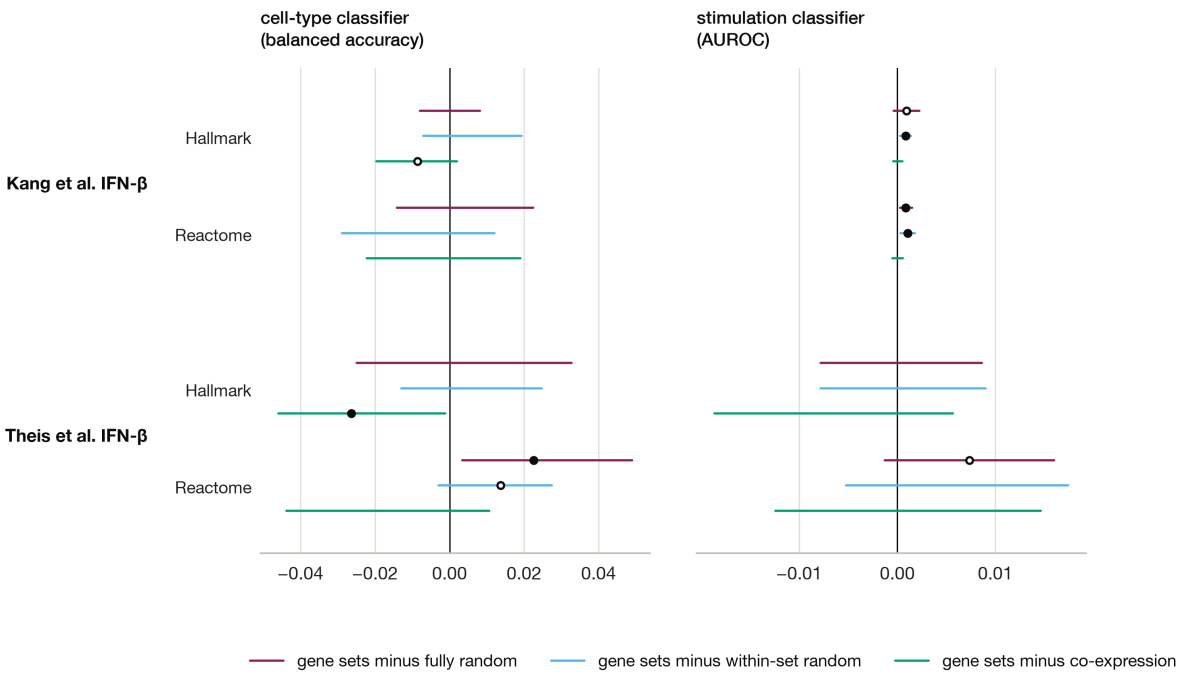
Supplementary Figure 4. Kang et al. interferon-β, Reactome sets. a to d, test loss, decoded response (Spearman correlation), cell-type classifier and stimulation classifier (one dot per seed, black dot at the mean).



Supplementary Figure 5. Theis et al. interferon-β, Hallmark sets. a to d, test loss, decoded response (Spearman correlation), cell-type classifier and stimulation classifier (one dot per seed, black dot at the mean).



Supplementary Figure 6. Theis et al. interferon-β, Reactome sets. a to d, test loss, decoded response (Spearman correlation), cell-type classifier and stimulation classifier (one dot per seed, black dot at the mean).



Supplementary Figure 7. Gene sets against the other masks in cell-type and stimulation classification, four benchmarks. 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. 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.

Supplementary Table 11. Decoded response as a share of the noise bound. Mean over 10 seeds of the decoded response divided by the noise bound of the dataset (Methods), for the Spearman and the Pearson correlation.

Data	Collectic	Correlatic	Noise bound	Vanilla	Gene sets	Fully random	Within-set random	Co-expression	Expression-matched
Kang et al. IFN-β	Hallmarl	Spearmar	0.715	78%	60%	58%	57%	65%	59%
Kang et al. IFN-β	Hallmark	Pearson	0.961	98%	91%	82%	78%	93%	85%
Kang et al. IFN-β	Reactor	Spearmar	0.715	78%	61%	57%	57%	65%	59%
Kang et al. IFN-β	Reactome	Pearson	0.961	98%	91%	76%	73%	93%	77%
Theis et al. IFN-β	Hallmarl	Spearmar	0.647	77%	59%	58%	56%	65%	58%
Theis et al. IFN-β	Hallmark	Pearson	0.853	93%	81%	75%	72%	82%	75%
Theis et al. IFN-β	Reactor	Spearmar	0.647	77%	59%	57%	57%	66%	58%
Theis et al. IFN-β	Reactome	Pearson	0.853	93%	79%	74%	73%	82%	74%

S5. Co-expression mask: full recipe

Co-expression mask. This mask tests whether pre-specified sets add anything beyond co-expressed genes found in the data. The groups are built from the data only, without condition or cell-type labels. 1. Expression Training cells only, stimulated and control together. Counts are scaled to the median total count and log-transformed ($\ln(1 + x)$), the same transform the encoder reads. 2. Correlation Pearson correlation between all 12,034 genes. A gene with no variance in training cells shows a correlation of 0 with every other gene. 3. Seed genes For each seed, 50 distinct seed genes are drawn at random from the 6,753 genes detected in $\geq 1\%$ of training cells. The nearest neighbours of rarer genes are noise, and these genes are excluded. 4. Groups 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). Group j adopts the size of Hallmark set j and none of its biology. Its latent is labelled "module j" in the code. Groups are built independently and may share genes. The groups differ from the Hallmark sets in ways that affect how the results are read. The median within-set correlation is 0.096 for the groups and 0.003 for Hallmark sets. In single PBMCs most members of a Hallmark set barely co-vary, whereas a group is built to co-vary. The groups are therefore a demanding comparison for the Hallmark sets. The groups reproduce the 5,568 memberships of the Hallmark sets but cover 2,338–3,233 genes (Hallmark: 3,187). The overlap concentrates on genes correlated with many others. Ribosomal-protein genes (RPS6, RPL3, …) and PTMA are in up to 28 groups per seed, whereas no Hallmark gene is in more than 10 sets. Several groups partly describe the same ribosomal axis. The 8 free latents reach the 8,801–9,696 unassigned genes.

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%). In one seed it grew around UBE2L6 and contains ISG15, IFI6, IFIT1, IFIT3, MX1, IRF7 and OAS1. The data recover the core interferon response without any gene set.

S6. Architecture and parameter counts

Every variant shares the same VAE, with an encoder of one hidden layer and a linear decoder (binary mask on its weight matrix in five of the six variants). The PyTorch module summary is as follows.

```
VAE(
  (encoder): Sequential(
    (0): Linear(in_features=12042, out_features=128, bias=True)
    (1): ReLU()
    (2): Dropout(p=0.1, inplace=False)
  )
  (mean): Linear(in_features=128, out_features=58, bias=True)
  (log_var): Linear(in_features=128, out_features=58, bias=True)
)
```

The encoder input is the 12,034 genes after a shifted-log transform (see Methods), plus a one-hot code for the 8 donors. These are the Kang et al. data. The Theis et al. data include about 11,340 genes and 12 donors. The encoder, $\mathbf{W}$ and $\mathbf{V}$ are smaller. The decoder has no layers of its own, only the three parameters $\mathbf{W}$, $\mathbf{V}$ and $\log\theta$ (Supplementary Table 12).

Supplementary Table 12. Parameters of the VAE. Shapes and counts for the Kang et al. data with the Hallmark sets.

Parameter	Shape	Role	Count
encoder	12,042 → 128	hidden layer	1,541,504
mean, log_var	128 → 58 (each)	mean and log-variance of the posterior $q(z \mid x, \mathbf{s})$	14,964
W	12,034 × 58	latent → gene weights, multiplied elementwise by the mask **M**	697,972
V	12,034 × 8	donor offsets (also each gene's baseline)	96,272
$\log \theta$	12,034	negative binomial dispersion per gene	12,034
total			**2,362,746**

The mask decides how many of the 697,972 weights in **W** can be trained (table in Methods, Model).

S7. Analysis choices and their reasons

Supplementary Table 13 documents all steps at which knowledge is applied or an assumption is introduced about the cells.

Supplementary Table 13. Analysis choices and their reasons.

Step	Choice	Reason
Data	Human Cytokine Dictionary (Theis et al.) for Q1c: 12 healthy donors, at most 1,200 cells per donor × condition group, genes shared with the Kang et al. data	Every donor appears in both conditions.
Cell identity	The genotype-based singlet calls and the cell-type labels of the authors, used as provided	A droplet that contains cells of two donors is identified from their genotypes, which is stronger evidence than an expression-based doublet score.
Cell quality control	Thresholds per condition × cell type	Cell types differ in RNA content, and stimulation shifts it.
Genes	All genes detected in at least 20 cells; no selection of highly variable genes	Selection can remove interferon-induced genes.
Split	Stratified by donor × condition × cell type	Every group appears in the training, validation and test cells.
Gene identity	Ensembl identifiers mapped to current HGNC symbols (1,256 symbols renamed since 2017; the 427 genes that HGNC no longer lists retain their 2017 symbol)	The gene set symbols are current.
Gene sets	Hallmark sets with at least 12 genes (all 50); for Q1b, 50 Reactome sets chosen from names, sizes, overlaps and the pathway hierarchy	Hallmark has one unambiguous target, the interferon-α response. Reactome is curated pathway by pathway and was not built from expression.
Fully random mask	Gene labels permuted within bins of expression level	Random genes are as highly expressed as the genes they replace.
Within-set random mask	Gene labels permuted within bins of expression level, among the genes that are in a set	The unassigned genes, and so the reach of the free latents, remain as in the Hallmark mask.

Step	Choice	Reason
Expression-matched random mask	Gene labels permuted among the genes that are in a set, within bins that cross expression level and stimulation effect	Every set retains its number of genes of each response strength.
Co-expression mask	Groups of co-expressed genes found in the data, with sizes matched to the Hallmark sets	The mask tests whether pre-specified sets add anything beyond groups found in the data (S5).
Free latents	Reach only the unassigned genes	With access to all genes, they absorbed the stimulation signal and the labelled latents became inactive.
Covariates	Donor in encoder and decoder; condition not used	The stimulation is the signal that the latents should find.
Likelihood	Negative binomial on raw counts, with the observed total counts as library size	Droplet counts are not zero-inflated[6].
Target latent	Hallmark interferon-α response	IFN-β is a type I interferon.
Orientation	Sign set so that the mean weight of the latent on its own genes is positive	A gene set that is switched on raises its own genes.
Scoring	AUROC per cell type on the test cells, with an unweighted mean over the scored types (7 for the Kang et al. data, 13 for the Theis et al. data)	The question is whether every cell type responds, and a latent should not score by separating cell types.
Genes of the decoded response	The genes detected in at least 1% of training cells (6,753 in the Kang et al. data)	The shifts of rarer genes rest on too few cells.
Megakaryocytes	Excluded from scores, retained in training	They are mostly platelets, which have no nucleus.

Literature search

Six Europe PMC queries (6 October 2026) searched for unsupervised single-cell models that compare curated gene sets with size-matched random gene sets: four over preprints (`SRC:PPR`, first published 3 October 2025 to 3 October 2026; 4, 7, 6 and 1 records, 18 distinct preprints) and two over published articles (`SRC:MED`, 3 October 2024 to 3 October 2026; 2 and 1 records).

```
E1 SRC:PPR AND FIRST_PDATE:[2025-10-03 TO 2026-10-03] AND (ABSTRACT:"single-cell" OR ABSTRACT:"
    single cell") AND (ABSTRACT:"gene program" OR ABSTRACT:"pathway" OR ABSTRACT:"gene set")
    AND (ABSTRACT:"random" OR ABSTRACT:"shuffled" OR ABSTRACT:"permuted") AND (ABSTRACT:"
    interpretable" OR ABSTRACT:"biologically informed")

E2 SRC:PPR AND FIRST_PDATE:[2025-10-03 TO 2026-10-03] AND (ABSTRACT:"single-cell" OR ABSTRACT:"
    single cell") AND ABSTRACT:"prior knowledge" AND (ABSTRACT:"autoencoder" OR ABSTRACT:"
    factor analysis" OR ABSTRACT:"latent")

E3 SRC:PPR AND FIRST_PDATE:[2025-10-03 TO 2026-10-03] AND (expiMap OR pmVAE OR "VEGA" OR "
    Spectra" OR slalom OR PLIER) AND (ABSTRACT:"single-cell" OR ABSTRACT:"single cell") AND (
    ABSTRACT:"gene program" OR ABSTRACT:"pathway" OR ABSTRACT:"gene set")

E4 SRC:PPR AND FIRST_PDATE:[2025-10-03 TO 2026-10-03] AND ABSTRACT:"plasma cell" AND (ABSTRACT
    :"single-cell" OR ABSTRACT:"single cell") AND (PRDM1 OR IRF4 OR BLIMP1)

E5 SRC:MED AND FIRST_PDATE:[2024-10-03 TO 2026-10-03] AND (ABSTRACT:"single-cell" OR ABSTRACT:"
    single cell") AND (ABSTRACT:"gene program" OR ABSTRACT:"pathway" OR ABSTRACT:"gene set")
    AND (ABSTRACT:"random" OR ABSTRACT:"shuffled" OR ABSTRACT:"permuted") AND (ABSTRACT:"
```

```
    interpretable" OR ABSTRACT:"biologically informed") AND (ABSTRACT:"autoencoder" OR ABSTRACT
    :"latent" OR ABSTRACT:"factor")

E6 SRC:MED AND FIRST_PDATE:[2024-10-03 TO 2026-10-03] AND (ABSTRACT:"single-cell" OR ABSTRACT:"
    single cell") AND ABSTRACT:"prior knowledge" AND (ABSTRACT:"autoencoder" OR ABSTRACT:"
    factor analysis" OR ABSTRACT:"latent") AND (ABSTRACT:"random" OR ABSTRACT:"shuffled" OR
    ABSTRACT:"permuted" OR ABSTRACT:"sparsity")
```

For each record the abstract was read, and the full texts of the three closest candidates were searched for random, shuffled or permuted gene sets. None compared a curated knowledge mask with a size-matched random mask in an unsupervised model. Tripso (Moullet et al.) relies on randomness only for seeds, token masking and data splits. MOFA-FLEX replaces 30% of the genes of each set with random non-members and tests the recovery of the original programs. scConcept employs same-size random gene sets only as the baseline for the marker-gene enrichment of one program and random concepts as a perturbation control. The remaining records were not models of this kind, such as a bulk gene-signature audit, gene regulatory network inference, foundation-model token selection and studies without a latent model. The earlier models VEGA and expiMap were checked in their full text.

References

[1] 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

[2] 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

[3] Waddell, S. J. et al. Dissecting interferon-induced transcriptional programs in human peripheral blood cells. *PLoS One* 5, e9753 (2010). doi:10.1371/journal.pone.0009753

[4] 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

[5] 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

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